Overview of CLEF QA Entrance Exams Task 2014

Anselmo Peñas[1], Yusuke Miyao[2], Álvaro Rodrigo[1], Eduard Hovy[3] and Noriko Kando[2]

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

Abstract. This paper describes the Entrance Exams task at the CLEF QA Track 2014. Following 2013 edition, the data set has been extracted from actual university entrance examinations including a variety of topics and question types. Systems receive a set of Multiple-Choice Reading Comprehension tests where the task is to select the correct answer among a finite set of candidates, according to the given text. Questions are designed originally for testing human examinees, rather than evaluating computer systems. Therefore, the data set challenges human ability to show their understanding of texts. Thus, questions and answers are lexically distant from their supporting excerpts in text, requiring not only a high degree of textual inference, but also the development of strategies for selecting the correct answer. As a novelty this year, data sets originally in English were manually translated into Russian, French, Spanish and Italian.

1 INTRODUCTION

Following 2013 edition, the Entrance Exams task at CLEF QA Track 2014 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. Despite the difficulty of the challenge, we believe we are building a real benchmark that will serve to measure real progress in the field during the next years.

With this goal in mind, CLEF and NTCIR started collaboration in 2013 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. The success of this coordination also owes to the standard data format and evaluation methodology followed in past editions.

2 TASK DESCRIPTION

Participant systems are asked to read a given document and answer a set of questions. Questions are given in multiple-choice format, with several options from which a single answer must be selected. 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 type 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.

3.1 Sources

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. As in the previous edition, we reduced the challenge to these Reading Comprehension exercises contained in the English exams, leaving other types of exercises available for future tasks.

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 reused as development data 12 examinations from last year campaign. Besides, we provided new 12 documents, 60 questions and 240 candidate answers to be validated.

3.2 Languages

As a novelty this year, data sets for development and testing originally in English were manually translated into Russian, French, Spanish and Italian. They are parallel translations of texts, questions and candidate answers.

In addition to the official data, we collected four more unoffcial translations into French. Despite they preserve original meaning, each translation has its particularities that produce different effects on systems performance: text simplification, lexical variation, different uses of anaphora, overall quality, etc. This data is extremely useful to get insights about systems and their level of inference. Synapse [3] reports some initial experiments with this unofficial data.

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 for each reading test as a whole are given. A test is considered to be passed if a c@1 score above 0.5 is reached. Then, the proportion of tests that are passed is given as a global score.

5 RESULTS

Table 1 enumerates the participating groups and their reference paper in CLEF 2014 Working Notes. Only LIMSI-CNRS has participated in the two editions and only one team (Synapse) has participated in a second language different than English (French).

Table 1. Participants and reference papers

SYNAPSE	Synapse Développement, France	Laurent et al. 2014 [2] [3]
DIPF	Technische Universität Darmstadt, Germany	Dhruva et al. 2014 [4]
CICNLP	Centro de Investigación en Computación Instituto Politécnico Nacional, Mexico	Gómez-Adorno et al. 2014 [5]
CSGS	Saarland University, Germany	Ostermann et al. 2014 [6]
LIMSI-CNRS	ILES – LIMSI, France	Gleize et al. 2014 [7]

Results are summarized in Tables 2 and 3 for the QA and for Reading perspectives respectively.

Table 2. Overall results for all runs, QA perspective

		# of questions ANSWERED				# of questions
RUN NAME	**C@1**	**RIGHT**	**WRONG**	**TOTAL**	**Prec.**	**UNANSWERED**
Synapse-French	0.59	33	23	56	0.59	0
Synapse-English	0.45	25	31	56	0.45	0
DIPF-7	0.38	21	35	56	0.38	0
cicnlp-8	0.38	21	35	56	0.38	0
cicnlp-7	0.36	20	36	56	0.36	0
csgs-1	0.36	20	36	56	0.36	0

csgs-2	0.36	16	25	41	0.39	15
cicnlp-2	0.34	19	37	56	0.34	0
cicnlp-3	0.3	17	39	56	0.30	0
cicnlp-4	0.3	17	39	56	0.30	0
cicnlp-1	0.29	16	40	56	0.29	0
cicnlp-6	0.29	16	40	56	0.29	0
DIPF-5	0.29	16	40	56	0.29	0
DIPF-3	0.29	16	40	56	0.29	0
LIMSI-4-HR	0.25	11	30	41	0.27	15
LIMSI-7	0.25	10	25	35	0.29	21
LIMSI-4	0.25	14	42	56	0.25	0
DIPF-6	0.25	14	42	56	0.25	0
Random	0.25	14	42	56	0.25	0
LIMSI-2-Inv	0.23	13	43	56	0.23	0
DIPF-4	0.23	13	43	56	0.23	0
cicnlp-5	0.23	13	43	56	0.23	0
LIMSI-2	0.2	7	16	23	0.30	33
DIPF-2	0.2	11	45	56	0.20	0
DIPF-1	0.2	11	45	56	0.20	0
LIMSI-1-dude1	0.18	10	46	56	0.18	0
LIMSI-3	0.16	6	20	26	0.23	30
LIMSI-5	0.15	6	28	34	0.18	22
LIMSI-1-dude	0.13	6	38	44	0.14	12
LIMSI-6	0.06	2	16	18	0.11	38

According to Table 2, the system with higher score (Synapse for French [3]) is the unique system that answered more questions correctly than incorrectly. Only few runs made use of the leaving questions unanswered option. In these cases, despite some systems reduced considerably the amount of incorrect answers, none of them could improve their overall c@1 score.

Table 3 shows results for 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	c@1	Pass	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24
Synapse-1Fr.	0.59	9/12	0.5	0.5	**0.75**	**0.83**	**0.67**	**0.8**	**1**	0.6	0.4	**0.4**	0.2	0.6
Synapse-1En.	0.45	7/12	0.5	0.5	0.5	0.67	0.33	0.6	0.33	0.4	**0.6**	0	0.2	0.6
DIPF-7	0.38	6/12	0.25	0.5	0.5	0.5	0.33	0.6	0	0.6	**0.6**	0	0	0.4
cicnlp-8	0.38	3/12	0.75	0.33	0.25	0.67	0.33	0.6	0.33	0.2	0.2	**0.4**	0.2	0.2
cicnlp-7	0.36	2/12	**1**	0.33	0.25	0.67	0.33	0.4	0.33	0.2	0	0.2	0.2	0.4
csgs-1	0.36	3/12	0.5	0.17	**0.75**	0.33	0	**0.8**	0	0.4	0.4	0.2	0.2	0.4
csgs-2	0.36	4/12	0.5	0	**0.75**	0.39	0	**0.8**	0	0.56	0.4	0	0	0.28
cicnlp-2	0.34	2/12	0.25	0.33	**0.75**	0.67	0.33	0.2	0	0.4	0	0.2	0.4	0.4
cicnlp-3	0.3	4/12	0.5	**0.67**	**0.75**	0.67	0	0	0	0	0.2	0	0.4	0.2
cicnlp-4	0.3	2/12	0.25	0.33	**0.75**	0.67	0	0.2	0	0.4	0	0	0.4	0.4
cicnlp-1	0.29	3/12	0.25	0.5	**0.75**	0.67	0	0	0	0	0	0.2	0.4	0.4

cicnlp-6	0.29	3/12	0.5	0.17	0.25	0.5	0.33	0.2	0	0.2	0.2	0	**0.6**	0.4
DIPF-5	0.29	3/12	0.25	0.5	0.25	0.5	**0.67**	0.4	0	0.4	0.2	0	0	0.2
DIPF-3	0.29	4/12	0	0.17	0.25	0.5	**0.67**	0.2	0	0.4	**0.6**	0	0	0.6
Average	0.27	-	0.24	0.26	0.37	0.42	0.23	0.34	0.07	0.29	0.27	0.12	0.18	0.31
LIMSI-4-HR	0.25	2/12	0	0.17	0.62	0	0.56	0.24	0	0.32	0.4	0	0.48	0.28
LIMSI-7	0.25	0/12	0	0	0.31	0.44	0	0.24	0	0.4	0.4	0.24	0.24	0
LIMSI- 4	0.25	2/12	0	0.17	0.5	0.5	0.33	0.2	0	0	0.4	0.2	0.4	0.2
DIPF-6	0.25	1/12	0.25	0.33	0.25	0.33	**0.67**	0.4	0	0.2	0	0.2	0	0.4
Median	0.25	-	0.25	0.33	0.25	0.5	0.33	0.24	0	0.2	0.24	0	0.2	0.4
Random	0.25	-	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
LIMSI-2Inv.	0.23	3/12	0	0	0	0	0	0.2	0	**0.8**	**0.6**	0	**0.6**	0.4
DIPF-4	0.23	1/12	0	0.33	0	0.17	0	**0.8**	0	0.2	0.4	0.2	0	0.4
cicnlp-5	0.23	1/12	0	0.17	0.25	0.67	0	0.4	0	0.4	0	0.2	0	0.4
LIMSI-2	0.2	2/12	0	0.5	0	0.67	0	0.28	0	0	0.36	0	0	0
DIPF-2-	0.2	0/12	0.25	0.33	0.25	0.33	0.33	0	0	0.2	0.4	0	0	0.2
DIPF-1-	0.2	0/12	0.25	0	0.25	0.33	0.33	0.4	0	0.2	0.4	0	0	0.2
LIMSI-1-dude1	0.18	1/12	0	0.17	0.5	0	0.33	0.2	0	0.2	0.2	0	0.2	0.4
LIMSI-3	0.16	1/12	0	0.5	0	0.44	0	0.28	0	0	0	0.24	0	0
LIMSI-5	0.15	0/12	0	0	0.31	0	0	0.24	0	0.2	0.2	0.24	0.24	0
LIMSI-1-dude	0.13	1/12	0.25	0	0	0	0	0.24	0	0.2	0	0.24	0	**0.64**
LIMSI-6	0.06	0/12	0	0	0	0	0	0	0	0.32	0.24	0	0	0

The results observed under the reading perspective are very encouraging. The three top systems were able to pass at least half of tests. As observed in Table 4, each test has a different degree of difficulty for the systems. There are three main reasons for that: the way the questions are formulated, the lexical gap between the text and the candidate answers, and the inherit difficulty of some questions for which wrong candidate answers seems to be closer to the supporting text in a light reading.

Table 4. Number of runs (only for English) that passed each test (out of 28), and maximum c@1 score achieved per test.

	T13	**T14**	**T15**	**T16**	**T17**	**T18**	**T19**	**T20**	**T21**	**T22**	**T23**	**T24**
# Runs pass	7	7	11	14	4	6	0	3	4	0	2	3
Max. score	1	0.67	0.75	0.67	0.67	0.8	0.33	0.8	0.6	0.4	0.6	0.64

An overall overview of systems' descriptions shows the importance given to answer ranking over validation. In fact, all participant systems relied on ranking methods except the LIMSI-CNRS group, which applied an approach based on validation.

We found two different approaches regarding the use of documents for finding the correct answer: on one hand, some systems work with the whole document while on the other hand, some systems select a set of promising text snippets using retrieval techniques. We do not see observation about the best performance of one approach over the other.

Some participants create hypotheses combining questions and candidates, and trying to match these hypotheses with the document excerpts.

All participants except [6] reported the use of coreference analysis in their systems, pointing out the importance of this information.

6 SUMMARY OF SYSTEMS

DIPF system [4] retrieves a set of sentences from the document that are likely to contain a correct answer according to a set of lexical and semantic similarity measures. Each candidate answer is combined with the question to form a hypothesis to be checked against the selected sentences. The final decision about the selected answer relies on a linear combination of two scores for each Text-Hypothesis pair: (1) the confidence score given by a state of the art RTE system and; (2) a combination of lexical and semantic similarity measures.

Synapse Dèveloppement [2][3] builds Clause Description Structures (CDS) for documents, questions and answers, and compares them in order to take the final decision. CDSs represent a rich structure containing semantic information of texts, as well as relations among the elements of the text. The system first removes candidate answers that do not match the expected answer type. Then it uses CDSs to compute the number of common elements and their proximity between documents and candidate answers. This value is used to rank the candidate answers and select the first one.

CICNLP system [5] combines questions with candidate answers to build hypotheses. First, the system generates graph representations for the hypotheses and documents based on syntactic analysis. Paths sharing initial and final nodes both in text and hypothesis are converted into linguistic features for vector representation. Finally, the system uses these vectors for computing the cosine similarity, and ranking the candidate answers.

CSGS system [6] is based on weighting the alignment of text sentences and question answers at token and chunk level.

LIMSI-CNRS system [7] relies on a validation approach in contrast to ranking methods used by other participants. First, the system uses the question and its expansion to retrieve passages of 3 to 5 sentences. Second, the system creates predicate-argument structures for passages and candidate answers, trying to align them at the word level using semantic relations. Then, the system applies a set of validation and invalidation rules. A candidate answer is validated if it fires all the validation rules and does not fire any invalidation rule. If there is more than one answer after the validation process, the system selects the answer with the highest alignment score. Validation and invalidation rules were made manually over information on subjects, predicates and arguments, as well as predicate truth values given by TruthTeller [8].

7 CONCLUSIONS

Last year exercise experience suggested the need to develop strategies to reject answers more than strategies to accept answers. One system started to develop this strategy but results aren't yet among the top performers. All systems except Synapse's for French select more incorrect answers than correct ones. This is really a measure of

progress in systems development. However, at the reading perspective evaluation, we have already three systems (two teams) able to pass at least half of reading tests.

Again, the Entrance Exams task shows that Question Answering is a task far from being solved. However, it provides a real benchmark able to assess real progress in the field along future years.

ACKNOWLEDGEMENTS

The collaboration has been developed in the framework of Todai Robot Project in Japan, and the CHIST-ERA Readers project in Europe (MINECO PCIN-2013-002-C02-01). 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.

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. Dominique Laurent, Baptiste Chardon, Sophie Negre and Patrick Seguela. English run of Synapse Développement at Entrance Exams 2014. *CLEF 2014 Working Notes*, Sheffield, 2014
3. Dominique Laurent, Baptiste Chardon, Sophie Negre and Patrick Seguela. French run of Synapse Développement at Entrance Exams 2014. *CLEF 2014 Working Notes*, Sheffield, 2014
4. Neil Dhruva, Oliver Ferschke and Iryna Gurevych. Solving Open-Domain Multiple Choice Questions with Textual Entailment and Text Similarity Measures. *CLEF 2014 Working Notes*, Sheffield, 2014
5. Helena Gómez-Adorno, Grigori Sidorov, David Pinto and Alexander Gelbukh. Graph Based Approach for the Question Answering Task Based on Entrance Exams. *CLEF 2014 Working Notes*, Sheffield, 2014
6. Simon Ostermann, Nikolina Koleva, Alexis Palmer and Andrea Horbach. CSGS: Adapting a short answer scoring system for multiple-choice reading comprehension exercises. *CLEF 2014 Working Notes*, Sheffield, 2014
7. Martin Gleize, Anne-Laure Ligozat and Brigitte Grau. LIMSI-CNRS@CLEF 2014: Invalidating Answers for Multiple Choice Question Answering. *CLEF 2014 Working Notes*, Sheffield, 2014
8. Lotan, A., Stern, A., Dagan, I.: Truthteller: Annotating predicate truth. In: Proceedings of NAACL-HLT 2013. pp. 752-757