Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation

Anselmo Peñas[1], Eduard Hovy[2], Pamela Forner,[3] Álvaro Rodrigo[4], Richard Sutcliffe[5], Corina Forascu[6], Caroline Sporleder [7]

[1-4] NLP&IR group, UNED, Spain (anselmo@lsi.uned.es; alvarory@lsi.uned.es)
[2] Information Sciences Institute of the University of Southern California, USA (hovy@isi.edu)
[3] CELCT, Italy (forner@celct.it)
[5] University of Limerick, Ireland (richard.sutcliffe@ul.ie)
[6] Al. I. Cuza University of Iasi, Romania (corinfor@info.uaic.ro)
[7] Saarland University, Germany (csporled@coli.uni-sb.de)

Abstract. This paper describes the first steps towards developing a methodology for testing and evaluating 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. This year a major innovation was introduced, as the traditional QA task was replaced by a new Machine Reading task whose intention was to ask questions which 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 the background collections provided by the organization. Beside the main task, also one pilot task was offered, namely, 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. This paper describes the preparation of the data sets, the creation of the background collections to allow systems to acquire the required knowledge, the metric used for the evaluation of the systems' submissions, and the results of this first attempt. Twelve groups participated in the task submitting a total of 62 runs in three languages: English, German and Romanian.

1. INTRODUCTION

Machine Reading (MR) is defined as a task that deals with the automatic understanding of texts. The evaluation of this "automatic understanding" can be approached in two ways: the first one is to define a formal language (target ontology), ask the systems to translate texts into the formal language representation, and then evaluate systems by using structured queries formulated in the formal language. The second approach is agnostic with any particular representation of the text. Systems are inquired about the text with natural language questions. The first option is approached by Information Extraction. The second is related to how Question Answering (QA) is being articulated during the last decade. In this evaluation we follow the second approach but with a significant change with respect to previous QA campaigns. Why?

By 2005 we realized that there was an upper bound of 60% of accuracy in systems performance, despite more than 80% of the questions were 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 pilot task called Answer Validation Exercise (AVE). The aim was to produce a change in QA architectures giving more responsibility to the validation step. In AVE we assumed there was a previous step of hypothesis over-generation and the hard work was 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 (Ferrucci et al., 2010).

After the three editions of AVE we tried to transfer our conclusions to the main QA task at CLEF 2009 and 2010. The first step was to introduce the option of leaving questions unanswered. This is related to the development of validation technologies. We needed a measure able to reward systems that reduce the number of questions answered incorrectly without affecting systems accuracy, by leaving unanswered the questions they estimated they couldn't answer. The measure was an extension of accuracy called c@1 (Peñas and Rodrigo, 2011), 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 using IR engines to retrieve relevant passages and then try to extract the exact answer from that. 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 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, focusing on the development of technologies able to work with a single document, and answer questions about it.

The idea of hypothesis generation and validation architecture is applicable to the new setting were only one document is considered, but of course the generation of hypotheses would be very limited if one only considers the given document. Systems should consider a large collection related to the given document in the task of hypothesis generation. Then, the validation must be performed according to the given document.

In the new setting, we started again decompounding the problem into generation and validation. Thus, in this first edition, we will test the systems only for the validation step. Together with the questions the organization provides a set of candidate answers. Besides, in this first edition, systems know there is one and only one correct answer among the candidates. This gives the evaluation the format of traditional Multiple Choice Reading Comprehension tests. From this starting point, a natural roadmap could be the following:

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 (e.g. about time and space).
    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 (e.g. expected actions of agents with a particular role, etc.)
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 were systems have to generate hypothesis considering the reference collection and provide the answer together with the set of documents that support the answer.

We are just at the beginning of this roadmap, giving space and resources for the evaluation of new QA systems with new architectures. The success of this new initiative is only measurable by the development of these new architectures able to produce a qualitative jump in performance. This vision will guide the concrete definition of the tasks year by year.

2. TASK DESCRIPTION

The QA4MRE 2011 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 five options, and only one correct answer. The detection of correct answers might require eventually 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 *cognitive 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 in the memory of the reader and it is sometimes difficult to clearly distinguish and reconstruct them in a system which needs additional knowledge and inference rules in order to understand the text and to give sensitive answers.

2.1 Main Task

By giving only a single document per test, systems are required to understand every statement and to form connections across statement in case the answer is spread over more than one sentence. Systems are requested to (i) understand the test questions, (ii) analyze 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.

Tests were divided into:

- 3 topics, namely “Aids”, “Climate change” and “Music and Society”
- Each topic had 4 reading test
- Each reading test consisted of one single document, with 10 questions and a set of five choices per question.

In global, the evaluation had in this campaign

- 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)

Test documents and questions were made available in English, German, Italian, Romanian, and Spanish. These materials were exactly the same in all languages, created using parallel translations.

2.2 Pilot Exercises

Beside the main task, also one pilot task was offered this year at QA4MRE; i.e. *Processing Modality and Negation for Machine Reading* [11]. It was coordinated by CLiPS, a research center associated with the University of Antwerp, Belgium. The task was aimed at evaluating whether systems are able to understand extra-propositional aspects of meaning like 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. Modality and negation interact to express extra-propositional aspects of meaning. More information at http://www.cnts.ua.ac.be/BiographTA/qa4mre.html

The Pilot task exploited the same topics and background collections of the main exercise. Test documents, instead, were specifically selected in order to ensure the properties required for the questionnaires. The pilot task was offered in English only.

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. At an early stage, a background collection related to the renewable energy domain was first released to participants together with some sample data. The background collection for the sample, of about 11,000 documents, was in English only. For the real test, three background collections - one for each of the topics – were released in all the languages involved in the exercise, i.e., English, German, Italian, Spanish and Romanian. Overall, fifteen large repositories as source of "background knowledge" were created to enable inferring information that is implicit in the text. These background collections are comparable (but not identical) topic-related (but not specialized) collections made available to all participants at the beginning of April by signing a license agreement. Thus, systems could "learn" and acquire knowledge in one language or several.

The only way to acquire big comparable corpora in the three domains we were interested, was crawling the web. Crawling refers to the acquisition of material specific to a given subject from the Web. The Web, with its vast volumes of data in almost any domain and language, offers a natural source for naturally occurring texts. To this end, a web crawler was specifically created by CELCT in order to gather domain-specific texts from the Web.

As for the distribution of documents among the collections, the final number of documents fetched for each language collection was different, but this is supposed to reflect the real distribution. Table 1 depicts the sizes of the corpora which were acquired and the number of documents contained in each language background collection for each of the three topics.

Table 1 : Size of the acquired background collections in the various languages for the three topics

TOPICS	DE		EN		ES		IT		RO	
	# docs	*KB*	*# docs*	*KB*	*# docs*	*KB*	*# docs*	*KB*	*# docs*	*KB*
AIDS	25,521	226,008	28,862	535,827	27,702	312,715	32,488	759,525	25,033	344,289
CLIMATE CHANGE	73,057	524,519	42,743	510,661	85,375	677,498	82,722	1238,594	51,130	374,123
MUSIC & SOCIETY	81,273	754,720	46,698	733,898	130,000	922,663	92, 036	1274,581	85,116	564,604

The corpora obtained from the process of crawling contain a set of documents which are related to the test documents. Unfortunately, the degree of noisy documents introduced is unknown.

As a final step, in order to ensure that each language background collection really contained documents which supported the inferences of the questions, each language organizer was also asked to manually search on the web for the documents, in their own language, which were to be manually added to each language collection. A list of the respective docs that should be looked for was provided by question creators to each language group.

Once all collections were ready in all languages, the zipped files were transferred to CELCT ftp server. All documents inside each collection were then re-numbered giving them a progressive unique identifier.

3.1 Keywords and Crawling

A web crawler is a relatively simple automated program, or script, that methodically scans or "crawls" through Internet pages to create an index of the data it's looking for.

The QA4MRE crawler is a flexible application designed to download a large number of documents from the World Wide Web around a specified list of keywords. It was developed using Google API, downloading documents in a ranked order, and obeying the Robot Exclusion Standard. After downloading, documents are converted in .txt format and each text is named according to the sources from which it has been downloaded, for example: "articles.latimes.com_68".

Keywords play a central role in the crawling process as they are used in acquiring the seed URLs. Before fixing the final set of keywords all people in charge of the creation of the respective language collection experimented with a preliminary pool of keywords and suggested changes to the others. Then, once the sets of keywords were standardised in English, they were translated into the other languages and loaded into CELCT's crawler. Keywords mustn't be too generic, and combination of keywords useful to restrict the domain helped to retrieve relevant documents. Synonyms or words which have very similar meaning – like for example, "climate change" and "climate variability"; "carbon dioxide" and "C02" – were kept as separate queries, as the documents which could be obtained could be different. Also, acronyms were always solved, – like for example Joint United Nations Programme on HIV (UNAIDS) – and were entered in the same query into the crawler.

In addition, as building a comparable corpus requires control over the selection of source texts in the various languages, each language group was asked to prepare a list of (trusted) web sites – indicatively a number of 40 – which were more likely to have plenty of documents related to the topic in their own language. This was required as a way to increase the number of relevant documents avoiding introducing noise (or virus files). The longer the list of domains was, the higher the number of documents which could be downloaded for each single query. Texts were drawn from a variety range of sources e.g.: newspapers, newswire, web, journals, blogs, Wikipedia entries, etc.

All keywords and all domains were entered in one crawling run. This solution allowed the removal of duplicate URLs retrieved making different queries, as the encountered URLs were kept in memory, so that every URL was visited only once. On average, it took 2-3 days to build one background collection for one topic.

Other parameters could also be set, namely the number of documents to be downloaded for each single query. By default it was set to 1000, since, due to Google restrictions, it is the maximum number of documents per query which can be downloaded for a specified source/domain. For the English language, this parameter was set to 500. In an attempt, to reduce the number of indices, and other useless files from the corpus lists, the documents which are too short were automatically discarded, by setting the minimum length of the document to 1000 characters. For the English language it was set to 1500.

4. TEST SET PREPARATION

As we have seen, the task this year was to answer a series of multiple choice tests, each based on a short document.

4.1 Test Documents

In order to allow participants to tune their systems, a set of pilot data was first devised. This consisted of three English documents concerned with the topic of renewable energy taken from Green Blog (http://www.green-blog.org/) together with three sets of questions, one for each document, and a background collection of about 11,000 documents. For each document there were ten multiple choice questions; each question had five candidate answers, one clearly correct answer and four clearly incorrect answers. The task of each system was therefore to choose one answer for each question, by analysing the corresponding test document in conjunction with the background collection.

Following the creation of the pilot data, attention was turned to the materials for the actual evaluation. The languages this year were English, German, Italian, Romanian and Spanish. The intention was to set identical questions for these five languages. This implied that we had access to a suitable parallel collection of documents so that each test document was exactly translated into each language of the task. Unfortunately, even after decades of interest in parallel corpora, very few publicly available high quality collections exist in these five languages. The main possibilities available to us were "Eurobabble" and technical manuals, but each was

somewhat unsuitable for the task. Another option was for us to commission special translations of selected documents in, say, English, just for the purposes of QA4MRE.

After some consideration, we took up a suggestion of Igal Gabbay to use documents taken from the Technology, Entertainment, Design (TED) conferences (www.ted.com). Each TED event consists of a series of invited presentations by prestigious speakers, from fields such as politics, entertainment and industry. The speakers are fluent, persuasive, and mostly speak from memory with no repetition or hesitation. Each talk lasts for twenty minutes or less and is aimed at a non-specialised but reasonably educated audience. The organisers provide for each talk a high-quality text transcription. In the case of the talks used, this ranges in length between 1125 and 3580 words. However, they also provide an infrastructure for the transcriptions to be translated by volunteers. These translations are carefully refereed and are generally of very high quality. The number of languages in which a talk is available varies, depending on its popularity, but is typically 20-40.

Table 2 : TED Test Documents

Topic	No.	Author	Title	Words
AIDS	1	Annie Lennox	Why I am an HIV/AIDS activist	1378
AIDS	2	Bono	Bono's call to action for Africa	3580
AIDS	3	Elizabeth Pisani	Sex, drugs and HIV -- let's get rational	3178
AIDS	4	Emily Oster	Emily Oster flips our thinking on AIDS in Africa	3299
Climate Change	5	Al Gore	Al Gore warns on latest climate trends	1235
Climate Change	6	Al Gore	Al Gore's new thinking on the climate crisis	3190
Climate Change	7	David Keith	David Keith's unusual climate change idea	3314
Climate Change	8	Lee Hotz	Inside an Antarctic time machine	1308
Music & Society	9	Adam Sadowsky	Adam Sadowsky engineers a viral music video	1125
Music & Society	10	Ben Cameron	The true power of the performing arts	2000
Music & Society	11	David Byrne	How architecture helped music evolve	2213
Music & Society	12	Jose Abreu	Jose Abreu on kids transformed by music	1679

From the perspective of QA4MRE, TED transcriptions have some good points and some bad ones. On the one hand, they are of high typographical and syntactic quality, they discuss clearly-defined topics, they are at a reasonable intellectual level, they are available translated accurately into many languages and they are of course publicly available. On the other hand, they are on the short side, and, length-for-length contain less facts amenable to the generation of questions than might be the case for other kinds of document. They may also contain jokes or digressions, or material which can only be comprehended in the context of film clips, photographs or recordings which are used in the talk but which of course do not appear in the transcription. Finally, the transcriptions can contain phrases such as "laughter", "applause" or "music" from time to time. These, of course, are describing events at the talk itself and are thus not a transcription of anything that was said.
Having decided on the source of documents, three topics were then chosen, AIDS, Climate Change, and Music and Society. For each topic, four TED talks were selected, each having transcripts available in English, German, Italian, Romanian and Spanish. Table 2 lists the selected talks. Ten multiple-choice questions were then devised for each talk. As in the pilot materials, a question always had five candidate answers from which to choose, with one clearly correct answer and four clearly incorrect answers.

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 were carefully checked by a referee to verify that they were clear, that the intended answer was clearly correct, that the intended answer was in the test document, and that the other candidate answers were clearly incorrect. Questions were modified accordingly. The English versions were then used to translate each question into each of the five languages of the task. The same process was used to translate each candidate answer (five per query) into the five languages.

The result of this process was a set of 120 questions in five languages, each with five multiple-choice answers, also in those five languages. The final step was to check that the answer to each question was in fact present in the test document for all the languages of the task. Occasionally, certain parts 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 made the question impossible to answer in that language. In such cases, the question had to be withdrawn from all languages and a new one devised to take its place.

In parallel with the above activity, a background collection was created for each of the three topics, as described in Section 3 above. The questions, test documents and background collections were now ready to be used in the QA4MRE task.

4.2 Questions

Unlike previous campaigns, where the aim was mainly to ask factoid questions involving the extraction of simple information (mainly Named Entities) from large collections of long documents, the intention in QA4MRE was to ask more searching questions which required a deep knowledge of individual short texts.
Concerning test queries, as is usual practice in the QA campaign, they were artificially constructed from portions of the text to match the criteria we wanted to test in this task.
The QA4MRE questions were also created taking into consideration different levels of difficulty. They may refer to:

- facts that (as in traditional QA evaluation) are explicitly present in the text
- facts that are explicitly present but are not explicitly related (for example, they do not appear in the same sentence, although any human would understand they are connected)
- facts that are not explicitly mentioned in the text, but that are one inferential step away (as in the RTE challenge)
- facts that are explicitly mentioned in the text but that require some inference to be connected to form the answer

Table 3 : Number of Questions which need background knowledge to be answered

number of questions : 120	**info from background collection required : 44**	**no extra knowledge is needed : 76**
		info from different paragraphs: 38
		answer in the same sentence/para. : 38

Out of the 120 questions given in the test set, 44 of them needed some extra information from the background collection in order to be answered, while for 76 questions the information present in the text document alone was enough to select the correct answer. More in details, as Table 3 shows, 38 questions had the answer contained in the same sentence/paragraph; while for 38 questions the system had to assemble information from different paragraphs in order to answer the question. In addition, questions were also posed so that the answers were not merely a mechanical repetition of the input question, but all kinds of textual inferences could be requested, i.e., lexical (acronym, synonymy, hyperonymy-hyponymy), syntactic (nominalization-verbalization, causative, paraphrase, active-passive), discourse (co-reference, anaphora ellipsis).

Table 4 : Examples of Questions

Type	Topic	Test	Question	Answers (Correct One in Bold)
CAUSE	Climate Change	5	What could be a consequence of a reduction in Arctic ice?	higher sea level / **more atmospheric pollution** / less drinking water / more fires / less droughts
COMPOSITE	Climate Change	7	What solution not tested by humans could contribute to reducing the climate change problem?	the use of renewable energies / the reduction of CO2 emissions / investment in climate change by the governments of developed countries / **the introduction of signed particles into the stratosphere** / the protection of the environment
DEGREE-OF-TRUTH	AIDS	1	Do people agree that governments should be committed to fighting AIDS?	**definitely yes** / definitely no / unknown / sometimes / only one person agrees
FACTOID-LOCATION	AIDS	2	Which African country did Bono Vox visit?	Somalia / Horn of Africa / Sudan / **Abyssinia** / Eritrea
FACTOID-NUMBER	Music & Society	9	How many times was OK Go's video viewed?	**more than 50 million** / fifty / a million / 10,000 / 85
FACTOID-PERSON	AIDS	3	Who wrote "People do stupid things. That's what spreads HIV"?	**Elizabeth Pisani** / Frankie / a friend of Elizabeth Pisani / the brother of Elizabeth Pisani / a drug addict
FACTOID-LIST	Music & Society	10	What are two difficulties associated with attending a live performance?	jeans and set curtain times / **parking and set curtain times** / internet and set curtain times / customisation and set curtain times / body types and set curtain times

FACTOID-TIME	Climate Change	7	When was a newspaper article published on climate change?	at the beginning of 2000 / in the 90s / in 1965 / **in the 1950s** / in 2075
FACTOID-UNKNOWN	Climate Change	8	What other information important for climate change is stored in Antarctic ice?	ozone gasses / a register of ocean currents / the amount of precipitation / crystals / **measurements of the Earth's temperature**
HYPO-THETICAL	Climate Change	6	What consequence would the use of renewable energies have in the US?	**new job opportunities** / a higher dependency on fossil fuels / a decrease in toxic dumps / a higher use of clean coal / less responsible use of energy
METHOD	AIDS	4	How are people infected by HIV?	through aerial transmission / through genetic transmission / through direct contact with infected people / through the faecal-oral route / **through sexual intercourse**
OPINION	AIDS	2	What is Bono's attitude towards the digital age?	sorrow / sadness / **enthusiasm** / indifference / anger
PURPOSE	Music & Society	11	Why did the Bayreuth Festspielhaus have a large orchestra pit?	to eat, drink and yell out / to be more intricate / to help Mozart / to suggest an encore / **to accommodate low-end instruments**
RESULT	Climate Change	6	Where has carbon from the Earth's atmosphere gone?	it is still in the atmosphere / **to form fossil fuels** / to be part of the arctic ice cap / to pollute the air / to create acid rain
WHICH--IS-TRUE	Music & Society	12	What is the worst thing about being poor?	the enjoyment of music / the satisfaction of playing / **the feeling of being no-one** / the lack of food / the lack of shelter

Concerning the types of questions which would be asked, it had originally been proposed that there would be four: FACTOID, CAUSE, HYPOTHETICAL and COMPOSITE. However, following the creation of the pilot materials, six further question types were suggested: DEGREE-OF-TRUTH, METHOD, OPINION, PURPOSE, RESULTS and WHICH-IS-TRUE. Furthermore, FACTOIDs are broken down into LOCATION, NUMBER-CALC, PERSON, STATED-LIST, TIME and UNKNOWN-TYPE Examples of the types can be seen in Table 4 with a breakdown by frequency in Table 5. Unlike in previous campaigns, questions were not required to fall into the ten types in a pre-determined distribution. As can be seen in Table 5, about half the questions (64 out of 120) were FACTOID, 17 were CAUSE and 16 were WHICH-IS-TRUE. There were between one and five instances of each of the remaining types.

Table 5 : Distribution of question types

Question type	**Total number of questions**
CAUSE	17
DEGREE-OF-TRUTH	1
COMPOSITE	2
FACTOID	64
HYPOTHETICAL	4
METHOD	5
OPINION	3
PURPOSE	4
RESULTS	4
WHICH-IS-TRUE	16
Total	120

Table 6 shows the proportion of correct answers and of NoA answers given by all systems to each different question type. Degree of truth seem to be the easiest type of question to be answered while composite and hypothetical questions appear to be the most difficult to be approached. However, system seem to be less confident in answering methods and opinion questions.

Table 6 : Percentage of Correct and NoA answers according to different question type

Question type	% of correct answers	% of NoA answers
CAUSE	0,18%	0,39%
DEGREE-OF-TRUTH	0,40%	0,40%
COMPOSITE	0,15%	0,30%
FACTOID *	0,30%	0,38%
HYPOTHETICAL	0,16%	0,31%
METHOD	0,28%	0,50%
OPINION	0,23%	0,49%
PURPOSE	0,24%	0,38%
RESULTS	0,31%	0,33%
WHICH-IS-TRUE	0,29%	0,37%

4.3 Tools and Infrastructure

Also this year, CELCT developed a series of infrastructures to help the management of the QA4MRE exercise. Many processes and requirements were to be dealt with:

- the need to develop a proper and coherent tool for the management of the data produced during the campaign, to store it and to make it re-usable, as well as to facilitate the analysis and comparison of results
- the necessity of assisting the different organizing groups in the various tasks of the data set creation and to facilitate the process of collection and translation of questions
- the possibility for the participants to directly access the data, submit their own runs (this also implied some syntax checks of the format), and later, get the detailed viewing of the results and statistics.

A series of automatic web interfaces were specifically designed for each of these purposes, with the aim of facilitating the data processing and, at the same time, showing the users only what they needed for the task they had to accomplish. So, the main characteristics of these interfaces are the flexibility of the system specifically centred on the user's requirements.
While designing the interfaces for question collection and translation one of the first issues which was to be dealt with, was the fact of having many assessors, a big amount of data, and a long process. So tools must ensure an efficient and consistent management of the data, allowing:

1. Edition of the data already entered at any time.
2. Revision of the data by the users themselves.
3. Consistency propagation ensuring that modifications automatically re-model the output in which they are involved.
4. Statistics and evaluation measures are calculated and updated in real time.

In particular, ensuring the consistency of data is a key feature in data management. For example, if a typo is corrected in the Translation Interface, the modification is automatically updated also in the GoldStandard files, in the Test Set files, etc.

5. EVALUATION

Participating systems could give one of two possible responses for each question in the test collection:

- To give one answer selected from the five candidate answers of the question
- not to answer the question if a system considered that it did not have enough evidences for selecting one of the candidate answers as the correct one. This option is called NoA answer. In order to evaluate the ability of validating its answers, the system could return in this case the candidate answer that it would select in case of having to answer the question.

Taking into consideration these two possible responses, each question receives one (and only one) of the three following assessments:

- *correct* if the system selected the correct answer among the five candidate ones of the given question

- *incorrect* if the system selected one of the wrong answers
- *NoA* if the system chose not to answer the question

The evaluation of the output given by participating systems was performed automatically by comparing the answers of systems against the gold standard collection with human-made annotations. No manual assessment was required.
The task developed this year allowed us to evaluate systems from two different perspectives:

1. A question-answering evaluation, as the traditional evaluation performed in past campaigns. In this evaluation, we just accounted answers without grouping them.
2. On the other hand, we can perform a reading-test evaluation, obtaining figures for each particular reading test, and as a part of a topic.

5.1 Evaluation Measure

The purpose of allowing NoA answers is to reduce the amount of incorrect responses, while keeping the number of correct ones, by leaving some questions unanswered. As the main evaluation measure for this year's campaign *c@1* was used, which takes into account the option of not answering certain questions. *c@1* was firstly introduced in ResPubliQA 2009 [8] and is fully described in (Peñas and Rodrigo, 2011). The formulation of c@1 is given in (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

c@1 acknowledges the option of giving NoA answers in the proportion that a system answers questions correctly, which is measured using *accuracy*. Thus, a higher *accuracy* over answered questions would give more value to unanswered questions, and therefore, a higher final *c@1* value.

As a secondary measure, we also provided scores according to *accuracy* (2), the traditional measure applied to QA evaluations that does not distinguish between answered and unanswered questions. We used also the candidate answer given to unanswered questions to obtain *accuracy* values.

$$accuracy = \frac{n_R + n_{UR}}{n} \quad (2)$$

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

A question-answering evaluation has been performed over the whole test collection. This evaluation measures the overall performance of a system, without analyzing the behaviour over a particular reading test. The information taken into account for each system at this level is:

- number of questions ANSWERED
  - number of questions ANSWERED with RIGHT answer
  - number of questions ANSWERED with WRONG answer

- number of questions UNANSWERED
  - number of questions UNANSWERED with RIGHT candidate answer

- number of questions UNANSWERED with WRONG candidate answer
- number of questions UNANSWERED with EMPTY candidate answer

More in detail, the evaluation at this level includes:

- Overall *c@1*(over the 120 questions of the test collection)
- *c@1* per topic (over the 40 questions of each topic)
- Overall *accuracy* (over the 120 questions of the test collection, considering also the candidate answers given to unanswered questions)
- Proportion of answers correctly discarded (see (3))

$$correctly\_discarded = \frac{n_{UW} + n_{UE}}{n_{UR} + n_{UW} + n_{UE}} \quad (3)$$

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-test 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 multiple-choice tests consisting of ten questions per document.

This evaluation is performed taking as reference the *c@1* values achieved for each test (one document with ten questions about it). Then, the c@1 values were aggregated at topic and global levels:

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

The *median c@1* has been provided under the consideration that it can be more informative at reading-test level than average values. This is because *median* is less affected by outliers than *average*, and therefore, it offers more information about the ability of a system to understand a text. For example, if we have three high *c@1* values in a topic, but the last one is very low, the *median* is not affected by this low result (because it is an isolated result in comparison with the other three), while *average* accounts for this bad behaviour.

5.4 Random Baselines

In order to offer some baselines for this task, it must be considered that participating systems can decide to answer or not to answer a given question. Then, we firstly propose the use of a random baseline where all the questions are answered. This baseline has five possibilities when trying to answer a question: it can select the correct answer to the question, or it can select one of the four incorrect answers. In this case, the overall result is 0.2 (both for accuracy and for c@1).

6. PARTICIPATION and RESULTS

Out of the 25 groups which had previously registered and signed the license agreement to download the background collections, a total of 12 groups participated in the QA4MRE tasks submitting 62 runs in 3 different languages (German, English, and Romanian). Table 7 shows the runs submitted in each language. No runs were submitted either in Italian, or - quite surprisingly - in Spanish (usually the second most chosen language). All runs were monolingual; no team attempted a cross-language task. This was probably due to the fact that crossing the language boundary is currently not core to the task, even though multiliguality is directly addressed through the provision of collections and tests in five languages.

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, web, Wikipedia, etc., or other types of inferences. All resources used to acquire the knowledge were to be listed in the submission file.

Beside specifying the resources used, systems were required to list also the document(s) and sentence(s) that helped them (directly or indirectly) to identify the correct answer. Such provenance was not used for formal evaluation, but for informal analysis and discussion.

Table 7 : Tasks and corresponding numbers of submitted runs.

	Target languages (corpus and answer)						
Source languages (questions)		DE	EN	ES	IT	RO	**Total**
	DE	11					**11**
	EN		42				**42**
	ES						**0**
	IT						**0**
	RO					**9**	**9**
	Total	**11**	**42**	**0**	**0**	**9**	**62**

As usual, the vast majority of the runs were in English, as Table 7 shows. The list of participating teams and the reference to their reports are shown in Table 8. Beside Europe, participants came also from USA, China and India.

Table 8: Teams with the reference to their reports

Team	**Reference**
Jadavpur University, India	Pakray et al.
Ca' Foscari University, Italy	-
LIMSI-CNRS, France	-
Universidade de Évora, Portugal	Saias and Quaresma
NEC Laboratories, USA	-
Daiict, India	Arora
AL.I.Cuza University, Romania	Iftene et al.
University of Hagen, Germany	Glockner et al.
Radboud University Nijmegen, The Netherlands	Verberne
University of Heidelberg, Germany	Babych et al.
UNED, Spain	Martinez-Romo and Araujo
Fudan University, China	Cao et al.

Table 9 illustrates the mean scores for each of the 12 reading tests considering all systems (all the values in the following tables refer to *c@1*) . This shows the difficulty of each particular test. Test 3 (Topic 1) at 0.09 is lower than all the others. So, this appeared to be a very hard test. On the contrary, Test 9 (Topic 3) at 0.32 is higher than all the others by 0.05. So, this test seems to be somewhat easier.

Table 9: Mean Scores for each Reading Test

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Test 11	Test 12
Mean	0.24	0.19	0.09	0.17	0.18	0.22	0.16	0.24	0.32	0.27	0.18	0.20

Concerning the overall difficulty of the exercize, Topic 3 was the easiest and Topic 1 was the hardest but the range of difficulty is not huge, as Table 10 demonstrates. So, the three topics look fairly balanced. Also, average performances do not exceed too much the random baseline (0.2).

Table 10: Mean Scores for each Topic

	Topic 1	Topic 2	Topic 3
Mean	0.18	0.20	0.25

The following three tables (14-15-16) show the best run for each participating group, reporting the mean of the tests for each topic. Except for one case, the overall mean is higher that the baseline.

Table 14 : Results for English

	Overall	Topic 1	Topic 2	Topic 3
RUN NAME	Mean	Mean	Mean	Mean
jucs1106enen	0.58	0.80	0.53	0.42
ifln1102enen	0.37	0.28	0.45	0.37
uaic1110enen	0.28	0.25	0.27	0.32
fdcs1102enen	0.27	0.27	0.20	0.34
uned1101enen	0.27	0.19	0.36	0.26
base1101enen	0.26	0.23	0.21	0.35
swai1101enen	0.25	0.24	0.21	0.29
iles1108enen	0.24	0.25	0.13	0.33
diue1102enen	0.20	0.12	0.31	0.17
random baseline	0.20	0.20	0.20	0.20
vens1101enen	0.18	0.17	0.06	0.31

Table 15: Results for German

	Overall	Topic 1	Topic 2	Topic 3
RUN NAME	Mean	Mean	Mean	Mean
uhei1102dede	0.23	0.18	0.34	0.16
loga1102dede	0.21	0.20	0.24	0.19
random baseline	0.20	0.20	0.20	0.20

Table 16: Results for Romanian

	Overall	Topic 1	Topic 2	Topic 3
RUN NAME	Mean	Mean	Mean	Mean
uaic1107roro	0.26	0.10	0.39	0.29
random baseline	0.20	0.20	0.20	0.20

As for system performances at the question-answering evaluation level we can generally see that only one team (jucs) is above 50%, showing a large room for improvement.
From a reading test perspective, in general no group passed the reading tests, and all system seem to be very close to random guessing. Overall results at reading test level, i.e., median, average, and standard deviation for all runs are given in Appendix 1.

Table 11: Results for English

System	c@1	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.95			0	0	0	0
jucs1106enen	0.57	58	40	22	0	0	22
jucs1107enen	0.47	52	57	11	0	0	11
ifln1102enen	0.37	42	71	7	0	0	7
ifln1105enen	0.35	40	73	7	0	0	7
ifln1101enen	0.34	32	56	32	0	0	32
ifln1104enen	0.33	31	57	32	0	0	32
jucs1104enen	0.32	38	82	0	0	0	0
jucs1105enen	0.32	38	82	0	0	0	0
uaic1110enen	0.29	25	47	48	12	34	2
fdcs1102enen	0.28	22	38	60	0	0	60
base1101enen	0.27	26	64	30	0	0	30
uned1101enen	0.27	24	53	43	0	0	43
fdcs1103enen	0.26	25	65	30	0	0	30
swai1101enen	0.26	24	62	34	0	0	34
iles1108enen	0.24	28	91	1	0	0	1
uned1109enen	0.24	20	47	53	0	0	53
iles1107enen	0.23	27	93	0	0	0	0
iles1110enen	0.22	26	94	0	0	0	0
diue1102enen	0.21	18	55	47	0	0	47
jucs1103enen	0.21	25	95	0	0	0	0
uned1102enen	0.21	17	46	57	0	0	57
iles1109enen	0.20	24	96	0	0	0	0
uned1103enen	0.20	16	44	60	0	0	60
random baseline	0.20						
iles1106enen	0.19	14	34	72	0	0	72

iles1103enen	0.18	00	98	0	0	0	0
vens1101enen	0.18	19	81	20	0	0	20
diue1101enen	0.17	15	59	46	0	0	46
iles1104enen	0.17	20	100	0	0	0	0
iles1105enen	0.17	20	100	0	0	0	0
swai1105enen	0.17	14	53	53	0	0	53
uned1105enen	0.17	13	37	70	0	0	70
jucs1101enen	0.16	19	101	0	0	0	0
jucs1102enen	0.16	19	101	0	0	0	0
uned1104enen	0.16	12	41	67	0	0	67
uned1106enen	0.15	11	34	75	0	0	75
iles1102enen	0.14	9	10	101	0	0	101
uned1107enen	0.14	10	29	81	0	0	81
swai1104enen	0.09	6	21	93	0	0	93
iles1101enen	0.08	5	6	109	0	0	109
swai1102enen	0.06	4	11	105	0	0	105
uned1108enen	0.03	2	14	104	0	0	104
swai1103enen	0.02	1	2	117	0	0	117

Table 12: Results for German

System	c@1	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.34			0	0	0	0
uhei1109dede	0.24	22	60	38	0	0	38
uhei1102dede	0.23	19	43	58	0	0	58
loga1101dede	0.22	21	67	32	0	0	32
loga1102dede	0.22	21	67	32	0	0	32
uhei1103dede	0.22	18	45	57	0	0	57
random baseline	0.20			0	0	0	0
uhei1106dede	0.19	13	20	87	0	0	87
uhei1104dede	0.18	14	43	63	0	0	63
uhei1108dede	0.18	14	36	70	0	0	70
uhei1105dede	0.17	13	43	64	0	0	64
uhei1107dede	0.16	11	14	95	0	0	95
uhei1101dede	0.13	9	18	93	0	0	93

Table 13: Results for Romanian

System	c@1	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.28			0	0	0	0
uaic1107roro	0.26	30	85	5	0	0	5
uaic1101roro	0.23	27	88	5	0	0	5
uaic1109roro	0.23	19	43	58	11	42	5
uaic1103roro	0.21	18	53	49	9	35	5
uaic1104roro	0.21	17	46	57	10	42	5
uaic1106roro	0.21	17	46	57	10	42	5
random baseline	0.20						
uaic1108roro	0.16	11	19	90	19	66	5
uaic1105roro	0.15	10	21	89	17	67	5
uaic1102roro	0.14	10	23	87	17	65	5

A summary of the applied methods and techniques reported by participants is given in Table 17-18-19 in Appendix 2.

8. RELATED WORK

The current state of development of the NLP technologies offers a good opportunity for proposing an evaluation of MR systems. The opportunity arises from the clear evolution of NLP systems towards a deeper level of text analysis that allows a better understanding of documents. In fact, the interest in MR among different research groups over the world has increased recently as the creation of the MR program at DARPA[1] testifies. The large community involved in Machine Reading is searching a way to evaluate their systems. But the problem of how to evaluate these machines is still an open research issue.

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)[2], and was continued at the Cross Language Evaluation Forum (CLEF)[3] in the EU, and at the NII-NACSIS Test Collection for IR Systems (NTCIR)[4] in Japan. Most of the questions used in these evaluations ask about facts (i.e. Who is the president of XYZ?) or dentitions (i.e. 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 [4]. 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) [5.6.7]. 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[5] with the aim of studying the impact of RTE systems in *Knowledge Base Population* (KBP)[6]. The objective of this pilot task is to validate the output of participant systems at the KBP slot illing task that was celebrated at the Text Analysis Conference (TAC)[7]. 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"[8]. 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 [7], where the authors proposed to use also RC tests. However, the objective of these tests was to extract correct answers from documents, which is similar to QA without an IR engine.

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

---

[1] http://www.darpa.mil/ipto/programs/mr/mr.asp
[2] http://trec.nist.gov/
[3] http://www.clef-campaign.org/
[4] http://research.nii.ac.jp/ntcir/
[5] http://www.nist.gov/tac/2010/RTE/index.html
[6] http://nlp.cs.qc.cuny.edu/kbp/2010/
[7] http://www.nist.gov/tac/2010/
[8] http://www.aclweb.org/anthology/W/W00/#0600

9. CONCLUSIONS

This year, the QA @ CLEF task was characterized by a major innovation, namely the transition from the traditional Question Answering (QA) task, proposed in the last eight QA challenges at CLEF, to a new evaluation focus on the reading of a single document. The main reason behind this choice was the feeling that most systems were ready to make a definitive move towards a deeper understanding of the text. Along the years, the QA challenges adopted simple questions which required almost no inferences to find the correct answers. These surface-level evaluations have promoted QA architectures based on Information Retrieval (IR) techniques, in which the final answer(s) is/are obtained after focusing on selected portions of retrieved documents and matching sentence fragments or sentence parse trees. No real understanding of documents was performed, since none was required by the evaluation. Machine Reading (MR), instead, requires the automatic understanding of texts at a deeper level, so this methodology encourages the development of systems able to perform a deep analyses of the text.

One way of evaluating the understanding of a text is to assess the ability to answer a set of questions about it. In particular, reading comprehension tests are designed to measure how well human readers understand what they read. Each text comes with a set of questions about information that is stated or implied in the text.

The objectives of the task are twofold: (i) to propose a task where a deeper level of understanding is required (ii) to extract the knowledge contained in texts as a way to improve the performance of systems where some kinds of reasoning are required. Hence, the development of MR technologies should be fostered and the number of groups interested in the task should increase. This is also an opportunity to create a common framework and community in the field of text understanding.

ACKNOWLEDGMENTS

Special thanks are also due Giovanni Moretti (CELCT. Trento. Italy) for the technical support in the management of all data of the campaign.
This work has been partially supported by the Research Network MA2VICMR (S2009/TIC-1542) and Holopedia project (TIN2010-21128-C02).

REFERENCES

1. Anselmo Peñas. Álvaro Rodrigo. Felisa Verdejo. Overview of the Answer Validation Exercise 2007. In C. Peters. V. Jijkoun. Th. Mandl. H. Müller. D.W. Oard. A. Peñas. V. Petras. and D. Santos. (Eds.): Advances in Multilingual and Multimodal Information Retrieval. LNCS 5152. September 2008.

2. Anselmo Peñas. Álvaro Rodrigo. Valentín Sama. Felisa Verdejo. Overview of the Answer Validation Exercise 2006. 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. CLEF 2006. Alicante. Spain. September 20-22. 2006. Revised Selected Papers.

3. Álvaro Rodrigo. Anselmo Peñas. Felisa Verdejo. Overview of the Answer Validation Exercise 2008. In C. Peters. Th. Mandl. V. Petras. A. Peñas. H. Müller. D. Oard. V. Jijkoun. D. Santos (Eds). Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum. CLEF 2008. Aarhus. Denmark. September 17-19. 2008. Revised Selected Papers.

4. Ido Dagan. Oren Glickman. and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. In Lecture Notes in Computer Science. volume 3944. pages 177–190. Springer. 2005.

5. Oren Etzioni. Michele Banko. and Michael J. Cafarella. Machine reading. In Proceedings of the 21st National Conference on Artificial Intelligence. 2006

6. Ellen M. Voorhees and Dawn M. Tice. The TREC-8 Question Answering Track Evaluation. In Text Retrieval Conference TREC-8. pages 83–105. 1999.

7. B. Wellner. L. Ferro. W. Greiff and L. Hirschman. Reading Comprehension Tests for Computer-based Understanding Evaluation. Nat. Lang. Eng. 12. 4. 305-334. 2006

8. Anselmo Peñas. Pamela Forner. Richard Sutcliffe. Álvaro Rodrigo. Corina Forascu. Iñaki Alegria. Danilo Giampiccolo. Nicolas Moreau. 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 Vol. I Text Retrieval Experiments*. Workshop of the Cross-Language Evaluation Forum. CLEF 2009. Corfu. Greece. 30 September - 2 October. Revised Selected Papers. Lecture Notes in Computer Science 6241. Springer-Verlag. 2010.

9. 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.

10. David Ferrucci. Eric Brown. Jennifer Chu-Carroll. James Fan. David Gondek. Aditya A. Kalyanpur. Adam Lally. J. William Murdock. Eric Nyberg. John Prager. Nico Schlaefer. and Chris Welty. 2010. Building Watson: An Overview of the DeepQA Project. AI Magazine. 31(3).

11. Roser Morante and Walter Daelemans. Annotating Modality and Negation for a Machine Reading Evaluation. CLEF 2011 Labs and Workshop - Notebook Papers. 19-22 September, Amsterdam, The Netherlands. Online Proceedings.

12. Juan Martinez-Romo and Lourdes Araujo. Graph-based Word Clustering Applied to Question Answering and Reading Comprehension Tests. CLEF 2011 Labs and Workshop - Notebook Papers. 19-22 September, Amsterdam, The Netherlands. Online Proceedings.

13. Adrian Iftene, Alexandru-Lucian Gînscă, Alex Moruz, Diana Trandabăt, Maria Husarciuc. Question Answering for Machine Reading Evaluation on Romanian and English. CLEF 2011 Labs and Workshop - Notebook Papers. 19-22 September, Amsterdam, The Netherlands. Online Proceedings.

14. Gaurav Arora. Cosine similarity as Machine Reading Technique. Question Answering for Machine Reading Evaluation on Romanian and English. CLEF 2011 Labs and Workshop - Notebook Papers. 19-22 September, Amsterdam, The Netherlands. Online Proceedings.

15. Suzan Verberne. Retrieval-based Question Answering for Machine Reading Evaluation. CLEF 2011 Labs and Workshop - Notebook Papers. 19-22 September, Amsterdam, The Netherlands. Online Proceedings.

16. Partha Pakray, Pinaki Bhaskar, Somnath Banerjee, Bidhan Chandra Pal, Alexander Gelbukh and Sivaji Bandyopadhyay. JU_CSE_TE: System Description QA4MRE@CLEF 2011. CLEF 2011 Labs and Workshop - Notebook Papers. 19-22 September, Amsterdam, The Netherlands. Online Proceedings.

17. Svitlana Babych, Alexander Henn, Jan Pawellek, and Sebastian Padò. Dependency-Based Answer Validation for German. CLEF 2011 Labs and Workshop - Notebook Papers. 19-22 September, Amsterdam, The Netherlands. Online Proceedings.

18. Ling Cao, Xipeng Qiu and Xuanjing Huang. Question Answering for Machine Reading with Lexical Chain. CLEF 2011 Labs and Workshop - Notebook Papers. 19-22 September, Amsterdam, The Netherlands. Online Proceedings.

19. Ingo Glockner, Bjorn Pelzer, and Tiansi Dong. The LogAnswer Project at QA4MRE 2011. CLEF 2011 Labs and Workshop - Notebook Papers. 19-22 September, Amsterdam, The Netherlands. Online Proceedings.

APPENDIX 1: Overall results at reading test level: Median, Average, and Standard Deviation for all runs

RUN_NAME	Overall Median	Overall Average	Overall Standard Deviation	Topic 1			Topic 2			Topic 3		
				Median	Average	Standard Deviation	Median	Average	Standard Deviation	Median	Average	Standard Deviation
base1101enen	0.27	0.26	0.17	0.22	0.23	0.10	0.22	0.21	0.16	0.44	0.35	0.24
diue1101enen	0.13	0.16	0.15	0.00	0.00	0.00	0.26	0.28	0.15	0.19	0.20	0.08
diue1102enen	0.15	0.2	0.14	0.14	0.12	0.08	0.27	0.31	0.17	0.14	0.17	0.08
fdcs1102enen	0.25	0.27	0.19	0.27	0.27	0.23	0.15	0.20	0.10	0.43	0.34	0.24
fdcs1103enen	0.22	0.25	0.17	0.24	0.27	0.07	0.14	0.17	0.06	0.34	0.33	0.29
ifln1101enen	0.38	0.32	0.20	0.25	0.26	0.15	0.31	0.28	0.22	0.41	0.42	0.23
ifln1102enen	0.35	0.37	0.16	0.23	0.28	0.11	0.45	0.45	0.24	0.40	0.37	0.10
ifln1104enen	0.27	0.31	0.18	0.21	0.23	0.11	0.24	0.29	0.16	0.41	0.42	0.23
ifln1105enen	0.35	0.35	0.13	0.23	0.28	0.11	0.40	0.40	0.18	0.40	0.37	0.10
iles1101enen	0.00	0.07	0.12	0.10	0.10	0.11	0.09	0.13	0.16	0.00	0.00	0.00
iles1102enen	0.09	0.12	0.17	0.09	0.20	0.28	0.17	0.13	0.09	0.00	0.05	0.09
iles1103enen	0.15	0.18	0.14	0.20	0.18	0.05	0.10	0.08	0.05	0.30	0.30	0.18
iles1104enen	0.20	0.17	0.11	0.20	0.18	0.05	0.05	0.08	0.10	0.20	0.25	0.10
iles1105enen	0.20	0.17	0.11	0.20	0.18	0.05	0.05	0.08	0.10	0.20	0.25	0.10
iles1106enen	0.15	0.18	0.16	0.21	0.25	0.24	0.16	0.16	0.02	0.08	0.12	0.15
iles1107enen	0.20	0.22	0.14	0.20	0.23	0.13	0.10	0.13	0.05	0.30	0.33	0.15
iles1108enen	0.20	0.24	0.12	0.25	0.25	0.13	0.10	0.13	0.05	0.37	0.33	0.09
iles1109enen	0.20	0.20	0.10	0.20	0.23	0.13	0.10	0.13	0.05	0.20	0.25	0.10
iles1110enen	0.20	0.22	0.10	0.25	0.25	0.13	0.10	0.15	0.10	0.25	0.25	0.06
jucs1101enen	0.15	0.16	0.13	0.15	0.15	0.06	0.15	0.18	0.17	0.15	0.15	0.17
jucs1102enen	0.15	0.16	0.13	0.15	0.15	0.06	0.15	0.18	0.17	0.15	0.15	0.17
jucs1103enen	0.20	0.21	0.12	0.15	0.18	0.10	0.20	0.25	0.17	0.20	0.20	0.08
jucs1104enen	0.30	0.32	0.15	0.25	0.30	0.14	0.20	0.25	0.19	0.40	0.40	0.08
jucs1105enen	0.30	0.32	0.15	0.25	0.30	0.14	0.20	0.25	0.19	0.40	0.40	0.08
jucs1106enen	0.74	0.58	0.37	0.81	0.80	0.18	0.68	0.53	0.36	0.39	0.42	0.49
jucs1107enen	0.45	0.48	0.28	0.81	0.80	0.18	0.20	0.25	0.19	0.40	0.40	0.08
loga1101dede	0.14	0.21	0.17	0.18	0.20	0.18	0.23	0.24	0.12	0.13	0.19	0.23
loga1102dede	0.14	0.21	0.17	0.18	0.20	0.18	0.23	0.24	0.12	0.13	0.19	0.23
swai1101enen	0.22	0.25	0.19	0.20	0.24	0.13	0.11	0.21	0.28	0.26	0.29	0.17
swai1102enen	0.00	0.06	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.17	0.01
swai1103enen	0.00	0.02	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.10
swai1104enen	0.00	0.08	0.11	0.00	0.05	0.09	0.00	0.00	0.00	0.17	0.20	0.08
swai1105enen	0.16	0.16	0.14	0.00	0.04	0.08	0.16	0.15	0.11	0.31	0.28	0.13

uaic1101roro	0.20	0.23	0.17	0.12	0.11	0.08	0.35	0.36	0.15	0.27	0.23	0.18
uaic1102roro	0.00	0.13	0.19	0.00	0.00	0.00	0.10	0.13	0.16	0.24	0.26	0.24
uaic1103roro	0.16	0.21	0.20	0.00	0.05	0.10	0.35	0.33	0.16	0.26	0.26	0.24
uaic1104roro	0.13	0.19	0.22	0.00	0.05	0.10	0.22	0.22	0.20	0.33	0.31	0.29
uaic1105roro	0.00	0.13	0.21	0.00	0.00	0.00	0.00	0.08	0.16	0.32	0.30	0.26
uaic1106roro	0.13	0.19	0.22	0.00	0.05	0.10	0.22	0.22	0.20	0.33	0.31	0.29
uaic1107roro	0.20	0.26	0.20	0.10	0.10	0.08	0.40	0.39	0.14	0.30	0.29	0.24
uaic1108roro	0.00	0.14	0.20	0.00	0.04	0.08	0.00	0.08	0.16	0.32	0.30	0.26
uaic1109roro	0.16	0.22	0.21	0.07	0.08	0.10	0.29	0.26	0.19	0.33	0.31	0.29
uaic1110enen	0.31	0.28	0.11	0.23	0.25	0.12	0.31	0.27	0.09	0.36	0.32	0.14
uhei1101dede	0.16	0.13	0.13	0.17	0.16	0.13	0.09	0.13	0.16	0.09	0.09	0.10
uhei1102dede	0.22	0.23	0.15	0.22	0.18	0.12	0.39	0.34	0.14	0.16	0.16	0.13
uhei1103dede	0.23	0.21	0.13	0.23	0.18	0.13	0.32	0.30	0.13	0.16	0.16	0.13
uhei1104dede	0.16	0.17	0.13	0.08	0.10	0.13	0.31	0.28	0.10	0.16	0.12	0.08
uhei1105dede	0.17	0.16	0.13	0.08	0.10	0.13	0.27	0.22	0.15	0.17	0.16	0.12
uhei1106dede	0.17	0.18	0.12	0.17	0.18	0.15	0.26	0.25	0.10	0.17	0.13	0.08
uhei1107dede	0.17	0.16	0.12	0.18	0.18	0.15	0.18	0.22	0.10	0.09	0.09	0.10
uhei1108dede	0.18	0.18	0.12	0.18	0.16	0.12	0.22	0.19	0.14	0.22	0.19	0.14
uhei1109dede	0.23	0.23	0.13	0.21	0.17	0.12	0.28	0.30	0.09	0.14	0.21	0.15
uned1101enen	0.27	0.27	0.13	0.19	0.19	0.08	0.35	0.36	0.11	0.23	0.26	0.14
uned1102enen	0.24	0.21	0.16	0.23	0.20	0.16	0.06	0.14	0.21	0.32	0.28	0.09
uned1103enen	0.17	0.19	0.18	0.24	0.23	0.09	0.00	0.11	0.21	0.21	0.23	0.22
uned1104enen	0.16	0.15	0.12	0.09	0.12	0.15	0.16	0.12	0.08	0.23	0.20	0.14
uned1105enen	0.16	0.15	0.16	0.08	0.09	0.10	0.15	0.12	0.08	0.23	0.26	0.24
uned1106enen	0.14	0.13	0.15	0.09	0.13	0.16	0.07	0.07	0.08	0.16	0.20	0.20
uned1107enen	0.16	0.13	0.11	0.09	0.09	0.10	0.08	0.08	0.09	0.21	0.23	0.08
uned1108enen	0.00	0.03	0.07	0.00	0.05	0.10	0.00	0.00	0.00	0.00	0.04	0.09
uned1109enen	0.21	0.23	0.18	0.20	0.21	0.07	0.41	0.33	0.23	0.09	0.16	0.21
vens1101enen	0.11	0.18	0.17	0.11	0.17	0.19	0.06	0.06	0.06	0.28	0.31	0.12

APPENDIX 2: SYSTEM DESCRIPTIONS

Table 17: Methods used by participating systems

System name	Question Analyses				Linguistic Processing																
	No Question Analyses	Manually done Patterns	Automatically acquired patterns	Other	Part Of Speech Tagging	Chunking	n-grams	Named Entity Recognition (NER)	Temporal expressions	Numerical expressions	Phrase transformations	Dependency analysis	Functions (sub. obj. etc)	Syntactic transformations	Semantic parsing	Semantic role labeling	Predefined Sets Of Relation	Frames	logic representation	Theorem prover	None	Other
base			x		x			x				x	x			x						
diue		x				x		x			x											
fdcs	x				x			x			x	x										
ifln	x											x										factoid extract.
iles		x			x			x				x	x	x								
jucs			x				x															Stem
loga		x						x											x	x		
swai	x				x																	
uaic		x						x	x	x				x								
uhei	x				x							x	x	x								
ued	x			Graph Analysis	x	x						x							x			
vens		x			x	x		x	x	x	x	x	x		x	x		x	x			

Table 18: Use of Knowledge by participating systems

System name	captured from the background collection	Knowledge Resources Used													Tools				
		Lexical DB	Thesaurus	Encyclopedia	Ontology	Collection of paraphrases	Word List	Gazzetteers	Categorial-Variation DB	Synonym-Acronym Dictionary	Dependency Similarity Dictionary	Proximity Similarity	Lexical Reference Rule-Base	Collection of word Knowledge propositions	Collection of entailment rules	Coreference Resolver	Named Entities Recognition	POS Tagger	Parser	Name Normalization
base	x																x	x		
diue	x																x			
fdcs		x														x	x	x	x	x
ifln	x																		x	
iles		x	x														x	x	x	
jucs															x		x			
loga		x	x		x										x	x	x		x	
swai	x																x			
uaic										x							x		x	
uhei	x																		x	
ued	x																	x	x	
vens		x														x	x	x	x	x

Table 19: Techniques used for the Answer Validation component

System name	No answer validation	Machine Learning	Web redundancies	Redundancies in the collection	Lexical similarity (term overlapping)	Syntactic similarity	Sematic similarity	Theorem prooving or similar	Other
base	x								
diue						x			
fdcs					x	x	x		
ifln				x					
iles	x								
jucs					x				
loga								x	
swai					x				
uaic					x				
uhei					x	x		x	
uned	x								
vens	x								
base									