Overview of QA4MRE at CLEF 2012: Question Answering for Machine Reading Evaluation

Anselmo Peñas[1], Eduard Hovy[2], Pamela Forner[3], Álvaro Rodrigo[4], Richard Sutcliffe[5], Caroline Sporleder[6], Corina Forascu[7], Yassine Benajiba[8], Petya Osenova[9]

[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] Saarland University, Germany (csporled@coli.uni-sb.de)
[6] Al. I. Cuza University of Iasi, Romania (corinfor@info.uaic.ro)
[8] Philips Research North America, USA (Yassine.Benajiba@philips.com)
[9] Bulgarian Academy of Sciences, Bulgaria (petya@bultreebank.org)

Abstract. This paper describes the Question Answering for Machine Reading (QA4MRE) task at the 2012 Cross Language Evaluation Forum. In the main task, systems answered multiple-choice questions on documents concerned with four different topics. There were also two pilot tasks, Processing Modality and Negation for Machine Reading, and Machine Reading on Biomedical Texts about Alzheimer's disease. This paper describes the preparation of the data sets, the definition of the background collections, the metric used for the evaluation of the systems' submissions, and the results. Eleven groups participated in the task submitting a total of 43 runs in seven languages.

1. INTRODUCTION

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

When reading a text, a human performs two processes, namely:

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

To assess the degree and types of understanding, we have the system answer questions about a given text. While the desired answer is usually also present in the test document (albeit perhaps in some non-obvious form), it may not be, or the reader may require additional background information to know what to search for, such as explicit and implicit references to entities, events, dates, places, situations, etc. pertaining to the topic.

In general, more prior background knowledge makes understanding and question answering easier. Computational resources such as wordnets, framenets, paraphrase lists, knowledge bases, etc., are aimed at making different kinds of prior knowledge available for the machine. In QA4MRE we add to these resources the possibility to acquire background knowledge from a large collection of related documents. The advantage is the opportunity to gather probability distributions linked to knowledge, and to explore distributional approaches to QA. We discuss background knowledge in Section 3.

The evaluation questions should be answerable by most humans without the need to explore a specific document of the background collection. Examples of inferences we allow are:

1. Linguistic inferences such as coreference, deictic references (like "then" and "here"), etc.);
2. Simple ontological inferences such as considering part-of relations or obtaining direct super-concepts for common objects;
3. Inferences considering causal relations or procedural steps in "life scripts" like visiting a restaurant or attending a concert;

4. Inferences that require composing several answers, in particular answering one part of the question using the background collection and then, with its answer, answering the other part of the initial question (e.g., "Who is the wife of the person who won the Nobel Peace Prize in 1992?").

2. TASK DESCRIPTION

In 2012, we had a main task and two pilot exercises.

Main Task This remained the same for participants. Background collections, test documents and reading tests were available in Arabic, Bulgarian, English, German, Italian, Romanian, and Spanish. In addition to last year's topics (AIDS, Climate Change, Music and Society), we included a topic on Alzheimer's disease. This new topic is related to a new pilot on Biomedical texts. The difference is that the reference collection for the main task is built from general public sources and for the pilot the source is the PubMed repository.

Having these two parallel exercises about the same topic but in different domains opens the door to evaluate research approaching the challenges of domain and language adaptation, the use of knowledge in one domain captured in the other, the differences in the background knowledge acquired, the differences between questions and answers in each domain, etc.

Pilot on Processing Modality and Negation for Machine Reading. This exercise is aimed at evaluating whether systems are able to understand extra-propositional aspects of meaning like modality and negation. Modality is a grammatical category for expressing the attitude of the speaker towards his/her statements, such as expressions of certainty, factuality, and evidentiality. Negation is a grammatical category that allows changing the truth value of a proposition. In the pilot, participants received some texs where they have to decide whether some events are Asserted, Negated, or Speculated. Our plan is to integrate modality and negation in the main task next year.

Machine Reading on Biomedical Texts about Alzheimer's disease. This exercise is aimed at setting questions in the Biomedical domain with a special focus on one disease, namely Alzheimer's. This pilot task explored the ability of a system to answer questions using scientific language. Texts were taken from PubMed Central related to Alzheimer's and from 66,222 Medline abstracts. In order to keep the task reasonably simple for systems, participants were given the background collection already processed with Tok, Lem, POS, NER, and dependency parsing.

The two pilot tasks are described in detail in dedicated papers in these proceedings.

2.1 Main Task

Tests were divided into:

- 4 topics, namely "Aids", "Climate change", "Music and Society" and "Alzheimer";
- Each topic had four reading tests;
- Each reading test consisted of one single document, with 10 questions and a set of five choices per question.

Overall, the following evaluation setting was proposed:

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

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

3. THE BACKGROUND COLLECTIONS

This is a very important element of the evaluation setting. It connects the task also with the research in Information Retrieval. The goal of reference/background collections is to contextualize the reading of a single

document related to the topic by collecting and fleshing out additional pertinent information. In the future this step may be done on the fly as a retrieval process once a single test text is provided. However, for now, we provide a carefully constructed background corpus for two main reasons: to allow more comparison among participant systems, and to focus on the Reading Comprehension problem. We believe it is important to develop a good methodology for building background collections for the evaluation task.

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 the 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.

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 rest of the 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 in the past edition 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 don't cover completely the topic, but also that the collections have some bias with respect to the real distribution of concepts. In this year's campaign, the assumption that the ideal background collection should include all relevant documents for the topic (and only them) is explicit, and we organizers 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 "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.

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

TOPICS	AR	BG	DE	EN	ES	IT	RO
	# docs *KB*	*# docs* *KB*	*# docs* *KB*	*# docs* *KB*	*# docs* *KB*	*# docs* *KB*	*# docs* *KB*
ALZHEIMER	19,278 docs 173,951 KB	19,412 docs 194,326 KB	18,506 docs 146,965KB	13,045 docs 254,924 KB	6,199 docs 42,899 KB	9,008 docs 60,819 KB	9,590 docs 121,413 KB

AIDS	8,790 docs 120,620 KB	17,102 docs 123,636 KB	10,399 docs 144,204 KB	12,280 docs 199,233 KB	6,344 docs 66,908 KB	3,690 docs 17,564 KB	3,793 docs 47,120 KB
CLIMATE CHANGE	10,151 docs 199,846 KB	32,459 docs 192,095 KB	6,501 docs 49,238 KB	13,424 docs 184,925 KB	5,185 docs 33,063 KB	3,839 docs 22,444 KB	6,035 docs 43,983 KB
MUSIC & SOCIETY	15,725 docs 265,546KB	24,585 docs 281,587 KB	6,639 docs 80,194 KB	7,785 docs 135,747 KB	4,628 docs 34.773 KB	3,525 docs 30,349 KB	3,571 docs 26,946 KB

Table 1 shows information about the background collections. Collections marked in violet are the extensions.

Next Table shows the keywords used for each topic. They are a sort of more concrete definition of each topic, giving an idea of the subtopics covered by the collection.

ALZHEIMER KEYWORDS	CLIMATE CHANGE KEYWORDS (EXTENSION)
Alzheimer's AND Alzheimer's disease	solar radiation
Alzheimer's drugs	carbon capture
Alzheimer's symptoms	fluorinated gases
Alzheimer's treatment	drought
Alzheimer's causes	heat-trapping gases
senile dementia	ground-Level ozone
memory loss	wind power
(memory testing OR neuropsychological tests) for Alzheimer	biofuel
brain disorder AND neurological disorder	gas emissions
plaques and tangles	biomass
Lewy bodies	AIDS KEYWORDS (EXTENSION)
mental confusion AND Alzheimer	
wandering AND Alzheimer	HIV/AIDS funding
irritability AND Alzheimer	AIDS global crisis
sundowning	TRIPS Agreement
depression AND Alzheimer	AIDS pharmaceutical industry
(language problems OR aphasia) AND Alzheimer	World Health Organization
(perception problems OR agnosia) AND Alzheimer	AIDS family planning
(disorder of motor planning OR apraxia) AND Alzheimer	AIDS pandemic
	AIDS life expectancy rate
personality changes AND Alzheimer	fighting AIDS
beta-amyloid	AIDS virology
(caregiving OR long-term care) AND Alzheimer	
nursing home AND Alzheimer	MUSIC AND SOCIETY KEYWORDS (EXTENSION)
(aging society OR geriatrics) AND Alzheimer	
healthcare costs AND Alzheimer	
cognitive reserve theory	music criticism
Auguste Deter	musicology
Danae Chambers	history of violin technique
Alzheimer's Association	music patronage
Alzheimer diagnosis	rock and roll
Alzheimers' associated disorders	history of song
Alzheimers' clinical features	electric musical instrument
Alzheimers' genetics	classical recording industry
Alzheimers' prevention	economics of classical music
Familial Alzheimer's	classical crossover music
Alzheimers' risk factors	
impact of Alzheimer's disease	
Neuropathology of Alzheimer's Disease	

4. TEST SET PREPARATION

This year the datasets was created for the following seven languages: Arabic, Bulgarian, English, German, Italian, Romanian and Spanish. The dataset was created following the methodology developed last year consisting of the following steps:

1. Four English documents were selected for each of the four topics (Aids, Climate Change, Music and Society, Alzheimer's). These were selected from copyright-free sources (see Table 2) and these represented the test documents against which questions were asked.

2. In order to have a set of identical questions for the seven languages above, we needed to have the selected test documents translated. For this purpose, expert translators were recruited form the Translation for Progress[1] platform for all languages. On the whole, 57 translators were contacted and asked to perform the translations in a couple of weeks' time. Most of the translations were of a high quality and were delivered within the agreed timescale.

3. To ensure that translations were faithful to the original document in both meaning and style and of good quality, all the documents were 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. Ten multiple-choice questions were then devised for each test document. A question always had five candidate answers from which to choose, with one clearly correct answer and four clearly incorrect answers.

5. 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.

6. The English versions were then used to translate each question into each of the seven languages of the task. The same process was used to translate each candidate answer (five per query) into the seven languages.

7. The result of this process was a set of 160 questions in seven languages, each with five multiple-choice answers, also in those seven 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.

Table 2: Test Documents

Topic	No.	Source	Author	Title	LICENSE	Words
AIDS	1	http://www.fpif.org/articles/the_dis-integration_of_us_global_aids_funding	Jodi L. Jacobson	"The Dis-Integration of U.S. Global AIDS Funding" (Washington, DC: Foreign Policy In Focus, March 3, 2003)	Creative commons Attribution	1350

[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.

AIDS	2	http://archive.icommons.org/articles/pipeline-patents-compulsory-licensing-and-the-costs-of-aids-treatment-in-brazil	Paula Martini	Pipeline patents, compulsory licensing and the costs of AIDS treatment in Brazil	Creative Commons Attribution	1147
AIDS	3	http://www.fpif.org/reports/hivaids_in_africa_time_to_stop_the_killing_fields	Chinua Akukwe and Melvin Foote,	"HIV/AIDS in Africa: Time to Stop the Killing Fields" (Washington, DC: Foreign Policy In Focus, October 6, 2005)	Creative Commons Attribution	2520
AIDS	4	http://www.fpif.org/articles/african_women_confront_bushs_aids_policy	Yifat Susskind	"African Women Confront Bush's AIDS Policy" (Washington, DC: Foreign Policy In Focus, December 2, 2005)	Creative Commons Attribution	1315
Climate Change	5	http://chevyvolt.cm.fmpub.net/#http://boingboing.net/2011/08/05/3-things-you-need-to-know-about-biofuels.html	Maggie Koerth-Baker	3 things you need to know about biofuels	Creative Commons Attribution Non-Commercial	2059
Climate Change	6	http://www.scidev.net/en/policy-briefs/brazil-climate-change-a-country-profile.html	Emilio Lèbre La Rovere and André Santos Pereira	Brazil & climate change: a country profile	Creative Commons Attribution	2300
Climate Change	7	http://www.energybulletin.net/node/51370	Maude Barlow	To curb climate change, we need to protect water	Creative Commons	1190
Climate Change	8	http://www.scidev.net/en/climate-change-and-energy/biofuels/opinions/reality-check-for-miracle-biofuel-crop.html	Miyuki Iiyama and James Onchieku	Reality check for 'miracle' biofuel crop	Creative Commons Attribution	1025
Music & Society	9	http://www.archive.org/stream/encyclopaediabri04chisrich/encyclopaediabri04chisrich_djvu.txt	Chisholm, Hugh (Ed.)	Charles Burney	Public Domain	1335
Music & Society	10	http://www.bos.frb.org/economic/nerr/rr2003/q2/requiem.htm	Julie Lee	Requiem for Classical Music	Reproduction of any information contained herein may be made without limitation as to number, provided that it is not distributed for the purpose of private gain and is appropriately credited to the Federal Reserve Bank of Boston	2656
Music & Society	11	http://en.wikipedia.org/wiki/Pop_music	Unknown	Pop Music	Public Domain	1373

Music & Society	12	http://www.gutenberg.org/files/14884/14884-h/14884-h.htm#page31	Henry C. Lahee	Famous Violinists of To-Day and Yesterday	Public Domain	1826
Alzheimer	13	http://knol.google.com/k/lara/alzheimer-s-disease/Ing3X-NE/g1JpHQ#	Bruce Miller; Lara Heflin,	Alzheimer's Disease	Creative Commons Attribution	3509
Alzheimer	14	http://knol.google.com/k/gloria-h-schneider/creativity-alzheimer-s-disease/1v6cy64kp9uk1/78#	Gloria Ha'o Schneider	Creativity & Alzheimer's Disease	Creative Commons Attribution	954
Alzheimer	15	http://knol.google.com/k/elder-care-elder-rage-know-the-warning-signs-of-alzheimer-s	Jacqueline Marcell	Caring for Aging Parents & Elder Rage: Know The Warning Signs of Alzheimer's!	Creative Commons Attribution	1731
Alzheimer	16	http://knol.google.com/k/stan-goldberg/it-s-only-alzheimer-s-not-the-bloody/32wlgicpxht73/5#	Stan Goldberg	It's Only Alzheimer's, Not the Bloody Plague!	Creative Commons Attribution	1079

4.1 Questions

For each text in the test set 10 multiple choice questions were created. Each question had five answer options. The 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. The exact breakdown of the number of questions per type in the test collection is provided in Table 3 below. Example questions:

PURPOSE: What is the aim of Obama‘s cap-and-trade policy?

METHOD: How could vast quantities of petrol be saved?

CAUSAL: What is the reason for the high price of solar energy?

FACTOID (time): When are bioethanol and biodiesel expected to become widely used?

WHICH-IS-TRUE: Which of the following goals is Europe committed to?

Table 3: Distribution of question types

Question type	**Total number of questions**
PURPOSE	27
METHOD	30
CAUSAL	36
FACTOID*	36
WHICH-IS-TRUE	31
TOTAL # of QUESTIONS	160

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.

Table 4 below shows a classification of the questions according to how much and what type of background knowledge they required. The table also provides the average c@1 obtained for each type of question. It can be seen that, unsurprisingly, the types of questions that require little knowledge and inference are generally answered more successfully. Questions requiring inference are by far the hardest, while it does not seem to make much difference whether the knowledge required is found within the test document or in the background collection.

Table 4: Classification of questions according to the knowledge required to answer them

Types of question	#of questions	c@1
NO EXTRA KNOWLEDGE REQUIRED	75	0.30
BACKGROUND KNOWLEDGE REQUIRED	46	0.28
INFERENCE REQUIRED	21	0.20
INFORMATION NEEDS TO BE GATHRED FROM DIFFERENT SENTENCES or PARAGRAPHS	20	0.27

4.2 Tools and Infrastructure

This year, CELCT developed a series of infrastructure components to help manage 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 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. The main characteristic of these interfaces is 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 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. Alteration 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. Real time calculation of statistics and evaluation measures.

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 Gold Standard files, in the Test Set files, etc.

5. EVALUATION

Since one of the objectives of the task is to assess the ability of systems to understand texts through their answers to questions about those texts, the evaluation focuses on measuring this understanding by computing the correctness of the responses given to the multiple-choice tests. Furthermore, we follow the line introduced in ResPubliQA 2009 [1] of promoting the development of systems able to reason about the correctness of their responses with the aim of reducing the amount of incorrect answers given as output. Thus, this year's evaluation remains quite similar to the one of the last edition.

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 is given when the system considers it is not able to find enough evidence about the correctness of candidate answers and it prefers not to answer the question instead of giving an incorrect answer. Thus, it gains some partial credit proportional to the performance shown with the answered questions. Moreover, the system can return as a hypothetical answer the candidate one that it would have been selected, which allows us 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 with human-made annotations. Therefore, no manual assessment was required, which reduces the effort of the evaluation once the collections have been created and facilitates the future development of systems. Each system's response to a question 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.

Given these assessments, we decided to evaluate systems from two different perspectives:

1. A question-answering approach, as in the traditional evaluation performed in past campaigns, where we just evaluate the ability of systems answering a set of questions.
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.

5.1 Evaluation Measure

We use *c@1* as the main evaluation measure for this year's campaign. *c@1* was first introduced in ResPubliQA 2009 [1] and is fully described in [2]. 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

c@1 acknowledges returning *NoA* answers 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 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.

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

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

In the Question Answering perspective we measure systems' performance over a set of questions without considering the ability of a system to understand a certain document. This is an approach similar to the one applied in QA@CLEF campaigns before 2010.

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.

This information is used for calculating the following scores:

- An overall *c@1* over the whole collection (a set of 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 multiple-choice tests with ten questions per document.

This evaluation is performed taking as reference the *c@1* scores achieved for each test (one document with its ten questions). Afterwards, 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* has been provided under the consideration that it can be more informative at reading 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.

This approach allows us to evaluate systems in a similar way to the manner new language learners are graded. Thus, we can consider that a system passes a test from this evaluation perspective if it achieves a score equal or higher than 0.5. In the case of obtaining an overall average *c@1* higher than 0.5, we say that the system passes this evaluation perspective.

5.4 Random Baselines

We propose here a simple baseline to which participants can be compared. Since participant systems can decide to answer or not to answer a given question, we must decide which behaviour must follow our baseline. For simplification purposes, the proposed baseline answers all the questions, randomly selecting each answer from the set of candidate ones.

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. Then, the overall result of this random baseline is 0.2 (both for *accuracy* and for *c@1*). Systems applying a certain kind of processing and reasoning should be able to outperform this baseline.

6. PARTICIPATION and RESULTS

From an initial amount of 25 groups that registered to the main task and signed the license agreement to download the background collections, 11 of them finally submitted at least one run, resulting in 43 runs in 7 languages (Arabian, Bulgarian, German, English, Spanish, Italian and Romanian). Table 7 shows the number of runs per language.

There were only 3 cross-lingual runs and all from the same group. The language with the highest amount of runs was, as usual, English with 20 submissions, while Spanish and Italian received only one run per language. Thus, no comparison in these two languages can be performed.

Tables 5 to 7 summarise the characteristics of the submissions.

Table 5: Overall participants and runs in QA4MRE tasks

REGISTERED PARTICIPANTS	PARTICIPANTS DOWNLOADING THE TEST SETS	PARTICIPANTS SUBMITTING RUNS	TOTAL NUMBER OF RUNS
38	24	21	88

Table 6: Participants and runs per tasks

NUMBER of PARTICIPANTS		NUMBER of RUNS	88
MAIN	11	MAIN	43
BIOMEDICAL about ALZHEIMER	7	BIOMEDICAL about ALZHEIMER	42
MODALITY AND NEGATION	3	MODALITY AND NEGATION	3 zip

Table 7: Runs submitted per language in the QA4MRE Main Task

		Target languages (corpus and answer)							
		AR	BG	DE	EN	ES	IT	RO	**Total**
Source langs (questns)	AR	4							**4**
	BG		5						**5**
	DE			3					**3**
	EN				20				**20**
	ES					**1**			**1**
	IT						1		**1**
	RO		1		1		1	6	**9**
	Total	**4**	**6**	**3**	**21**	**1**	**2**	**6**	**43**

Table 8 below shows the percentage of correct and NoA answers for different question types. Percentages of correct answers overall for Purpose, Factoid and Which-is-true are very similar at around 25%. Method and Causal are not much lower at 22.24% and 20.86%. NoA scores are similar over different question types; the highest is Which-is-true at 17.32% and the lowest is Method at 15.56%. Hence, while Method and Causal might be a bit more difficult than the other question types, possibly due to the fact that they tend to require more inference, overall the question types were quite balanced with respect to difficulty.

Table 8: Percentage of Correct and NoA answers according to different question type

Question type	**% of correct answers**	**% of NoA answers**
PURPOSE	25.23%	17.14%
METHOD	22.24%	15.56%
CAUSAL	20.86%	17.70%
FACTOID*	25.25%	16.79%
WHICH-IS-TRUE	25.28%	17.32%

Table 9 shows the average results for each one of the proposed 16 reading comprehension tests according to *c@1*. The Table shows that, except for Test 8, the mean value was higher than the proposed baseline, while only half of them were higher in the previous edition [2].

The mean values for all the tests where under 0.5, the value needed to pass the evaluation from the reading perspective. This result suggests that systems are still far away from obtaining satisfactory results according to this perspective.

Table 9: Mean Scores for each Reading Test

Topic 1				Topic 2				Topic 3				Topic 4			
Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Test 11	Test 12	Test 13	Test 14	Test 15	Test 16

[2] It must be mentioned that there were 12 tests in QA4MRE 2011

Average	0.34	0.28	0.31	0.33	0.29	0.27	0.21	0.16	0.24	0.26	0.29	0.34	0.28	0.31	0.33	0.29

We can see in Table 10 how results across tests in the same topic are more similar than in 2011, which suggests that this year's collections are more homogenous. On the other hand, Table 10 shows the mean scores per topic. The scores across topics seem remarkably similar except for Topic 1 (AIDS), which seems to be a bit easier.

Table 10: Mean Scores for each Topic

	Topic 1	Topic 2	Topic 3	Topic 4
Average	0.32	0.24	0.25	0.24

Table 11 shows the results of all the submitted runs grouped by language. It can be seen how most of the systems were able to beat the baseline (only 8 runs performed lower), with at least a system per language able to do so. This amount is higher than in the previous edition (where only half of the systems outperformed the baseline), which is further evidence of the improvement of systems' performance in this edition.
Systems were able to find answers for more than 36% of questions in all languages (combination row in Table 11) except Spanish where there was only one run (also cross-language runs have been considered).
Considering all languages, 99% of questions received at least one correct answer by at least one system.

Table 11: c@1 in participating systems according to the language

System name	AR	BG	DE	EN	ES	IT	RO
Combination	38.75	66.87	36.87	93.75		46.50	55.62
jucs12013enen				0.65			
vulc12014enen				0.40			
vulc12034enen				0.38			
onto12031itit						0.35	
vulc12024enen				0.35			
l2fi12041enen				0.34			
onto12081roro							0.34
l2fi12031enen				0.33			
diue12024enen				0.31			
onto12021enen				0.31			
onto12071bgbg		0.30					
btbn12011bgbg		0.29					
diue12012enen				0.29			
onto12011bgbg		0.28					
onto12091dede			0.28				
onto12101eses					0.28		
uaic12082enen				0.28			
loga12023dede			0.26				
uaic12092enen				0.26			
loga12011dede			0.25				
uaic12024roro							0.25
uaic12062enen				0.25			
uaic12042roro							0.24
uaic12014roro							0.23
uaic12034roro							0.23

uaic12102enen				0.23			
idrq12021arar	0.21						
l2fi12021enen				0.21			
uaic12072enen	0.21						
btbn12031bgbg		0.20					
fdcs12042enen				0.20			
l2fi12011enen				0.20			
baseline	0.20	0.20	0.20	0.20	0.20	0.20	0.20
fdcs12021enen				0.19			
mira12021arar	0.19						
uaic12052roro							0.19
mira12011arar	0.15						
fdcs12032enen				0.14			
idrq12011arar	0.13						
btbn12021bgbg		0.12					
fdcs12011enen	0.12						

The best results were obtained in English, where the highest score was obtained by *jucs12013*enen with 0.65. This value is 25 percentage points higher than the next system (*vulc12014enen* at 0.40). In fact, *jucs12013enen* was the only system able to pass the evaluation according to the reading perspective. This system obtained *c@1* values over 0.60 for all the topics except for Topic 2 (Climate change). We can consider it a very good result if we compare that system with a person over such complex questions.

Regarding cross language runs, all of them were from the *onto* group over different target languages with Romanian as source, which does not allow to make any comparison. All these runs obtained the same result (0.29 of *c@1*).

Table 12: c@1 in participating systems (cross-lingual) according to the language

System name	AR	BG	DE	EN	ES	IT	RO
onto12041robg		0.29					
onto12051roen				0.29			
onto12061roit						0.29	

Table 13 compares the performance of systems in these two editions of QA4MRE. There has been an overall improvement across all runs between last year and this year (0.21 increased to 0.26), as well as an improvement across best runs of each participant group (0.28 to 0.32).

Table 13: Average Scores over all runs and over best runs

	over all runs	over all best runs
QA4MRE 2012	0.26	0.32
QA4MRE 2011	0.21	0.28

6.1 Analysis of the Use of External Knowledge

This task tries also to promote the use and combination of external sources of knowledge in order to help answering questions as it has been said above. In order to study it, we asked participants to report the resources employed to assist in answering the questions and we summarise this information in Table 14.

Table 14: Categorisation of runs, depending on the resources used

Types of runs	#of runs	Average c@1
No external resource is used (only the test document)	24	0.24
Only the test document and the associated background collection are used	10	0.22
The test document and other resources are used, but not the background collection	2	0.45
The test document together with both the background collection and other resources are used	7	0.30
TOTAL of runs	**43**	**0.24**

53% of the submitted runs did not employ any kind of external resources, while 23% used only the background collection. The remainder of runs used additional resources, either with or without using the background collection. These observations suggest that the inclusion of such external sources and their exploitation is not yet widely adopted. Moreover, as shown in Table 14, more detailed information about the external sources used for each participant can be seen in Table 16 of Appendix 3.

A subsequent analysis of questions revels that questions requiring no extra knowledge were not much easier than the others. In fact, some of them seem to be considerably harder than some questions that require external resources. This observation suggests that in order to answer questions, the fact of having to compose two or more parts to form an answer is harder than just matching a single piece of text. However, whether the pieces of the answer are in the main text or in a background resource collection does not make much difference. It is more relevant for the performance how difficult the pieces are to match.

6.2 Analysis of Systems

Table 15 in Appendix 3 summarizes the set of techniques that participants have reported are being used in their systems. There was only one system that did not perform any kind of question analysis, while most of the other systems employed questions patterns, with a high proportion of systems acquiring them automatically.

Regarding the linguistic processing, the most popular techniques were PoS tagging, the use of NER tools and dependency parsers, which were also some of the most applied techniques in previous editions. However, very few systems explore the use of deeper techniques relying on semantics, while only one relied on logic representation and a theorem prover.

Those two systems applying the most different techniques (*jucs* and *idrq*) were the ones that best performed in their languages (English and Arabic respectively). However, system *vulc*, which performs very well in English, reported only the use of phrase transformations. Therefore, it does not seem to be very clear which is the best combination of techniques in order to obtain a good performance. Evaluation frameworks such as the one presented in this paper must be used by researchers for exploring and answering such questions.

7. CONCLUSIONS

While this year's results show some improvement compared to last year, specially respect to the respective baselines, the majority of systems are still far from being able to pass a Reading Comprehension test. Nevertheless, best systems are, in general, very close to achieve this goal.

When we defined the task we kept in mind three main ideas: that we are developing a validation technology able to determine if a particular answer is correct or not; that knowledge is crucial for understanding; and that a large set of documents related to a topic could be an additional source of background knowledge. We discuss each in turn:

The first question is whether the technology developed so far is just ranking the options or it is validating them. The difference is important: What happens if we don't provide the options? Most systems use a kind of similarity measure or they don't use validation at all. Thus, more than validating the answers, systems are ranking them. This leads to the need of a change for next campaign. Again, the option of gaining partial credit by leaving some questions unanswered and reduce the number of incorrect answers is not enough. We need to introduce an explicit assessment of the ability to reject candiate answers when they are incorrect. This could be done easily in our framework if we introduce a significant portion of questions where none of the options are correct and a last

option in all questions "None of the answers above are correct". If a significant portion of questions (up to 40%) have no correct answer among its options, this will give as a new baseline to beat: a dummy system that always chose there is no correct answer as default.

About the second and third issues, it seems that the use of external resources help to improve results, but this is not so clear in the case of background collections. Although we have refined the methodology to build the background collections this year, most participants don't seem to know how to gather usable background knowledge from it. Moreover, it seems that the use of other external resources benefit more than the use of the background collections. We need to decide on this issue because the organization is spending a lot of resources in creating theses collections[3].

ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministry of Science and Innovation, through the project Holopedia (TIN2010-21128-C02), and the Regional Government of Madrid, through the project MA2VICMR (S2009/TIC1542).

This work has been partially supported by the PROMISE Network of Excellence (258191).

Special thanks are due Giovanni Moretti (CELCT, Trento, Italy) for the technical support in the management of all data and evaluation scripts of the campaign.

We would also like to acknowledge the volunteer translators that contributed to the creation of the dataset: Mercedes Marta Moreno; Juan Manuel Pérez Rojas; Adriana Pedemonte; Sophie; Ana Casillas Tomasin; Mayra Alvarez; MARÍA SOL ACCOSSATO; Natalia Steckel; María Constanza Galli; Fiorela; Hanna; Yessica V. Apolo Martínez; Fatima Alvarez; Alberto Mengibar Martin; Taras Giovanna; Pamela Aikpa Gnaba; Danielle; Marco Menegazzi; Alfredo Lo Bello; Chiara S.; Martina Scarano; Nunzio; Francesca Rubino; Katia G; Lucia Zirattu; Camilla Cosmelli; Sara Colombo; Chiara Gavasso; Katie M; Antti; Saskia Scharnowski; Jeffrey Bunce; Gabriele Mark; Melanie Liebchen; Helena Knaup; Judith Müller; Kathrin Meier; Anika Abel; Eva Wagle-Fopp; Franziska Bioh; Irina Rata; Nadia Bucurenci; Daniela Arsinel; Cristina Manoli; luminita Isaic; Nicoleta Mihaita; Mohamed ElGohary; Dina Awadallah; Nawel; Amal; Sarah; Shameem; Rabie Mustapha; Difaf Sharba; Amani; Khebouri Amina; Manel Rada.

REFERENCES

1 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.

2. 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.

3. Hugo Rodrigues, Luísa Coheur, Ana Cristina Mendes, Ricardo Ribeiro, and David Martins De Matos. Testing Lexical Approaches in QA4MRE. In Proceedings of CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, 17-20 September, 2012, Rome, Italy

4. Cross-Language Answer Validation. Valentin Zhikov, Laura Tolosi, Petya Osenova, Kiril Simov, and Georgi Georgiev. In Proceedings of CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, 17-20 September, 2012, Rome, Italy

[3] Note this year we Google's API wasn't available for research purposes.

5. IDRAAQ: New Arabic Question Answering System Based on Query Expansion and Passage Retrieval. Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. In Proceedings of CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, 17-20 September, 2012, Rome, Italy

6. DI@UE in CLEF2012: Question Answering Approach to the Multiple Choice QA4MRE Challenge. José Saias and Paulo Quaresma. In Proceedings of CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, 17-20 September, 2012, Rome, Italy

7. Enhancing a Question Answering System with Textual Entailment for Machine Reading Evaluation. Adrian Iftene, Alexandru-Lucian Gînscă, Mihai Alex Moruz, Diana Trandabat, Maria Husarciuc, and Emanuela Boroş. In Proceedings of CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, 17-20 September, 2012, Rome, Italy

8. Bulgarian Question Answering for Machine Reading. Kiril Simov, Petya Osenova, Georgi Georgiev, Valentin Zhikov and Laura Tolosi. In Proceedings of CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, 17-20 September, 2012, Rome, Italy

9. The LogAnswer Project at QA4MRE 2012. Ingo Glöckner and Björn Pelzer In Proceedings of CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, 17-20 September, 2012, Rome, Italy

10. An Entailment-Based Approach to the QA4MRE Challenge. Peter Clark, Phil Harrison, and Xuchen Yao. In Proceedings of CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, 17-20 September, 2012, Rome, Italy

11. Arabic QA4MRE at CLEF 2012: Arabic Question Answering for Machine Reading Evaluation. Omar Trigui, Lamia Hadrich Belguith, Paolo Rosso, Hichem Ben Amor, and Bilel Gafsaoui. In Proceedings of CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, 17-20 September, 2012, Rome, Italy

12. Question Answering System for QA4MRE@CLEF 2012. Pinaki Bhaskar, Partha Pakray, Somnath Banerjee, Samadrita Banerjee, Sivaji Bandyopadhyay, and Alexander Gelbukh. In Proceedings of CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, 17-20 September, 2012, Rome, Italy

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

RUN_NAME	Overall c@1	c@1 topic_1	c@1 topic_2	c@1 topic_3	c@1 topic_4
btbn12011bgbg	0,29	0,25	0,25	0,35	0,33
btbn12021bgbg	0,12	0,10	0,03	0,18	0,18
btbn12031bgbg	0,20	0,23	0,15	0,23	0,20
diue12012enen	0,29	0,33	0,28	0,14	0,40
diue12024enen	0,31	0,40	0,25	0,20	0,40
fdcs12011enen	0,12	0,16	0,07	0,16	0,08
fdcs12021enen	0,19	0,30	0,14	0,18	0,12
fdcs12032enen	0,14	0,23	0,07	0,16	0,08
fdcs12042enen	0,20	0,30	0,14	0,18	0,12
idrq12011arar	0,13	0,25	0,18	0,05	0,05
idrq12021arar	0,21	0,36	0,19	0,08	0,17
jucs12013enen	0,65	0,77	0,33	0,64	0,76
l2fi12011enen	0,20	0,20	0,23	0,10	0,28
l2fi12021enen	0,21	0,28	0,16	0,30	0,11
l2fi12031enen	0,33	0,41	0,38	0,29	0,26
l2fi12041enen	0,34	0,44	0,35	0,32	0,28
loga12011dede	0,25	0,26	0,07	0,42	0,23
loga12023dede	0,26	0,29	0,11	0,42	0,23
mira12011arar	0,15	0,16	0,20	0,08	0,16
mira12021arar	0,19	0,23	0,25	0,15	0,15
onto12011bgbg	0,28	0,35	0,28	0,28	0,20
onto12021enen	0,31	0,43	0,25	0,33	0,25
onto12031itit	0,35	0,45	0,35	0,28	0,33
onto12041robg	0,29	0,23	0,30	0,38	0,25
onto12051roen	0,29	0,23	0,30	0,38	0,25
onto12061roit	0,29	0,23	0,30	0,38	0,25
onto12071bgbg	0,30	0,43	0,25	0,30	0,23
onto12081roro	0,34	0,40	0,28	0,35	0,35
onto12091dede	0,28	0,33	0,25	0,23	0,30
onto12101eses	0,28	0,33	0,30	0,30	0,20
uaic12014roro	0,23	0,18	0,27	0,20	0,26
uaic12024roro	0,25	0,28	0,27	0,20	0,26
uaic12034roro	0,23	0,26	0,28	0,18	0,21
uaic12042roro	0,24	0,25	0,26	0,18	0,26
uaic12052roro	0,19	0,16	0,29	0,08	0,21
uaic12062enen	0,25	0,28	0,25	0,24	0,23
uaic12072enen	0,21	0,29	0,24	0,18	0,08
uaic12082enen	0,28	0,34	0,25	0,27	0,23
uaic12092enen	0,26	0,33	0,24	0,25	0,20
uaic12102enen	0,23	0,33	0,20	0,18	0,19
vulc12014enen	0,40	0,55	0,40	0,40	0,25
vulc12024enen	0,35	0,55	0,33	0,25	0,28
vulc12034enen	0,38	0,57	0,40	0,23	0,33
Average	0,26	0,32	0,24	0,25	0,24
Median	0,26	0,29	0,25	0,23	0,23

Standard Dev	0,09	0,13	0,09	0,11	0,11

APPENDIX 2: Overall results at READING TEST level: Median, Average, and Standard Deviation for all runs

RUN_NAME	c@1 r id_1	c@1 r id_2	c@1 r id_3	c@1 r id_4	c@1 r id_5	c@1 r id_6	c@1 r id_7	c@1 r id_8	c@1 r id_9	c@1 r id_10	c@1 r id_11	c@1 r id_12	c@1 r id_13	c@1 r id_14	c@1 r id_15	c@1 r id_16
btbn12011bgbg	0,30	0,30	0,20	0,20	0,20	0,01	0,40	0,30	0,20	0,50	0,50	0,20	0,40	0,30	0,20	0,40
btbn12021bgbg	0,10	0,30	0,00	0,00	0,10	0,00	0,00	0,00	0,30	0,20	0,10	0,10	0,10	0,20	0,30	0,10
btbn12031bgbg	0,20	0,20	0,30	0,20	0,10	0,30	0,10	0,10	0,10	0,30	0,20	0,30	0,10	0,20	0,30	0,20
diue12012enen	0,50	0,20	0,20	0,40	0,30	0,20	0,40	0,20	0,00	0,30	0,00	0,20	0,60	0,30	0,40	0,30
diue12024enen	0,70	0,30	0,20	0,40	0,20	0,30	0,30	0,20	0,20	0,30	0,10	0,20	0,60	0,30	0,60	0,10
fdcs12011enen	0,00	0,16	0,32	0,16	0,00	0,00	0,00	0,20	0,15	0,00	0,00	0,42	0,34	0,00	0,00	0,00
fdcs12021enen	0,36	0,26	0,33	0,24	0,15	0,12	0,00	0,20	0,24	0,00	0,00	0,39	0,32	0,16	0,00	0,00
fdcs12032enen	0,00	0,26	0,42	0,24	0,00	0,00	0,00	0,20	0,15	0,00	0,00	0,42	0,34	0,00	0,00	0,00
fdcs12042enen	0,33	0,26	0,30	0,30	0,15	0,12	0,00	0,20	0,24	0,00	0,00	0,39	0,32	0,16	0,00	0,00
idrq12011arar	0,32	0,16	0,32	0,18	0,30	0,34	0,00	0,00	0,00	0,00	0,19	0,00	0,00	0,00	0,00	0,18
idrq12021arar	0,52	0,30	0,28	0,30	0,30	0,42	0,00	0,00	0,26	0,00	0,00	0,00	0,00	0,19	0,30	0,14
jucs12013enen	0,77	0,70	0,78	0,80	0,42	0,48	0,15	0,17	0,60	0,55	0,65	0,78	0,84	0,90	0,66	0,60
l2fi12011enen	0,00	0,40	0,20	0,20	0,50	0,20	0,10	0,10	0,20	0,10	0,10	0,00	0,40	0,30	0,30	0,10
l2fi12021enen	0,50	0,10	0,30	0,22	0,12	0,10	0,30	0,11	0,44	0,10	0,33	0,33	0,10	0,10	0,12	0,10
l2fi12031enen	0,50	0,60	0,22	0,30	0,70	0,10	0,30	0,40	0,30	0,40	0,20	0,24	0,10	0,55	0,30	0,10
l2fi12041enen	0,60	0,40	0,33	0,40	0,60	0,20	0,30	0,30	0,20	0,60	0,20	0,24	0,10	0,50	0,30	0,20
loga12011dede	0,24	0,12	0,30	0,36	0,15	0,00	0,15	0,00	0,77	0,32	0,30	0,14	0,15	0,16	0,36	0,17
loga12023dede	0,36	0,12	0,30	0,36	0,15	0,13	0,15	0,00	0,77	0,32	0,30	0,14	0,15	0,16	0,36	0,17
mira12011arar	0,16	0,00	0,16	0,32	0,32	0,48	0,00	0,00	0,16	0,00	0,00	0,16	0,48	0,00	0,00	0,16
mira12021arar	0,20	0,30	0,10	0,30	0,40	0,20	0,20	0,20	0,30	0,10	0,00	0,20	0,20	0,20	0,20	0,00
onto12011bgbg	0,20	0,50	0,20	0,50	0,30	0,40	0,30	0,10	0,20	0,30	0,60	0,00	0,20	0,30	0,10	0,20
onto12021enen	0,60	0,20	0,20	0,70	0,40	0,20	0,30	0,10	0,20	0,40	0,50	0,20	0,20	0,30	0,10	0,40
onto12031itit	0,40	0,30	0,60	0,50	0,40	0,40	0,30	0,30	0,20	0,40	0,40	0,10	0,50	0,30	0,20	0,30
onto12041robg	0,10	0,20	0,30	0,30	0,30	0,40	0,30	0,20	0,30	0,50	0,60	0,10	0,40	0,20	0,30	0,10
onto12051roen	0,10	0,20	0,30	0,30	0,30	0,40	0,30	0,20	0,30	0,50	0,60	0,10	0,40	0,20	0,30	0,10
onto12061roit	0,10	0,20	0,30	0,30	0,30	0,40	0,30	0,20	0,30	0,50	0,60	0,10	0,40	0,20	0,30	0,10
onto12071bgbg	0,60	0,30	0,30	0,50	0,30	0,10	0,40	0,20	0,20	0,30	0,70	0,00	0,40	0,20	0,20	0,10
onto12081roro	0,40	0,60	0,10	0,50	0,40	0,30	0,20	0,20	0,40	0,50	0,40	0,10	0,40	0,40	0,30	0,30
onto12091dede	0,40	0,10	0,40	0,40	0,30	0,30	0,20	0,20	0,30	0,20	0,30	0,10	0,10	0,40	0,40	0,30
onto12101eses	0,30	0,40	0,30	0,30	0,20	0,40	0,40	0,20	0,20	0,30	0,50	0,20	0,30	0,20	0,20	0,10
uaic12014roro	0,20	0,30	0,10	0,12	0,33	0,33	0,30	0,11	0,20	0,10	0,30	0,20	0,30	0,33	0,12	0,28
uaic12024roro	0,40	0,30	0,20	0,22	0,33	0,33	0,30	0,11	0,20	0,10	0,30	0,20	0,30	0,33	0,12	0,28
uaic12034roro	0,40	0,30	0,10	0,22	0,33	0,33	0,30	0,12	0,20	0,10	0,30	0,10	0,30	0,33	0,12	0,00
uaic12042roro	0,22	0,22	0,26	0,28	0,20	0,50	0,20	0,11	0,24	0,12	0,24	0,11	0,33	0,12	0,33	0,24
uaic12052roro	0,28	0,00	0,18	0,17	0,26	0,50	0,15	0,13	0,00	0,00	0,14	0,14	0,13	0,16	0,26	0,26
uaic12062enen	0,20	0,10	0,50	0,30	0,44	0,20	0,22	0,00	0,12	0,40	0,28	0,11	0,14	0,16	0,24	0,33
uaic12072enen	0,24	0,00	0,55	0,36	0,42	0,13	0,32	0,00	0,00	0,39	0,16	0,12	0,17	0,00	0,15	0,00
uaic12082enen	0,33	0,24	0,48	0,30	0,16	0,12	0,33	0,36	0,12	0,45	0,44	0,00	0,12	0,17	0,28	0,30
uaic12092enen	0,30	0,20	0,55	0,30	0,12	0,11	0,30	0,42	0,11	0,33	0,40	0,13	0,11	0,14	0,22	0,30
uaic12102enen	0,26	0,14	0,48	0,39	0,00	0,12	0,32	0,36	0,00	0,34	0,39	0,00	0,15	0,17	0,00	0,30
vulc12014enen	0,70	0,50	0,60	0,40	0,50	0,70	0,30	0,10	0,50	0,40	0,50	0,20	0,20	0,10	0,40	0,30
vulc12024enen	0,60	0,60	0,40	0,60	0,50	0,60	0,00	0,20	0,20	0,30	0,30	0,20	0,20	0,10	0,40	0,40
vulc12034enen	0,70	0,60	0,50	0,50	0,50	0,80	0,10	0,20	0,20	0,30	0,30	0,10	0,20	0,30	0,40	0,40
Average	0,34	0,28	0,31	0,33	0,29	0,27	0,21	0,16	0,24	0,26	0,29	0,17	0,27	0,23	0,24	0,20
Median	0,32	0,26	0,30	0,30	0,30	0,30	0,30	0,20	0,20	0,30	0,30	0,14	0,20	0,20	0,26	0,18

Standard Dev	0,20	0,17	0,16	0,15	0,16	0,19	0,13	0,11	0,17	0,18	0,20	0,15	0,17	0,16	0,15	0,14

APPENDIX 3: SYSTEM DESCRIPTIONS

Table 15: 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
Btb				analyzes the questions and possible answers	x		x					x										Lemmatization
Diue			x			x		x														
Fdcs			x		x			x														
Idrq			x		x	x	x					x	x		x	x	x	x				Conceptual Graphs; Similarity scoring
Jucs			x		x	x	x	x	x	x		x	x	x	x	x						
l2fi	x				x												x					
Loga		x												x					x	x		
Mira	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Onto			x				x															Some basic rules for identification of abbreviations prior to sentence splitting. Similarity metric: string overlap.
Uaic		x			x			x	x	x	x			x								
Vulc	x			weighted lexical matching, paraphrasing, sentential proximity			x					x	x	x								

Table 16: Use of Knowledge by participating systems

System name	Knowledge Resources Used														Tools				
	captured from the background collection	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
btb							x	x										x	x	
diue	x																x		x	x
fdcs		x															x	x		
idrq		x			x		x			x	x	x						x	x	
jucs	x															x	x	x	x	x
l2fi										x								x		
loga		x			x					x						x			x	
mira	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
onto	x																			x
uaic	x				x					x					x		x	x	x	x
vulc		x				x	x								x				x	

Table 17: 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	System Description
btb	x								System analyzes the questions and the answers and it is looking for overlapping n-grams from the analyses of the question and the answers. The answer with greatest overlapping is selected.
diue	x								The first run is based on surface text analysis, with no grammatical nor semantic processing.
fdcs									The system uses redundancies in the collection
idrq				x	x	x	x		The system "IDRAAQ" is developed for Arabic QA integrates three levels of processing: - Keyword-based level: uses a Arabic WordNet based Query Expansion module - Structure-based level: consists in measuring the Density Distance N-gram Model of candidate
jucs					x	x	x		It uses Lexical, Syntactic, Semantic level.
l2fi	x				x				The system uses lexical distances such as Word Proximity and similarity measures to select the candidate answer more related with the question. The system also uses Latent Semantic Analysis to extract latent topics from the test documents
loga								x	The system parses questions, answers and documents. Coreference resolution is applied to the document representation. The system constructs a hypothesis from question and answer parse. It then tries to prove the hypothesis from the logical document
mira	-	-	-	-	-	-	-	-	-

onto					x				The system relies on approximate string matching - an approach that makes it easily adaptable to any language without significant modifications. The analysis comprises two phases: 1. Extraction of sentences that match best the query and answer string
uaic					x	x	x		Romanian system uses syntactic similarity and the English system uses semantic similarity.
vulc	x								The system uses a simple principle: An answer A is likely to be the right answer to the question Q if a sentence like A in the document is very close (within 1 or 2 sentences) to a sentence Sq like Q in the document.