Overview of QA4MRE 2013 Entrance Exams Task

Anselmo Peñas[1], Yusuke Miyao[2], Eduard Hovy[3], Pamela Forner[4] and Noriko Kando[2]

[1] NLP&IR group, UNED, Spain (anselmo@lsi.uned.es)
[2] National Institute of Informatics, Japan {yusuke,kando}@nii.ac.jp
[3] Carnegie Mellon University, USA (hovy@cmu.edu)
[4] CELCT, Italy (forner@celct.it)

Abstract. This paper describes the Question Answering for Machine Reading (QA4MRE) Entrance Exams at the 2013 Cross Language Evaluation Forum. The data set of this task is extracted from actual university entrance examinations as-is, and therefore includes a variety of topics in daily life. Another unique feature of the Entrance Exams task is that questions are designed originally for testing human examinees, rather than evaluating computer systems. Therefore, the data set is expected to have a natural distribution of human ability for reading and understanding texts.

1 INTRODUCTION

The Entrance Exams task at CLEF 2013 QA4MRE is focused on solving Reading Comprehension tests of English examinations. 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.

In QA4MRE, tests are usually made in an artificial way by organizers, in order to test properly systems performance on a controlled set of question types and a defined level of inference.

In such scenarios, the question arises how the performance of systems on artificial tests compares to their performance when confronted with real human tests. We believe that finding a real benchmark able to test real systems performance over the time offers great value to assess real progress in the field along the future years.

With this goal in mind, CLEF and NTCIR started collaboration around the idea of testing systems against University Entrance Exams, the same exams humans have to pass to enter University. The data set was prepared and distributed by NTCIR, while other organization efforts, including announcements, collecting and evaluating submissions, etc. were managed by CLEF. This style of the organization reduced the workload of each side, since the NTCIR side is already familiar with the contents of the data and its copyright issues, while the CLEF side has already established other organization processes such as submission management and evaluation. The success of this coordination also owes to the standard data format and evaluation methodology, which were also adopted for this pilot task. The next round of this task is expected to be organized in a similar manner.

2 TASK DESCRIPTION

The form of the task is essentially the same as the QA4MRE Main Task. Participant systems are asked to read a given document and answer questions. Questions are given in multiple-choice format, with several options from which a single answer must be selected.

A crucial difference from the other QA4MRE tasks is that background text collections are not provided. Systems have to answer questions by referring to "common sense knowledge" that high school students who aim to enter the university are expected to have. Another important difference is that we do not intend to restrict question types. Any types of reading comprehension questions in real entrance exams will be included in the test data.

3 DATA

Japanese 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 that humans are evaluated to enter the University. In this first campaign we reduced the challenge to Reading Comprehension exercises contained in the English exams.

The data set is extracted from standardized English examinations for university admission in Japan. Exams are created by the Japanese National Center for University Admissions Tests.

Original examinations include various styles of questions, such as word filling, grammatical error recognition, sentence filling, etc.

One of such styles is reading comprehension; a test provides a text that describes some daily life situation, and questions about the text are asked. Since this type of questions is suitable for the QA4MRE lab, we extracted questions of this type automatically from XML files of the examination data, and converted the XML annotations to fit the standard format of QA4MRE.

For each examination, one text is given, and five questions on the given text are asked. Each question has four choices. For this year campaign, we selected 10 examinations, one of which was delivered as development data while the others were provided as final test data. That is, we provided 9 documents, 46 questions[1] and 184 choices.

4 EVALUATION

Scoring of the output produced by participant systems was performed automatically by comparing the answers of systems against the gold standard collection with annotations made by humans. No manual assessment was performed.

Each test receives an evaluation score between 0 and 1 using c@1 [1]. This measure, used in previous CLEF QA Tracks, encourages systems to reduce the number of incorrect answers while maintaining the number of correct ones by leaving some questions unanswered. Systems received evaluation scores from two different perspectives:

1. **At the question-answering level:** correct answers are counted individually without grouping them
2. **At the reading-test level:** figures both for each reading test as a whole are given.

5 RESULTS

During registration, 27 different groups showed interest in the task. Out of them, 10 groups fulfilled the data agreements, and finally, only 5 teams submitted runs. Despite their interest in the task, some groups expressed that the difficulty of the tests exceeded the current state of the art in the field and decided not to participate. Table 1 enumerates the participating groups and their reference paper in CLEF 2013 Working Notes.

[1] One test document was accompanied with 6 questions exceptionally.

Table 1. Participants and reference papers

NIIJ	National Institute of Informatics, Japan	Li et al. 2013 [2]
JUCS	Jadavpur University, India	Banerjee et al. 2013 [3]
NARA	Nara Institute of Science and Technology, Japan	Arthur et al. 2013 [4]
CMU	Carnegie Mellon University, United States	-
LIMS-CNRS	ILES – LIMSI, France	-

Results are summarized in Tables 2 and 3 for the QA and for Reading perspectives respectively. According to Table 2, the system with higher score (jucs [3]) is the one that answered incorrectly less questions. It is also the unique system that answered more questions correctly than incorrectly, finding a better balance with leaving some questions unanswered. This indicates that their modules to detect whether they have enough evidence about the correctness of the answer are working pretty well.

Table 2. Overall results for all runs, QA perspective

RUN NAME	**C@1**	**# of questions ANSWERED**			**# of questions UNANSWERED**
		RIGHT	**WRONG**	**Total**	
jucs	0.42	13	10	23	23
NIIJ-3	035	16	30	46	0
NIIJ-5	0.33	15	31	46	0
NIIJ-4	0.25	8	19	27	19
Random	0.25	12	34	46	0
NIIJ-2	0.24	11	35	46	0
lims-cnrs-1	0.24	11	35	46	0
NIIJ-1	0.22	7	17	24	22
nara	0.22	10	35	45	1
lims-cnrs-2	0.20	9	37	46	0
cmu	0.10	4	33	37	9

Table 3 shows results under the reading perspective. First column corresponds to systems run id, second column to the overall c@1 obtained, third column shows the number of tests that the systems have passed if we consider the threshold of 0.5, and the rest of columns correspond to the c@1 value for each particular test.

Table 3. Overall results for all runs, reading perspective

Run	Over-all	Pass	T1	T2	T3	T4	T5	T6	T7	T8	T9
jucs	0.42	4/9	0.00	0.25	0.24	0.72	0.28	0.64	0.00	0.64	0.84
NIIJ-3	0.35	3/9	0.40	0.50	0.60	0.20	0.20	0.40	0.60	0.20	0.00
NIIJ-5	0.33	2/9	0.20	0.67	0.40	0.20	0.00	0.40	0.60	0.20	0.20
AVERAGE	0.26	-	0.21	0.36	0.28	0.23	0.13	0.27	0.29	0.18	0.27
NIIJ-4	0.25	0/9	0.24	0.39	0.28	0.28	0.24	0.28	0.32	0.00	0.00
RANDOM	0.25	-	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
NIIJ-2	0.24	1/9	0.20	0.67	0.40	0.20	0.00	0.00	0.20	0.00	0.40
lims-cnrs-1	0.24	0/9	0.40	0.17	0.20	0.00	0.20	0.40	0.40	0.20	0.40
MEDIAN	0.24	-	0.21	0.32	0.28	0.20	0.10	0.34	0.26	0.10	0.24
NIIJ-1	0.22	0/9	0.28	0.58	0.28	0.28	0.00	0.00	0.00	0.00	0.28
nara	0.22	1/9	0.22	0.00	0.40	0.20	0.00	0.20	0.20	0.60	0.20
lims-cnrs-2	0.20	0/9	0.20	0.17	0.00	0.20	0.40	0.40	0.40	0.00	0.00
cmu	0.10	0/9	0.00	0.19	0.00	0.00	0.00	0.00	0.20	0.00	0.40

JUCS [3] report very good results using a system based on Textual Entailment and answer ranking. One particularity of this system is that it only answered 23 questions out of the 46. From these 13 were right and 10 wrong. This strategy is rewarded by c@1, since that provides partial credit when no answer is given instead of an incorrect one. It is worth noticing the difference in score among different tests. In particular, authors report that the difference depends on the type of questions of tests 1 and 7.

The NIIJ system [2] also performed above average and random baseline. It is also based on Textual Entailment after combining relevant sentences, questions, and answers. In their case, the best run answered all questions, being 16 correct answers and 30 incorrect ones.

Results also show that systems based only on statistical analysis of words alone can’t perform the kind of inferences required to solve the tests.

6 CONCLUSIONS

The dataset together with results suggest something very interesting: the need to develop strategies to reject answers more than strategies to accept answers. In one hand, the dataset shows that in some cases, the way to select the correct answer is by discarding the other candidates. In the other hand, most systems still select more incorrect answers than correct ones, while a measure of progress in systems development is, precisely, the reduction in selecting wrong answers.

The Entrance Exams task shows that Question Answering is a task far from being solved. This is true even for the simplified scenario where only one text is given and a set of options are provided as candidate answers to the question.

Results also show that systems based only on statistical analysis of words alone can't perform the kind of inferences required to solve the tests. In other words, that systems based only on textual similarity can't address the challenge.

Finally, we think that Entrance Exams provides a real benchmark able to assess real progress in the field along future years.

7 ACKNOWLEDGEMENTS

The collaboration has been developed in the framework of Todai Robot Project in Japan, and the CHIST-ERA Readers project in Europe. The Todai Robot Project is a grand challenge headed by NII, and aims to develop an end-to-end AI system that can solve real entrance examinations of universities in Japan integrating heterogeneous AI technologies, such as natural language processing, situation understanding, math formula processing or vision processing.

8 REFERENCES

1. 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, 2011
2. Xinjian Li, Tian Ran, Ngan L.T. Nguyen, Yusuke Miyao and Akiko Aizawa. Question Answering System for Entrance Exams in QA4MRE. *CLEF 2013 Working Notes*, 2013
3. Somnath Banerjee, Pinaki Bhaskar, Partha Pakray, Sivaji Bandyopadhyay and Alexander Gelbukh. Multiple Choice Question (MCQ) Answering System for Entrance Examination. *CLEF 2013 Working Notes*, 2013
4. Philip Arthur, Graham Neubig, Sakriani Sakti, Tomoki Toda and Satoshi Nakamura. NAIST at the CLEF 2013 QA4MRE Pilot Task. *CLEF 2013 Working Notes*, 2013