Overview of QA4MRE Main Task at CLEF 2013

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

[1] School of CSEE, University of Essex, UK (rsutcl@essex.ac.uk)
[2] NLP&IR group, UNED, Spain (anselmo@lsi.uned.es; alvarory@lsi.uned.es)
[3] Carnegie Mellon University, USA (hovy@cmu.edu)
[4] CELCT, Italy (forner@celct.it)
[5] Al. I. Cuza University of Iasi, Romania (corinfor@info.uaic.ro)
[6] Philips Research North America, USA (Yassine.Benajiba@philips.com)
[7] Bulgarian Academy of Sciences, Bulgaria (petya@bultreebank.org)

Abstract. This paper describes the Question Answering for Machine Reading (QA4MRE) Main Task at the 2013 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, Machine Reading on Biomedical Texts about Alzheimer's disease, and Japanese Entrance Exams. 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. We introduced two novelties this year: auxiliary questions to evaluate systems level of inference, and a portion of questions where none of the options were correct. Nineteen groups participated in the task submitting a total of 77 runs in five languages.

1 INTRODUCTION

The QA4MRE Lab focuses on the reading of single documents and the identification of the correct answers to a set of questions. Questions are in the form of multiple choices, 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. 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.

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 co-reference, 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 2013, we had three exercises.

1. **Main Task.** This remained the same for participants. Background collections, test documents and reading tests were available in Arabic, Bulgarian, English, Romanian, and Spanish. There were four topics: AIDS, Alzheimer's Disease, Climate Change and Music and Society. As was the case last year, there is also a pilot task on Alzheimer's disease. 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. Following the pilot task last year on Processing Modality and Negation, these aspects were incorporated into questions within the main task.

2. **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.
3. **Entrance Exams.** University Entrance Exams include questions formulated at various levels of complexity and test a wide range of capabilities. The challenge of "Entrance Exams" aims at evaluating systems under the same conditions humans are evaluated to enter the University. In this first campaign we will reduce the challenge to Reading Comprehension exercises contained in the English exams. More types of exercises will be included in subsequent campaigns (2014–2016) in coordination with the "Entrance Exams" task at NTCIR. Exams are created by the Japanese National Center for University Admissions Tests. The "Entrance Exams" corpus is provided by NII's Todai Robot Project and NTCIR.

In this paper we describe the Main task. The two other tasks are described in detail in dedicated papers in these proceedings.

2.1 Main Task

Tests were divided into:
- 4 topics, namely “Aids”, “Alzheimer”, “Climate change” and “Music and Society”;
- Each topic had four reading tests;
- Each reading test consisted of one single document, with 15 Main questions (six having no answer in the text) and a set of five choices per question. The last of the five choices was always “None of the above”. In addition, one or more Auxiliary questions could be asked, each of which was a simplification of a Main question (see discussion later).

Overall, the following evaluation setting was proposed:
- 16 test documents (4 documents for each of the four topics),
- 240 Main questions (15 questions for each document),
- 1200 choices/options (5 for each question).

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

2.2 What’s new this year?

We introduced two novelties this year: (i) auxiliary questions to evaluate systems level of inference, and (ii) a portion of questions where none of the options were correct.

With respect to auxiliary questions, they correspond to Main questions where a deliberate simplification is done by removing one inference step. The idea was that if a system answered a Main question incorrectly but the corresponding

Auxiliary question correctly, it suggests that the system was near to answering the question but could not perform the inference step. In a similar way, if a system answers the main question but not the simplified one, this indicates a lack in the inference process. Hence this approach could be used to pinpoint the exact shortcomings of a system.

With respect to questions without correct answers among candidates, the idea is to test the ability to reject candidate answers when they are incorrect. We implemented this change by introducing in our tests a portion of questions where none of the options are correct and including a new last option in all questions: "None of the answers above is correct".

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, etc. 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 2011, 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. Besides, participants had available the collections used in 2011 (see Table 2).

Table 2. Size of the background collections in the various languages for all 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

Table 3 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.

Table 3. Queries used to build the reference collections

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
	gas emissions
brain disorder AND neurological disorder	biomass
plaques and tangles	
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
personality changes AND Alzheimer	AIDS life expectancy rate
beta-amyloid	fighting AIDS
(caregiving OR long-term care) AND Alzheimer	AIDS virology
nursing home AND Alzheimer	
(aging society OR geriatrics) AND Alzheimer	MUSIC AND SOCIETY KEYWORDS (EXTENSION)
healthcare costs AND Alzheimer	
cognitive reserve theory	music criticism
Auguste Deter	musicology

Danae Chambers Alzheimer's Association Alzheimer diagnosis Alzheimers' associated disorders Alzheimers' clinical features Alzheimers' genetics Alzheimers' prevention Familial Alzheimer's Alzheimers' risk factors impact of Alzheimer's disease Neuropathology of Alzheimer's Disease	history of violin technique music patronage rock and roll history of song electric musical instrument classical recording industry economics of classical music classical crossover music

4 TEST SET PREPARATION

This year the datasets was created for the following five languages: Arabic, Bulgarian, English, Romanian and Spanish. The dataset was created following the methodology developed in previous years and consisting of the following steps:

1. Four English documents were selected for each of the four topics (Aids, Alzheimer's, Climate Change, Music and Society). These were selected from various sources (see Table 4) and comprised the test documents against which questions were asked. The documents for the first three topics were chosen from copyright-free sources. The documents for Music and Society were selected from Grove Music Online (http;//www.oxfordmusiconline.com) by kind permission of the Editor in Chief, Editor and Oxford University Press. This source was chosen because of its exceptional scholarly quality, as well as the very large choice of articles available on music of all kinds.

2. In order to have a set of identical questions for the five 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.

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

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

6. Once the questions had been composed in the language of the original author, each was then translated into English. The English versions of the questions and candidate answers 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.

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

7. The English versions 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.

8. The result of this process was a set of 240 Main questions and 44 Auxiliary 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.

Table 4. Test Documents

Topic	No.	Source	Author	Title	LICENSE	Words
Alzheimer	1	http://blog.kylebarlow.com/2012/04/of-mice-and-men-alzheimers-cure-for-our.html	Kyle Barlow	What's life? Of mice and men: an Alzheimer's cure for our murine brethren	Creative Commons Attribution-NonCommercial 3.0	1159
Alzheimer	2	http://www.insight.mrc.ac.uk/2012/10/19/fighting-alzheimers-disease-get-the-immune-system-on-board/	James Fuller	Fighting Alzheimer's disease? Get the immune system on board	Creative Commons Attribution	859
Alzheimer	3	http://www.ted.com/talks/alanna_shaikh_how_i_m_preparing_to_get_alzheimer_s.html	Alanna Shaikh	How I'm preparing to get Alzheimer's	Attribution-NonCommercial-NoDerivs	1109
Alzheimer	4	http://www.alz.co.uk/icaniwill/library/people-with-dementia/living-with-early-memory-loss/financial-challenges	Mike Donohue	Financial challenges faced by person with dementia	Creative Commons AttributionShareAlike	2320
Music & Society	5	Grove Music Online at http://www.oxfordmusiconline.com	Jerald C. Graue, Thomas Milligan	Johann Baptist Cramer	Copyright Oxford University Press, used with permission	1749
Music & Society	6	Grove Music Online at http://www.oxfordmusiconline.com	Geeta Dayal, Emily Ferrigno	Electronic Dance Music	Copyright Oxford University Press, used with permission	2040
Music & Society	7	Grove Music Online at http://www.oxfordmusiconline.com	Mervyn Cooke	Film Music -Hollywood	Copyright Oxford University Press, used with permission	1712
Music & Society	8	Grove Music Online at http://www.oxfordmusiconline.com	Thomas Christensen	Disciplines of Musicology - Analytic Traditions	Copyright Oxford University Press, used with permission	1255
Climate Change	9	http://www.fpif.org/articles/latin_america_climate_change_swing_states	Janet Redman	"Latin America: Climate Change Swing States" (Washington, DC: Foreign Policy In Focus, July 22, 2010)	Creative Commons Attribution	2335
Climate Change	10	http://www.fpif.org/articles/global_warming_its_all_about_energy	Michael Klare	"Global Warming: It's All About Energy" (Washington, DC: Foreign Policy In Focus, February 15, 2007)	Creative Commons Attribution	1347
Climate Change	11	http://www.fpif.org/reports/ozone_depletion_global_warming	Jessica Vallette Revere	"Ozone Depletion & Global Warming" (Washington, DC: Foreign Policy In Focus, October 12, 2005)	Creative Commons Attribution	2364

Climate Change	12	http://www.fpif.org/articles/preventing_a_blowout_in_the_arctic	Julia Heath	"Preventing a Blowout in the Arctic" (Washington, DC: Foreign Policy In Focus, February 15, 2012)	Creative Commons Attribution	1827
AIDS	13	http://www.fpif.org/articles/how_to_stop_aids_now	By Caiti Schroering	"How to Stop AIDS Now" (Washington, DC: Foreign Policy In Focus, August 21, 2007)	Creative Commons Attribution	1308
AIDS	14	http://www.fpif.org/articles/curing_aids_policy_of_greed_and_dogma	Yifat Susskind	"Curing AIDS Policy of Greed and Dogma" (Washington, DC: Foreign Policy In Focus, November 30, 2006)	Creative Commons Attribution	1191
AIDS	15	http://www.fpif.org/articles/aids_in_africa_and_black_america	Kwei Quartey	"AIDS in Africa and Black America" (Washington, DC: Foreign Policy In Focus, October 11, 2012)	Creative Commons Attribution	1124
AIDS	16	http://www.fpif.org/articles/aids_appointee_shows_that_business_still_rules_the_roost	Jim Lobe	"AIDS Appointee Shows that Business Still Rules the Roost" (Washington, DC: Foreign Policy In Focus, July 3, 2003)	Creative Commons Attribution	1067

4.1 Questions

For each text in the test set 10 multiple choice questions were created. Each question had five answer options. The fifth option was always 'None of the above'. 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 5 below. Example questions:

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

Table 5. Distribution of question types

Question type	**Total number of questions**
PURPOSE	31
METHOD	44
CAUSAL	48
FACTOID*	80
WHICH-IS-TRUE	81
TOTAL # of QUESTIONS	284

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 6 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 6. Classification of questions according to the knowledge required to answer them

Types of question	**#of questions**	**c@1**
Same sentences	119	0.33
Background knowledge required	45	0.30
Information needs to be gathered from difference sentences of paragraphs	120	0.22

Table 7. Numbers of questions that are auxiliary, have no answer, or contain modality or negation

Types of question	**#of questions**	**c@1**
AUXILIARY QUESTIONS	44	0.48
NO CORRECT ANSWER	39	0.05
MODALITY AND NEGATION	28	0.21

Table 7 below shows a breakdown of questions which are auxiliary (see below), have no correct answer, or contain modality or negation in either the question or the answer.

4.2 Auxiliary Questions

In the first two years of the QA4MRE task, questions required a deep understanding of the text. However, since they were multiple-choice, the answer was simply judged as correct or incorrect. In the case of a correct question, it was impossible to judge whether the answer had been chosen at random or derived from a valid process of deduction. Similarly, if the answer was incorrect, it was impossible to judge why. To address this latter issue, an experiment was conducted this year in which Auxiliary questions were posed in addition to Main questions. Each Auxiliary question corresponded to a Main question and was a deliberate simplification of it which removed one inference step. The idea was that if a system answered a Main question incorrectly but the corresponding Auxiliary question correctly, it suggests that the system was near to answering the question but could not perform the inference step. Hence this approach could be used to pinpoint the exact shortcomings of a system.

In the main, three forms of simplification were used, hypernym replacement, noun phrase synonymy, and verbal entailment. Moreover, simplification could be made to the question or to the correct answer. Here are some examples.

Hypernym Replacement

In this example, the simplification is made to the question:
Q (main): What has been offered to the President of the United States if he signs the Kyoto Protocol?
Q (aux): What has been offered to Obama if he signs the Kyoto Protocol?
Supporting text:

> Perhaps most surprising was Stern's stop in Quito, Ecuador. The United States slashed $2.5 million of support when Ecuador submitted a letter that it would not join the accord. In response, Ecuadorian Foreign Minister Ricardo Patiño offered the United States $2.5 million if Obama signed the Kyoto Protocol.

Here, the hypernym "President of the United States" has been replaced by its hyponym "Obama" in the Auxiliary question. The supporting text refers to Obama and not President of the United States. Thus in order to answer the Main question, a system must infer that Obama is the President. This inference is not needed for the corresponding Auxiliary question since Obama is actually mentioned in both the question and the document.

Noun Phrase Synonymy

Q: What sort of music was written for Hollywood films in the Golden Age?
Supporting text:

> The conventions of the "classical" Hollywood film score in the Golden Age - essentially a *leitmotif-based symphonic romanticism* with narrative orientation, the music almost always subordinated to the primacy of the visual image and dialogue - prevailed in scores by other expatriate musicians.

Here, the simplification is made to the answer, while the wording of the Main question and Auxiliary question remains the same:
A (main): music for orchestra with strong melodies
A (aux): music embodying leitmotif-based symphonic romanticism

The Main question can only be answered by deducing that "music for orchestra with strong melodies" is largely synonymous with "leitmotif-based symphonic romanticism"; i.e., "symphonic" implies that the music is for full orchestra while "leitmotif-based" implies the use of strong easily-recognised melodies, associated with ideas or characters in the film.

Verbal Entailment

Q: What was Cramer's attitude towards the music of Bach?
Supporting text (with added italics):

> He may have been introduced to Das wohltemperirte Clavier as early as 1787, and *he developed a lifelong fascination for Bach*.

Again, the simplification is made to the answer:
A (main): he admired Bach all his life
A (aux): he developed a lifelong fascination for Bach

Here, the Main question can only be answered by deducing that "he admired Bach all his life" is entailed by the supporting text "he developed a lifelong fascination for Bach". In the Auxiliary question, the answer is a substring of the supporting text, so no entailment is needed.

In all, 44 Auxiliary questions were composed, seventeen containing a simplification of the question and 27 containing a simplification of the correct answer. Analysis of the results concerning these questions can be found in Section 6.1.

5 EVALUATION

This task has the aim of promoting a change in QA architectures giving more importance to the validation step over the IR component in order to improve results. This is why we have been proposing from 2009 to evaluate system confidence by introducing the possibility of leaving questions unanswered [1]. Thus, systems might reduce the amount of incorrect answers while keeping the proportion of correct ones.

However, the analysis of last editions has shown how systems rely more on ranking than in validation of candidate answers. These systems calculate the similarity of each candidate answer with a combination of the question and certain snippets of the document and return the most similar answer. Hence, systems have not shown nor developed their ability discarding incorrect answers. Besides, it is not clear the behavior in case of not providing the candidate answers.

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

This modification does not affect the output of participants since 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, taking into account that one candidate is “None of the answers above is correct”
- A *NoA* answer. This response should be given if the system considers it is not able to find enough evidences about the correctness of candidate answers and it prefers not to answer the question instead of giving an incorrect answer. Moreover, the system can return as a hypothetical answer the candidate one that it would have been selected, which allows to give some feedback about its validation performance.

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

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

Evaluation of systems given was given from two perspectives following the format of last editions:

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 and rank systems according to the final value given by a measure.
2. A reading-test evaluation, obtaining figures for each particular reading test and topics. This perspective permits us to evaluate whether a system was able to understand a document and to what degree. More in detail, we evaluate if the system is able to pass each test, in a similar way to humans with RC tests. This is a kind of evaluation studied with more detail in the pilot Entrance Exams task.

5.1 Evaluation Measure

We keep *c@1* as the main evaluation in this edition. *c@1* was introduced in ResPubliQA 2009 [1] and is fully described in [2]. The formulation of *c@1* is given in Formula (1).

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

where

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

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

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

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

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

The Question Answering perspective is focused on measuring systems' performance over a set of questions without considering the ability of a system to pass tests associated with documents. This is an approach similar to the one applied in QA@CLEF campaigns before 2010.

Then, the information considered for each system at this level is:

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

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

The following scores are calculated from this information:

- An overall *c@1* score over the whole collection (the set with 160 questions),
- A *c@1* score for each topic (40 questions for each topic),
- An overall *accuracy* score (over the 160 questions of the test collection, considering also the candidate answers given to unanswered questions as it has been explained above),
- The proportion of answers correctly discarded (see Formula (3)) in order to evaluate the validation performance.

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

where:

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

5.3 Reading Perspective Evaluation

The objective of the reading perspective evaluation is to offer information about the performance of a system "understanding" the meaning of each single document. This understanding is evaluated by means of multiple-choice tests with ten questions per document. That is, each system has to pass a test about a given document similar to the evaluation of RC of new language learners. As we said above, this kind of evaluation is studied more in detail in a pilot task of this edition.

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

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

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

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

5.4 NCA Baseline

After previous years' experience, we realized that advancing the state of the art requires systems ability to decide whether all candidate answers were incorrect or not. In this way, systems able to take this decision should be rewarded over systems that just rank answers. For this reason, we introduced an additional option "none of the above answers are correct" (NCA), that was the correct option in the 39% of questions. Thus, this is the baseline for a dummy system that always return NCA.

5.5 Random Baseline

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

6 PARTICIPATION

From an initial amount of 39 groups that registered for the main task and signed the license agreement to download the background collections, 19 of them finally submitted at least one run, resulting in 77 monolingual runs in four languages (Bulgarian, English, Spanish, and Romanian). There were no Arabic runs this year and neither were there any cross-lingual runs. Tables 8-10 show a characterization of runs.

Table 8. Overall participants and runs in QA4MRE tasks

REGISTERED PARTICIPANTS	PARTICIPANTS DOWNLOADING THE TEST SETS	PARTICIPANTS SUBMITTING RUNS	TOTAL NUMBER OF RUNS
39	32	19	77

Table 9. Participants and runs per tasks

NUMBER of PARTICIPANTS	19	NUMBER of RUNS	77
MAIN	11	MAIN	54
BIOMEDICAL about ALZHEIMER	3	BIOMEDICAL about ALZHEIMER	13
ENTRANCE EXAM	5	ENTRANCE EXAM	10

Table 10. Runs submitted per language in the QA4MRE Main Task

	Target languages (corpus and answer)						
Source langs (questns)		AR	BG	EN	ES	RO	Total
	AR						0
	BG		10				10
	EN			36			36
	ES				3		3
	RO					5	5
	Total	0	10	36	3	5	54

7 RESULTS

7.1 Reading Perspective

Table 11 shows the average results for each one of the 16 reading comprehension tests according to *c@1*. The Table shows that, except for Test 3, the mean value was higher than the baseline of 0.2, a similar situation to last year.

Table 11. 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
Average	0.38	0.27	0.20	0.26	0.24	0.29	0.32	0.37	0.26	0.24	0.25	0.29	0.24	0.30	0.25	0.24

However, the mean values for all the tests were still under 0.5, which is the score needed to pass the evaluation from the reading perspective. This is the same result as last year and suggests that systems are still far away from obtaining satisfactory results according to this perspective.

Table 12. Mean Scores for each Topic

	Topic 1 Alzheimer	Topic 2 Music	Topic 3 Climate Change	Topic 4 Aids
Average	0.28	0.31	0.26	0.26

Table 12 shows the mean scores per topic. The scores across topics are within five percentage points which seems very consistent in difficulty, especially as the topics themselves are so diverse. This year the Music questions were slightly easier whereas last year (2012) the AIDS questions were slightly easier than the others.

Apendix I and II show these results for al submitted runs

7.2 Question Answering perspective

Table 13 shows the results of all submitted runs grouped by language. Most of the systems were able to beat the baseline (only 5 runs performed lower, compared to 8 in 2012), with at least a system per language able to do so. Once again, this amount is higher than in 2011, when only half of the systems outperformed the baseline. So there seems to be a progressive improvement in systems' performance over the years. Considering all languages, 93% of questions received at least one correct answer by at least one system.

Table 13. c@1 in participating systems according to language

System name	BG	EN	ES	RO
jucs1302enen		0.59		
Baseline NCA	0.39	0.39	0.39	0.39
kule1304enen		0.35		
kule1305enen		0.35		
kule1302enen		0.34		
buap1301enen		0.33		
kule1303enen		0.33		
kule1306enen		0.33		
nara1301enen		0.33		
buap1304enen		0.32		
buap1305enen		0.32		
kule1308enen		0.32		
buap1303enen		0.31		
buap1309enen		0.31		
kule1309enen		0.31		
buap1306eses			0.30	
csui1302enen		0.30		
csui1304enen		0.30		
csui1305enen		0.30		
csui1307enen		0.30		
kule1307enen		0.30		
nara1302enen		0.30		
csui1303enen		0.29		
csui1306enen		0.29		
csui1308enen		0.29		
csui1309enen		0.29		
buap1307eses			0.28	
buap1308eses			0.28	
csui1310enen		0.28		
lims1304enen		0.28		
lims1305enen		0.28		
onto1305bgbg		0.28		
onto1307bgbg		0.27		
onto1308bgbg		0.27		
evma1303enen		0.26		

lims1302enen		0.26		
lims1303enen		0.25		
onto1306bgbg		0.25		
uaic1303roro		0.25		
buap1302enen		0.24		
buap1310enen		0.24		
evma1301enen		0.24		
vens1301enen		0.24		
vens1302enen_LATE_RUN		0.24		
uaic1302roro				0.23
uaic1306roro				0.23
onto1301bgbg	0.22			
onto1302bgbg	0.22			
onto1303bgbg	0.22			
onto1304bgbg	0.22			
Baseline 1	0.20	0.20	0.20	0.20
btbn1301bgbg	0.19			
evma1302enen		0.19		
btbn1302bgbg	0.18			
uaic1304roro				0.17
uaic1305roro			0.17	
Baseline 2	0.16	0.16	0.16	0.16

The best results were obtained in English, where the highest score was obtained by *jucs1302enen* with 0.59. This value is 24 percentage points higher than the next system (kule1304enen at 0.35). In 2012 the jucs group also submitted the best run, *jucs12013enen* with a score 0.65. So, for a second year running, jucs was the only system able to pass the evaluation according to the reading perspective. Moreover, their individual scores were well over 0.5 in every topic except Topic 3, Climate Change (where their score was 0.3).

Table 14 shows the distribution of correct and incorrect answers for all runs.

Table 14. Distribution of questions answered correctly, incorrectly and unanswered

		# of questions ANSWERED			# of questions UNANSWERED			
RUN_NAME	C@1 ALL questions	Total	# RIGHT	# WRONG	Total	# with RIGHT candidate answer	# with WRONG candidate answer	# with EMPTY candidate answer
jucs1302enen	0.59	225	138	87	59	0	0	59
NCA baseline	0.39	-	-	-	-	-	-	-
kule1304enen	0.35	265	93	172	19	0	0	19
kule1305enen	0.35	284	100	184	0	0	0	0
kule1302enen	0.34	284	96	188	0	0	0	0
buap1301enen	0.33	264	88	176	20	4	16	0
kule1303enen	0.33	242	82	160	42	0	0	0
kule1306enen	0.33	284	93	191	0	0	0	0

nara1301enen	0.33	270	88	182	14	0	0	14
buap1304enen	0.32	284	91	193	0	0	0	0
buap1305enen	0.32	240	79	161	44	13	31	0
kule1308enen	0.32	257	83	174	27	0	0	27
buap1303enen	0.31	284	87	197	0	0	0	0
buap1309enen	0.31	284	88	196	0	0	0	0
kule1309enen	0.31	284	87	197	0	0	0	0
buap1306eses	0.30	274	83	191	10	1	9	0
csui1302enen	0.30	245	76	169	39	0	0	39
csui1304enen	0.30	246	75	171	38	0	0	38
csui1305enen	0.30	222	70	152	62	0	0	62
csui1307enen	0.30	230	71	159	54	0	0	54
kule1307enen	0.30	238	74	164	46	0	0	46
nara1302enen	0.30	267	80	187	17	0	0	17
csui1303enen	0.29	244	72	172	40	0	0	40
csui1306enen	0.29	230	70	160	54	0	0	54
csui1308enen	0.29	234	60	165	50	0	0	50
csui1309enen	0.29	236	70	166	48	0	0	48
buap1307eses	0.28	282	80	202	2	0	2	0
buap1308eses	0.28	282	79	203	2	0	2	0
csui1310enen	0.28	233	67	166	51	0	0	51
lims1304enen	0.28	284	79	205	0	0	0	0
lims1305enen	0.28	284	80	204	0	0	0	0
onto1305bgbg	0.28	240	69	171	44	3	41	0
onto1307bgbg	0.27	277	76	201	7	0	7	0
onto1308bgbg	0.27	284	76	208	0	0	0	0
evma1303enen	0.26	249	64	181	39	0	0	39
lims1302enen	0.26	284	73	211	0	0	0	0
lims1303enen	0.25	284	72	212	0	0	0	0
onto1306bgbg	0.25	284	72	212	0	0	0	0
uaic1303roro	0.25	270	68	202	14	1	11	2
buap1302enen	0.24	284	68	216	0	0	0	0
buap1310enen	0.24	284	67	217	0	0	0	0
evma1301enen	0.24	224	57	167	60	0	0	60
vens1301enen	0.24	274	65	209	10	1	2	7
vens1302enen_LATE_RUN	0.24	284	68	216	0	0	0	0
uaic1302roro	0.23	162	45	117	122	24	96	2
uaic1306roro	0.23	162	45	117	122	24	96	2
onto1301bgbg	0.22	276	62	214	8	0	8	0
onto1302bgbg	0.22	284	62	222	0	0	0	0
onto1303bgbg	0.22	281	63	218	3	0	3	0
onto1304bgbg	0.22	284	63	221	0	0	0	0
btbn1301bgbg	0.19	284	53	231	0	0	0	0
evma1302enen	0.19	219	44	175	65	0	0	65

btbn1302bgbg	0.18	284	51	233	0	0	0	0
uaic1304roro	0.17	255	44	211	29	7	22	0
uaic1305roro	0.17	185	36	149	99	15	84	0

Table 15 shows the difference in performance for each type of question. Notice that consistently, all systems perform better on the auxiliary questions that require less inference than main questions. Also results over questions with modality and negation are lower for all runs than the score obtain for all questions.

However, the most important result is that scores show how systems can't decide whether there is a correct answer or not among candidates. This is a very important challenge that we have to continue addressing in future.

Table 15. c@1 in participating systems considering auxiliary questions and main questions, questions without correct answer and questions with modality and negation

RUN_NAME	c@1 on ALL questions	c@1 on main questions	c@1 on auxiliary questions	c@1 on NCA questions	c@1 on Mod/Neg questions
jucs1302enen	0.59	0.55	0.74	0.46	0.38
kule1304enen	0.35	0.28	0.72	0.12	0.32
kule1305enen	0.35	0.28	0.75	0.11	0.30
kule1302enen	0.34	0.28	0.66	0.14	0.30
buap1301enen	0.33	0.27	0.67	0.04	0.28
kule1303enen	0.33	0.26	0.71	0	0.32
kule1306enen	0.33	0.30	0.47	0.20	0.27
nara1301enen	0.33	0.28	0.57	0.23	0.18
buap1304enen	0.32	0.25	0.68	0.11	0.20
buap1305enen	0.32	0.24	0.69	0.04	0.19
kule1308enen	0.32	0.29	0.50	0.10	0.26
buap1303enen	0.31	0.24	0.68	0	0.23
buap1309enen	0.31	0.28	0.56	0.20	0.27
kule1309enen	0.31	0.28	0.48	0.09	0.27
buap1306eses	0.30	0.27	0.51	0	0.27
csui1302enen	0.30	0.26	0.55	0	0.16
csui1304enen	0.30	0.25	0.55	0	0.16
csui1305enen	0.30	0.25	0.59	0	0.17
csui1307enen	0.30	0.26	0.53	0	0.25
kule1307enen	0.30	0.26	0.51	0	0.27
nara1302enen	0.30	0.24	0.59	0	0.15
csui1303enen	0.29	0.24	0.55	0	0.20
csui1306enen	0.29	0.25	0.54	0	0.25
csui1308enen	0.29	0.25	0.50	0	0.21
csui1309enen	0.29	0.25	0.48	0	0.23
buap1307eses	0.28	0.24	0.55	0.02	0.10
buap1308eses	0.28	0.23	0.55	0	0.10
csui1310enen	0.28	0.24	0.48	0	0.23
lims1304enen	0.28	0.23	0.52	0.04	0.10
lims1305enen	0.28	0.23	0.57	0	0.13
onto1305bgbg	0.28	0.26	0.39	0	0.25

onto1307bgbg	0.27	0.24	0.43	0	0.15
onto1308bgbg	0.27	0.24	0.43	0	0.13
evma1303enen	0.26	0.24	0.33	0.14	0.23
lims1302enen	0.26	0.22	0.48	0.04	0.10
lims1303enen	0.25	0.20	0.52	0.04	0.10
onto1306bgbg	0.25	0.23	0.36	0	0.2
uaic1303roro	0.25	0.23	0.38	0	0.38
buap1302enen	0.24	0.20	0.43	0	0.17
buap1310enen	0.24	0.19	0.48	0	0.20
evma1301enen	0.24	0.23	0.32	0.14	0.14
vens1301enen	0.24	0.22	0.33	0.17	0.08
vens1302enen _LATE_RUN	0.24	0.21	0.33	0.18	0.12
uaic1302roro	0.23	0.21	0.30	0	0.42
uaic1306roro	0.23	0.21	0.30	0	0.42
onto1301bgbg	0.22	0.19	0.41	0	0.11
onto1302bgbg	0.22	0.18	0.41	0	0.10
onto1303bgbg	0.22	0.19	0.43	0	0.17
onto1304bgbg	0.22	0.18	0.43	0	0.17
btbn1301bgbg	0.19	0.19	0.18	0.20	0.20
evma1302enen	0.19	0.18	0.25	0.06	0.09
btbn1302bgbg	0.18	0.17	0.25	0.08	0.10
uaic1304roro	0.17	0.17	0.20	0.03	0.21
uaic1305roro	0.17	0.16	0.21	0.04	0.14

Finally, Table 16 compares the performance of systems in the three editions of QA4MRE. Results show how introducing NCA questions together with modality and negation made the task more difficult this year.

Table 16. Average Scores over all runs and over best runs for 2013, 2012, and 2011

	over all runs	over all best runs
QA4MRE 2013	0.24	0.27
QA4MRE 2012	0.26	0.32
QA4MRE 2011	0.21	0.28

7.3 Unanswered Questions

Table 17 below shows the percentage of correct and NoA answers for different question types over the last three years. Percentages of correct answers are lowest for Causal questions at 22.56% with Purpose following at 24.19%. Which-is-true and Factoid are similar at 25.44% and 25.92%, while the highest is Method at 30.64%. Similar to last year, the Causal questions are the hardest to answer. This corresponds to the intuition that the need for inference in such questions can cause difficulties for systems. However, while Method questions were the most difficult questions after Causal ones last year, this time around the Method questions seems to be the easiest. It may be that easier Method questions were set this year. NoA scores are similar across question types although, interestingly, the number is lowest at 6.32% for Causal questions even though these were the hardest. It seems therefore that for Causal questions, systems were less inclined to withhold their answers than for other types, but then in answering such questions they were less successful than for other types.

Table 17. Percentage of Correct and NoA answers according to different question type shown over the last three years

2013 Data		
Question type	**% of correct answers**	**% of NoA answers**
PURPOSE	24.19	8.42
METHOD	30.64	9.89
CAUSAL	22.56	6.32
FACTOID*	25.92	9.30
WHICH-IS-TRUE	25.44	9.55

2012 Data		
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

2011 Data		
Question type	**% of correct answers**	**% of NoA answers**
CAUSE	18	39
DEGREE-OF-TRUTH	40	40
COMPOSITE	15	30
FACTOID *	30	38
HYPOTHETICAL	16	31
METHOD	28	50
OPINION	23	49
PURPOSE	24	38
RESULTS	31	33
WHICH-IS-TRUE	29	37

7.4 Analysis of Auxiliary Questions

As stated above, various Auxiliary questions were added to the test set, each such question being a simplification of a particular Main question. Simplifications took three main forms: hypernym replacement, noun phrase synonymy and verbal entailment: 16 were hyponym replacement (HYP), 18 were noun phrase synonymy (NPS), and 10 were verbal entailment (VEN).

In total there were 44 Auxiliary questions, seventeen being simplifications of Main questions themselves and 27 being simplifications of the correct answers to Main questions. The simplifications were designed to identify the (in)ability of a system to perform specific inferences; essentially, the main question required the inference while the auxiliary one did not. In consequence, we were looking for instances where systems found auxiliary questions easier to answer than their main counterparts. We identified two means of studying the data. First, we looked at how many *systems* answered main questions correctly as against how many answered the corresponding auxiliary ones correctly (Tables 15 and 17). We expected to see more systems answering auxiliary questions correctly, if indeed they were easier to answer. Secondly we looked at how many main-auxiliary *question pairs* had the property that more systems answered the auxiliary question correctly than answered the main question correctly. These results are shown in Tables 16 and 18. We were expecting most pairs to have this property.

Overall, our expectations were fulfilled since Auxiliary question simplification led to a score increase in 36 out of 44 cases (81.82%): 13 out of 17 for question simplification (Table 16) and 23 out of 27 for answer simplification (Table 18). Where scores increased, they did so strongly, by 244.21%: in other words there were about two-and-a-half times as many correct responses on average for Auxiliary questions relative to their corresponding Main questions, in cases where there was any increase. This appears to support our hypothesis that certain key inferences were causing systems to get answers

wrong. There was a big score difference between simplifying the Question and simplifying the Answer. Question simpli-simplification led to a 58.96% increase, while Answer simplification gave 348.91%. It appears that answer simplification makes a question much easier to answer than question simplification.

Question Simplification. The breakdown of this by simplification type is shown in Tables 18 and 19. These tables only consider main questions that have an Auxiliary counterpart. Table 18 shows the counts of Main questions correct, the equivalent Auxiliary questions correct and the percentage difference. These figures are then broken down by simplification type (HYP, NPS or VEN) in the last nine columns. Recall that question type and domain are two different ways of breaking down the same set of question pairs.

For operational reasons there no auxiliary questions for the Alzheimer's topic. Table 18 shows that the overall percentage difference between main and auxiliary in correct answers was 25%. Concerning topic, Climate was the lowest (9%) and Aids the highest (39%). Concerning question type, Cause was the highest at 52% while Purpose, Fact and True fall in the range 20-30%. This suggests that Cause questions require the most complex reasoning. Interestingly, Method questions were worse in the auxiliary case. Problems in formulating the auxiliary questions could be the reason here.

Table 19 shows by question-aux pairs of a particular type, in how many of the group falling into that type the Auxiliary question has more correct answers than the Main question. The figures are counts of question pairs, not counts of correct answers. So, for example, there were six main questions that were of Factoid type (in the Auxiliary question pilot) and for five of these, there were more correct answers for the corresponding Auxiliary questions. Generally this table shows a trend of increase in the number of correct answers to an auxiliary question relative to the main question. The numbers are of course small so it is hard to identify trends within question type or topic.

Table 18. Question Simplification: Counts of Main correct, Aux correct and Percent difference

		Overall			Aux HYP			Aux NPS			Aux VEN		
		main	aux	%	main	aux	%	main	aux	%	main	aux	%
Q type	PURP	5	6	20	0	0	0	0	0	0	5	6	20
	METHOD	39	33	-15	0	0	0	39	33	-15	0	0	0
	CAUSE	21	32	52	21	32	52	0	0	0	0	0	0
	FACT	134	170	27	82	109	33	52	61	17	0	0	0
	TRUE	135	176	30	56	71	27	79	105	33	0	0	0
	Total	334	417	25	159	212	33	170	199	17	5	6	20
Domain	ALZ	0	0	0	0	0	0	0	0	0	0	0	0
	AIDS	145	202	39	102	133	20	43	69	60	0	0	0
	CLIMA	108	118	9	28	43	54	75	69	-8	5	6	20
	MUSIC	81	97	20	29	36	24	52	61	17	0	0	0
	Total	334	417	25	159	212	33	170	199	17	5	6	20

Table 19. Question Simplification: Counts of Questions where Aux better than Main

		All Main-Aux Qs		Aux HYP		Aux NPS		Aux VEN	
		# aux better	total	# aux better	total	# aux better	total	# aux better	total
Q type	PURP	1	1	0	0	0	0	1	1
	METHOD	1	2	0	0	1	2	0	0
	CAUSE	1	1	1	1	0	0	0	0
	FACT	5	6	3	4	2	2	0	0
	TRUE	5	7	2	4	3	3	0	0
	Total	13	17	6	9	6	7	1	1
Domain	ALZ	0	0	0	0	0	0	0	0
	AIDS	6	8	3	5	3	3	0	0
	CLIMA	4	6	2	3	1	2	1	1
	MUSIC	3	3	1	1	2	2	0	0
	Total	13	17	6	9	6	7	1	1

Answer Simplification. The breakdown of scores by answer simplification type is shown in Tables 20 and 21 which are analogous to Tables 18 and 19. Table 20 shows that the overall percentage difference between main and auxiliary in correct answers was 122%, considerably more than for question simplification. Concerning topic, Climate was once again the lowest (18%) and Aids the highest (179%), with Music close behind (167%). Concerning question type, Cause was now the *lowest* at 41% while the highest was True (187%). True (i.e. which-is-true) questions often ask for a difficult choice between statements about the text, statements that can take many different forms. So it is reasonable to expect a big improvement here. Concerning Table 21, this once again shows a trend of increasing scores for the auxiliary questions.

Table 20. Answer Simplification: Counts of Main correct, Aux correct and Percent difference

		Overall			Aux HYP			Aux NPS			Aux VEN		
		main	aux	%	main	aux	%	main	aux	%	main	aux	%
Q type	PURP	0	0	0	0	0	0	0	0	0	0	0	0
	METHOD	83	216	160	0	0	0	46	116	152	37	100	170
	CAUSE	79	111	41	1	2	100	20	16	-20	58	93	60
	FACT	65	130	100	49	87	78	16	43	169	0	0	0
	TRUE	70	201	187	23	29	26	33	69	109	14	103	636
	Total	297	658	122	73	118	62	115	244	112	109	296	172
Domain	ALZ	0	0	0	0	0	0	0	0	0	0	0	0
	AIDS	39	109	179	16	25	56	23	84	265	0	0	0
	CLIMA	94	111	18	57	93	63	37	18	-51	0	0	0
	MUSIC	164	438	167	0	0	0	55	142	158	109	296	172
	Total	297	658	122	73	118	62	115	244	112	109	296	172

Table 21. Answer Simplification: Counts of Questions where Aux better than Main

		All Main-Aux Qs		Aux HYP		Aux NPS		Aux VEN	
		# aux better	total	# aux better	total	# aux better	total	# aux better	total
Q type	PURP	0	0	0	0	0	0	0	0
	METHOD	6	6	0	0	3	3	3	3
	CAUSE	4	5	1	1	0	1	3	3
	FACT	5	6	3	3	2	3	0	0
	TRUE	8	10	2	3	3	4	3	3
	Total	23	27	6	7	8	11	9	9
Domain	ALZ	0	0	0	0	0	0	0	0
	AIDS	4	4	1	1	3	3	0	0
	CLIMA	5	8	5	6	0	2	0	0
	MUSIC	14	15	0	0	5	6	9	9
	Total	23	27	6	7	8	11	9	9

To summarize, the aim was to see if simplifications of a question would increase a system's performance. The indications are that this actually occurred. The implication is that the Auxiliary question approach could be used to dig deeper into the exact workings of a system and in particular the performance of individual components within that system, while keeping with the multiple choice answer format which allows complex questions but still permits automatic evaluation. However, this was a pilot study only and reservations should be noted: Firstly, there were only 44 Auxiliary questions out of 240 Main questions which is only 18.33%, and 44 is not a big enough number to comprise a representative sample; Secondly the distribution of simplification types was not controlled; Thirdly, the exact nature of simplifications was not that closely specified or validated, as it is a very complex matter and this was a small part of the project; Finally, in some cases, the simplification substituted a direct substring of the text, which systems could then match using string processing and hence answer correctly. Such a substitution might possibly not be pinning down the lack of an inference at all but simply turning QA into string comparison.

Subject to the above remarks, this pilot study did seem to identify strong effects and to set out a framework which could be refined in future evaluation frameworks. This suggests a more systematic study and analysis of Question and

Answer simplification in future years, using additional simplification operations, and hence allowing system builders to pinpoint the strengths and weaknesses of their systems.

7.5 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. This year participants were allowed to submit a maximum of 10 runs. Run 01 had to be produced using only the Background Collection provided by the organizations—no external resources were allowed. Participants that did not use the Background Collections submitted their runs starting from number 02. Runs 02 to 10 were permitted to make use of any additional resources. Out of 11 groups, 6 submitted also the first run. These runs can be seen below in Table 22 (extracted from Table 13 above).

Table 22. Scores in '1' runs compared with best other runs

RUN_NAME (...1 Runs)	c@1 on main questions	RUN_NAME (other runs)	c@1 on main questions
nara1301enen	0.28	nara1302enen	0.24
buap1301enen	0.27	buap1309enen	0.28
evma1301enen	0.23	evma1303enen	0.24
vens1301enen	0.22	vens1302enen	0.21
btbn1301bgbg	0.19	btbn1302bgbg	0.17
onto1301bgbg	0.19	ontol1302bgbg	0.18
Average:	0.23		0.22

On the left pair of columns are the submitted '1' runs that were only permitted to use the Background Collection and no other source of knowledge. On the right pair of columns are the best non-'1' runs submitted by the same groups. It is clear that the average c@1 scores for the '1' runs (0.23) and the non-'1' runs (0.22) are very similar. Viewed individually, nara, vens, btbn and ontol were all better when using just the Background Collection, with the biggest difference being nara which scored 0.24 with non-'1' and 0.4 more (0.28) with '1'. nara was also the best scoring run in this group (but not the best overall as some groups did not submit '1' runs) and did seem to gain some benefit from the Background Collection. On the other hand, buap and evma were worse in '1' than non-'1' though the difference was only 0.1 in each case. The differences, whether increases or decreases, are small except for nara, so it is hard to decide whether the Background Collections are beneficial to systems or not.

Generally the use of the Background Collections on the one hand, and how to measure such use on the other hand, remain unanswered questions. Systems could be asked to 'prove' that they have used a background document by for example quoting a supporting passage from it, but it is hard to prevent such use from being reverse engineered once the system has first found the required information elsewhere. In addition, the extraction of simple supporting passages is not the only valid use to which a background collection can be put; the use could be more intangible, such as extracting statistical data or causal rules.

Finally, in considering the '1' runs it is important to remember that the best runs overall by c@1 on main questions were jucs1302enen (0.55) and kule1306enen (0.3), neither of which submitted '1' runs.

7.6 Analysis of Systems

The table in Appendix 3 summarises the set of techniques that participants have reported are being used in their systems. A more detailed explanation of each system is given by participants in the Working Notes.

Most of systems perform question analysis as it was shown in last editions. However, while in the last edition questions patterns were automatically obtained, this year's participants seem to prefer to create patterns manually.

The most common techniques applied for processing texts were, as usual, PoS tagging, NER and dependency analysis in a lower proportion. Nevertheless, participants did not report the application of deeper analysis techniques, except the vens system, which uses semantic role labelling for its logic representation. Therefore, it seems systems continue relying on lexical and simple syntactic analysis, which do not allow all the phenomena in language to be captured and limit the final performance of systems.

8 CONCLUSIONS

The task this year was significantly harder than in previous years, due to the introduction of NCA questions, and questions with modality and negation issues. While this year's results show some improvement compared to first year, especially with 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 achieving this goal.

The NoA option (i.e., system unable to determine with enough confidence an answer) shows an interesting trend. Comparing the overall NoA performance over the three years of QA4MRE. It is quite striking that the percentage of NoA answers returned by systems seems to halve every successive year: from around 40% in 2011 to under 20% to under 10% this year. This may suggest that systems are becoming more confident in their answering ability and hence more reluctant to use NoA answers unless they are sure these are appropriate. But the fact is that in most cases systems increase the proportion of wrong answer when they decide to give answer.

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.

Regarding the second and third issues, results suggest that the use of external resources helps in general to improve results, although not so clearly in the case of Background Collections. Most participants do not seem to know how to gather usable background knowledge from these collections, while it seems that other external resources provide greater benefit. We need to decide whether to continue collection Background Collections, since the organization is spending a lot of resources doing so every year[2]. Somehow, we expected to gather some attention to Open Information Extraction and similar research fields aimed at acquiring knowledge from textual sources to enable textual inferences.

The first question is whether the technology developed so far is just ranking the options or is actually 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 led us to introduce a change this year: an explicit assessment of the ability to reject candidate answers when they are incorrect, using the "None of the answers above are correct" option. Maybe due to the novelty or to a surprise effect, the fact is that systems performed consistently worse over these questions. Given the fact that 39% of questions were of this type, none of the systems except one, was able to achieve this baseline.

It is important to notice the difference between NCA questions and NoA responses. Systems should use NoA response when the risk of choosing a wrong answer is high. In order to choose NCA option as response, systems must be able to find evidences about the incorrectness of the candidate answer. This must lead research towards the development of the ability to reject answers more than the ability to accept them. This is in accordance to the main QA scenario were we expect some hypothesis over-generation that answer validation modules must manage. For this reason, we'll work in future about how the evaluation methodology can reward systems with this desirable feature.

ACKNOWLEDGMENTS

Anselmo Peñas and Alvaro Rodrigo work has been partially supported by the Spanish Government (MINECO) in the framework of CHIST-ERA program (READERS project), and the Regional Government of Madrid, through the project MA2VICMR (S2009/TIC1542). Pamela Forner work has been partially supported by the PROMISE Network of Excellence (258191). Special thanks are due to Giovanni Moretti (CELCT, Trento, Italy) for the technical support in the management of all data and evaluation scripts of the campaign.

[2] Note Google's API wasn't available for research purposes. This significantly increased our collection work.

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.

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

Run	C @ 1 ALL QUESTIONS	C@ 1 topic 1	C@ 1 topic 2	C@1 topic 3	C@ 1 topic 4
jucs1302enen	0,59	0,68	0,73	0,30	0,66
kule1304enen	0,35	0,33	0,39	0,35	0,33
kule1305enen	0,35	0,33	0,37	0,35	0,35
kule1302enen	0,34	0,33	0,38	0,32	0,31
buap1301enen	0,33	0,29	0,35	0,31	0,37
kule1303enen	0,33	0,29	0,39	0,32	0,31
kule1306enen	0,33	0,37	0,33	0,35	0,26
nara1301enen	0,33	0,36	0,37	0,32	0,25
buap1304enen	0,32	0,32	0,42	0,31	0,22
buap1305enen	0,32	0,27	0,39	0,36	0,25
kule1308enen	0,32	0,31	0,36	0,35	0,26
buap1303enen	0,31	0,33	0,36	0,30	0,24
buap1309enen	0,31	0,27	0,33	0,30	0,33
kule1309enen	0,31	0,28	0,33	0,34	0,26
buap1306eses	0,30	0,29	0,39	0,24	0,27
csui1302enen	0,30	0,33	0,39	0,22	0,27
csui1304enen	0,30	0,33	0,38	0,21	0,28
csui1305enen	0,30	0,36	0,38	0,22	0,23
csui1307enen	0,30	0,33	0,36	0,25	0,25
kule1307enen	0,30	0,26	0,37	0,33	0,24
nara1302enen	0,30	0,29	0,39	0,28	0,23
csui1303enen	0,29	0,30	0,38	0,21	0,25
csui1306enen	0,29	0,34	0,35	0,23	0,25
csui1308enen	0,29	0,33	0,36	0,23	0,21
csui1309enen	0,29	0,33	0,34	0,23	0,26
buap1307eses	0,28	0,31	0,32	0,22	0,29
buap1308eses	0,28	0,31	0,29	0,23	0,29
csui1310enen	0,28	0,33	0,33	0,22	0,25
lims1304enen	0,28	0,25	0,33	0,24	0,28
lims1305enen	0,28	0,28	0,29	0,24	0,31
onto1305bgbg	0,28	0,26	0,21	0,32	0,33
AVERAGE	0,28	0,28	0,31	0,26	0,26
MEDIAN	0,28	0,29	0,33	0,25	0,26
onto1307bgbg	0,27	0,26	0,22	0,29	0,33
onto1308bgbg	0,27	0,25	0,22	0,28	0,32
evma1303enen	0,26	0,29	0,30	0,26	0,18

lims1302enen	0,26	0,27	0,31	0,22	0,24
lims1303enen	0,25	0,18	0,28	0,28	0,25
onto1306bgbg	0,25	0,23	0,19	0,28	0,31
uaic1303roro	0,25	0,28	0,22	0,22	0,28
buap1302enen	0,24	0,23	0,32	0,24	0,15
buap1310enen	0,24	0,22	0,32	0,26	0,14
evma1301enen	0,24	0,24	0,26	0,26	0,21
vens1301enen	0,24	0,28	0,31	0,18	0,17
vens1302enen_ LATE_RUN	0,24	0,28	0,29	0,16	0,22
uaic1302roro	0,23	0,31	0,14	0,19	0,28
uaic1306roro	0,23	0,31	0,14	0,19	0,28
onto1301bgbg	0,22	0,12	0,17	0,29	0,30
onto1302bgbg	0,22	0,12	0,17	0,28	0,29
onto1303bgbg	0,22	0,14	0,23	0,24	0,27
onto1304bgbg	0,22	0,13	0,23	0,24	0,26
btbn1301bgbg	0,19	0,22	0,17	0,19	0,18
evma1302enen	0,19	0,19	0,26	0,20	0,11
btbn1302bgbg	0,18	0,23	0,18	0,12	0,19
uaic1304roro	0,17	0,18	0,16	0,24	0,08
uaic1305roro	0,17	0,19	0,17	0,24	0,07
STANDARD DEVIATION	0,06	0,08	0,10	0,05	0,08

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

Run	c@1	C@1 r_1	C@1 r_2	C@1 r_3	C@1 r_4	C@1 r_5	C@1 r_6	C@1 r_7	C@1 r_8	C@1 r_9	C@1 r_10	C@1 r_11	C@1 r_12	C@1 r_13	C@1 r_14	C@1 r_15	C@1 r_16
jucs1302enen	0,59	0,80	0,83	0,60	0,48	0,65	0,56	0,88	0,80	0,39	0,26	0,19	0,37	0,43	0,68	0,86	0,65
kule1304enen	0,35	0,45	0,33	0,28	0,27	0,28	0,47	0,35	0,47	0,49	0,29	0,28	0,33	0,29	0,31	0,33	0,37
kule1305enen	0,35	0,40	0,33	0,33	0,27	0,25	0,47	0,35	0,42	0,50	0,33	0,28	0,30	0,33	0,33	0,33	0,39
kule1302enen	0,34	0,47	0,33	0,27	0,27	0,25	0,47	0,35	0,47	0,44	0,28	0,28	0,30	0,28	0,28	0,33	0,33
buap1301enen	0,33	0,23	0,24	0,14	0,56	0,26	0,29	0,45	0,37	0,33	0,22	0,43	0,28	0,37	0,44	0,22	0,43
kule1303enen	0,33	0,48	0,30	0,16	0,21	0,29	0,47	0,39	0,45	0,35	0,26	0,29	0,36	0,31	0,26	0,31	0,37
kule1306enen	0,33	0,47	0,27	0,40	0,33	0,25	0,21	0,35	0,53	0,39	0,22	0,39	0,40	0,17	0,39	0,22	0,28
nara1301enen	0,33	0,57	0,17	0,23	0,43	0,20	0,32	0,50	0,47	0,35	0,26	0,33	0,32	0,17	0,17	0,29	0,39
buap1304enen	0,32	0,53	0,33	0,07	0,33	0,45	0,37	0,45	0,42	0,28	0,33	0,39	0,25	0,22	0,22	0,22	0,22
buap1305enen	0,32	0,30	0,36	0,08	0,36	0,35	0,39	0,47	0,33	0,35	0,31	0,43	0,35	0,34	0,25	0,18	0,26
kule1308enen	0,32	0,42	0,27	0,30	0,24	0,30	0,22	0,37	0,55	0,35	0,26	0,43	0,35	0,18	0,35	0,22	0,28
buap1303enen	0,31	0,47	0,27	0,20	0,40	0,35	0,37	0,35	0,37	0,28	0,33	0,28	0,30	0,28	0,22	0,22	0,22
buap1309enen	0,31	0,27	0,33	0,20	0,27	0,25	0,37	0,40	0,32	0,28	0,22	0,39	0,30	0,22	0,50	0,17	0,44
kule1309enen	0,31	0,40	0,27	0,27	0,20	0,25	0,21	0,35	0,53	0,33	0,28	0,39	0,35	0,22	0,33	0,22	0,28
buap1306eses	0,30	0,33	0,23	0,20	0,40	0,40	0,37	0,50	0,32	0,12	0,35	0,17	0,32	0,19	0,39	0,18	0,33
csui1302enen	0,30	0,51	0,33	0,20	0,28	0,40	0,23	0,42	0,52	0,22	0,07	0,29	0,28	0,28	0,41	0,26	0,08
csui1304enen	0,30	0,51	0,33	0,20	0,28	0,35	0,23	0,42	0,52	0,22	0,07	0,25	0,28	0,28	0,41	0,26	0,14
csui1305enen	0,30	0,51	0,38	0,21	0,36	0,35	0,20	0,42	0,52	0,19	0,07	0,28	0,31	0,23	0,34	0,26	0,08
csui1307enen	0,30	0,45	0,27	0,25	0,34	0,25	0,39	0,30	0,49	0,26	0,14	0,31	0,26	0,16	0,39	0,20	0,22
kule1307enen	0,30	0,37	0,23	0,19	0,24	0,30	0,22	0,40	0,55	0,35	0,26	0,35	0,36	0,12	0,31	0,23	0,28
nara1302enen	0,30	0,36	0,25	0,16	0,36	0,32	0,37	0,45	0,42	0,23	0,32	0,23	0,32	0,11	0,17	0,29	0,33
csui1303enen	0,29	0,45	0,28	0,20	0,28	0,30	0,33	0,42	0,49	0,22	0,07	0,25	0,28	0,23	0,35	0,25	0,15
csui1306enen	0,29	0,48	0,28	0,25	0,34	0,23	0,37	0,31	0,49	0,26	0,14	0,26	0,26	0,16	0,37	0,20	0,22
csui1308enen	0,29	0,48	0,33	0,23	0,28	0,31	0,35	0,36	0,41	0,25	0,07	0,32	0,28	0,17	0,20	0,23	0,22
csui1309enen	0,29	0,48	0,28	0,21	0,34	0,23	0,33	0,33	0,44	0,23	0,14	0,26	0,28	0,15	0,37	0,26	0,22
buap1307eses	0,28	0,33	0,20	0,38	0,33	0,25	0,32	0,40	0,32	0,06	0,33	0,28	0,20	0,44	0,33	0,22	0,17
buap1308eses	0,28	0,33	0,20	0,38	0,33	0,25	0,32	0,35	0,26	0,06	0,33	0,28	0,25	0,44	0,33	0,22	0,17
csui1310enen	0,28	0,48	0,28	0,21	0,34	0,23	0,29	0,33	0,44	0,19	0,14	0,26	0,28	0,15	0,32	0,26	0,22
lims1304enen	0,28	0,33	0,20	0,27	0,20	0,20	0,37	0,40	0,37	0,28	0,28	0,17	0,25	0,33	0,33	0,17	0,28
lims1305enen	0,28	0,33	0,27	0,20	0,33	0,25	0,26	0,35	0,32	0,22	0,28	0,22	0,25	0,33	0,39	0,22	0,28
onto1305bgbg	0,28	0,28	0,32	0,00	0,40	0,00	0,29	0,07	0,44	0,32	0,43	0,27	0,28	0,35	0,29	0,35	0,31
AVERAGE	0,28	0,38	0,27	0,20	0,26	0,24	0,29	0,32	0,37	0,26	0,24	0,25	0,29	0,24	0,30	0,25	0,24
MEDIAN	0,28	0,38	0,27	0,20	0,27	0,25	0,29	0,35	0,37	0,27	0,26	0,26	0,28	0,22	0,30	0,23	0,22
onto1307bgbg	0,27	0,33	0,28	0,07	0,33	0,21	0,11	0,25	0,32	0,28	0,39	0,25	0,25	0,35	0,28	0,33	0,35
onto1308bgbg	0,27	0,33	0,27	0,07	0,33	0,20	0,11	0,25	0,32	0,28	0,39	0,22	0,25	0,33	0,28	0,33	0,33
evma1303enen	0,26	0,48	0,14	0,19	0,34	0,29	0,24	0,26	0,37	0,18	0,18	0,25	0,44	0,07	0,29	0,22	0,12
lims1302enen	0,26	0,33	0,27	0,27	0,20	0,20	0,32	0,35	0,37	0,22	0,28	0,11	0,25	0,28	0,28	0,17	0,22
lims1303enen	0,25	0,33	0,20	0,20	0,00	0,15	0,37	0,35	0,26	0,33	0,28	0,17	0,35	0,28	0,33	0,22	0,17
onto1306bgbg	0,25	0,27	0,27	0,00	0,40	0,05	0,26	0,05	0,42	0,28	0,39	0,22	0,25	0,33	0,28	0,33	0,28
uaic1303roro	0,25	0,47	0,21	0,23	0,21	0,15	0,16	0,17	0,45	0,17	0,28	0,25	0,20	0,39	0,17	0,23	0,33
buap1302enen	0,24	0,20	0,27	0,20	0,27	0,30	0,37	0,35	0,26	0,17	0,28	0,28	0,25	0,22	0,11	0,11	0,17
buap1310enen	0,24	0,13	0,20	0,27	0,27	0,30	0,26	0,45	0,26	0,17	0,22	0,28	0,35	0,17	0,06	0,17	0,17
evma1301enen	0,24	0,34	0,23	0,20	0,19	0,34	0,27	0,07	0,33	0,39	0,12	0,06	0,46	0,14	0,25	0,33	0,06
vens1301enen	0,24	0,33	0,33	0,33	0,13	0,32	0,26	0,20	0,47	0,17	0,06	0,28	0,23	0,12	0,17	0,11	0,28
vens1302enen_LATE_RUN	0,24	0,40	0,27	0,33	0,13	0,30	0,32	0,10	0,47	0,22	0,00	0,17	0,25	0,28	0,28	0,11	0,22
uaic1302roro	0,23	0,51	0,28	0,22	0,20	0,08	0,20	0,08	0,19	0,07	0,21	0,17	0,29	0,25	0,19	0,25	0,39

uaic1306roro	0,23	0,51	0,28	0,22	0,20	0,08	0,20	0,08	0,19	0,07	0,21	0,17	0,29	0,25	0,19	0,25	0,39
onto1301bgbg	0,22	0,21	0,14	0,07	0,07	0,05	0,05	0,30	0,28	0,29	0,44	0,18	0,25	0,22	0,59	0,28	0,12
onto1302bgbg	0,22	0,20	0,13	0,07	0,07	0,05	0,05	0,30	0,26	0,28	0,44	0,17	0,25	0,22	0,56	0,28	0,11
onto1303bgbg	0,22	0,20	0,21	0,07	0,07	0,25	0,16	0,30	0,21	0,17	0,39	0,17	0,25	0,17	0,50	0,28	0,12
onto1304bgbg	0,22	0,20	0,20	0,07	0,07	0,25	0,16	0,30	0,21	0,17	0,39	0,17	0,25	0,17	0,50	0,28	0,11
btbn1301bgbg	0,19	0,33	0,33	0,00	0,20	0,15	0,21	0,25	0,05	0,22	0,11	0,33	0,10	0,22	0,17	0,17	0,17
evma1302enen	0,19	0,17	0,23	0,20	0,18	0,34	0,33	0,00	0,32	0,19	0,12	0,00	0,46	0,07	0,18	0,19	0,00
btbn1302bgbg	0,18	0,27	0,20	0,33	0,13	0,05	0,32	0,25	0,11	0,28	0,17	0,06	0,00	0,11	0,28	0,17	0,22
uaic1304roro	0,17	0,38	0,20	0,08	0,07	0,07	0,22	0,25	0,11	0,28	0,22	0,11	0,35	0,07	0,06	0,12	0,06
uaic1305roro	0,17	0,42	0,17	0,10	0,00	0,07	0,27	0,29	0,00	0,32	0,19	0,07	0,36	0,08	0,00	0,15	0,07
STANDARD DEVIATION	0,06	0,12	0,10	0,11	0,12	0,12	0,10	0,14	0,14	0,10	0,11	0,10	0,07	0,10	0,13	0,11	0,12

APPENDIX 3: SYSTEM DESCRIPTIONS

Table 23. 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	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		x			x			x				x										
buap	x				x							x										Stemmer, lemmatisation
csui			x		x	x	x	x		x												Anaphora resolution
evma		x			x			x		x			x									
jucs			x		x	x	x	x	x	x		x	x									
kule	x				x								x									
limsi		x		Machine learning	x	x					x	x		x			x					
nara		x			x			x														Anaphora Resolution
onto			x				x				x											Stemming, stop-word filtering
uaic		x					x	x	x													
vens				deep question linguistic processing	x	x		x				x	x			x			x			

Table 24. 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
btb		x					x			x			x					x	x	x
buap							x									x		x	x	
csui		x															x	x	x	
evma							x			x							x	x	x	
jucs	x						x								x		x	x	x	
kule							x												x	
limsi		x																x	x	
nara	x																x	x		
onto						x				x										x
uaic					x										x	x	x	x		
vens		x	x					x		x						x	x	x	x	x

Table 25. 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					x			The system search for answers on the basis of dependency triples.
buap					x	x	x		The system uses information retrieval techniques and a graph based representation to find similarity features vector between the answers and support text extracted from the documents.
csui	x								The system performs two different approaches to determine the answer of a question. For factoid question, the system will use the list of named entities obtained from the passage that is relevant to the query.
evma				x					This year we are testing a new system. The approach is similar to that used in the previous year, based on superficial analysis of the text, supplemented with POS and NER.
jucs	x				x	x			The system used textual entailment based answer validation technique. Per topic, one multi-document summary was generated form the background collection provided by the organiser. Then the summary is also used to select the correct answer.
kule						x			Simple system employing set similarity metrics.
limsi					x	x			
nara		x							This system uses a combination of several lexical similarity features, with weights trained using thresholded minimum error rate training.
onto	x				x				The system relies on direct term matching. Some transformations are applied on the text, question and answer strings prior to analysis. These involve stemming, stop-word filtering, and enrichment of sentences with synonyms and paraphrases.
uaic	x								The system is based on previous year's system and additionally we use a Coreference Resolver.
vens					x		x		Specialized version of GETARUNS: it does complete semantic analysis but uses a less restricted version of the parser. Semantic relations are not totally transformed, only predicate argument and modifier relations are memorized in the discourse model.