Overview of CLEF QA Entrance Exams Task 2015

Álvaro Rodrigo[1], Anselmo Peñas[1], Yusuke Miyao[2],
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 2015. Following the last two editions, 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.

1 INTRODUCTION

Following the 2013 and 2014 editions, the Entrance Exams task at CLEF QA Track 2015 is focused on solving Reading Comprehension (RC) 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 UNED. 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 Entrance Examinations. Original examinations include various styles of questions, such as word filling, grammatical error recognition, sentence filling, etc.

One of such styles is reading comprehension, where a test provides a text that describes some daily life situation, and questions about the text. As in the previous edition, we reduced the challenge to these Reading Comprehension exercises contained in the English exams.

For each examination, one text is given and some (between 4 and 8) questions about the given text are asked. Each question has four choices, with only one correct answer. For this year campaign, we reused as development data 24 examinations from previous campaigns, with a total number of 115 questions and 460 candidates. Besides, we provided a new test set of 19 documents, 89 questions and 356 candidate answers to be validated.

3.2 Languages

Test data sets, originally in English, were manually translated into German, Russian, French, Spanish and Italian [1]. They are parallel translations of texts, questions and candidate answers. All these collections represent a benchmark for evaluating systems in different languages.

In addition to the official data, we collected several unofficial translations for each language. These collections have the same meaning that the original collection, but they use different words, expressions, syntax, semantics and anaphora, which produce collections with a different level of difficulty. The study of results over these variations should offer useful conclusions about systems' performance and the main issues for current technologies.

4 EVALUATION

We obtain the score of each system comparing the answers of systems against the gold standard collection with annotations made by humans. This is an automatic evaluation where we do not need manual assessments.

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:** firstly we obtain scores for each reading test. Then, we consider that a system passes a test if the score is at least 0.5. Finally, we account for the number of passed tests. A system passes the task if it passes more than a half of tests.

[1] Development data was translated to the same languages in the previous edition.

5 RESULTS

Table 1 enumerates the participating groups and their reference paper in CLEF 2015 Working Notes. Although the number of participant groups was the same than last year, they presented fewer systems (only 18 runs). Only LIMSI-CNRS has participated in the three editions, while two teams, CICNLP and SYNAPSE, took part also in the last edition and only one team (Synapse) has participated in a second language different than English (French).

Table 1. Participants and reference papers

Group ID	Group Name	#runs	Reference paper
SYNAPSE	Synapse Développement, France	2	Laurent et al. 2015 [2]
NTUNLG	National Taiwan University, Taiwan	3	-
CICNLP	Centro de Investigación en Computación Instituto Politécnico Nacional, Mexico	8	-
CoMIC	Universitä Tübingen, Germany	1	Ziai 2015 [4]
LIMSI-CNRS	ILES – LIMSI, France	4	Gleize et al. 2015 [3]

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

Table 2 shows that only the two systems from Synapse [2] gave more correct answers than incorrect ones and obtained a *c@1* score greater than 0.5. In fact, Synapse obtained also the best results in the previous edition. While French results remain similar, English results raise from a *c@1*score of 0.45 in the last edition, to a score of 0.58 in this edition. Furthermore, the LIMSI group has improved also its performance with respect to the previous edition, while CICNLP obtained similar scores.

Overall results were lower in this edition. This may mean that the current collection was more complex, but participants did not reported if they performed better over past collections. This is why we would find interesting the proposition of some baseline systems based on lexical and syntactic similarity able to offer reference scores for collections.

Table 2. Overall results for all runs, QA perspective

RUN NAME	C@1	# of questions ANSWERED				# of questions UNANSWERED
		RIGHT	WRONG	TOTAL	Prec.	
Synapse-English	0.58	52	37	89	0.58	0
Synapse-French	0.56	50	39	89	0.56	0
LIMSI-2	0.36	32	57	89	0.36	0
LIMSI-1	0.34	30	59	89	0.34	0
LIMSI-3	0.31	28	61	89	0.31	0
LIMSI-4	0.31	28	61	89	0.31	0
cicnlp-8	0.3	27	62	89	0.3	0
cicnlp-2	0.29	26	63	89	0.29	0
NTUNLG-2	0.29	26	63	89	0.29	0
CoMiC-1	0.29	26	63	89	0.29	0
cicnlp-3	0.28	25	64	89	0.28	0
cicnlp-5	0.28	25	64	89	0.28	0
cicnlp-4	0.27	24	65	89	0.27	0
cicnlp-6	0.26	23	66	89	0.26	0
cicnlp-1	0.26	23	66	89	0.26	0
Random	0.25	22	67	89	0.25	0
NTUNLG-3	0.24	21	68	89	0.24	0
NTUNLG-1	0.22	17	57	74	0.23	15
cicnlp-7	0.21	19	70	89	0.21	0

Only one system (NTUNLG-1) decided to leave some questions unanswered, while several systems of two participants did it in the previous edition. In fact, this is why NTUNLG-1 was the only system with different c@1 and precision scores. This is because c@1 gives the same score that accuracy and precision if all the questions are answered.

Therefore, there have been fewer systems returning unanswered questions every year. However, the reduction of unanswered questions did not bring a reduction of incorrect answers. It seems that neither the campaign nor the evaluation measure have been able to promote this change in systems. Hence, we must think about new ways of promoting such change.

Synapse reported additional experiments where they left unanswered some questions and they increased precision over answered questions. However, they obtained fewer correct answers and lower *c@1* scores. This is because the objective of c@1 is to acknowledge systems able to reduce incorrect answers while keeping the number of correct answers, but systems are not able to do so.

On the other hand, Table 3 shows results for the reading perspective. First column corresponds to systems run id, second column

Table 3. Overall results for all runs, reading perspective

Run	c@1	Pass	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19
Synapse-En	0.58	16/19	0.25	**0.5**	0.33	**0.5**	0.5	**1**	**0.8**	0.67	**0.6**	**0.8**	0.4	0.5	**0.57**	**0.75**	0.5	**0.5**	**0.67**	**0.75**	0.67
Synapse-Fr	0.56	16/19	0.25	**0.5**	**0.5**	0.25	0.5	0.67	0.6	**1**	**0.6**	**0.8**	0.4	**0.75**	**0.57**	0.5	0.5	**0.5**	0.5	0.5	**0.83**
LIMSI-2	0.36	8/19	0	0.25	**0.5**	0.25	0.5	0.67	0.2	0.33	0.2	0.6	0.6	0.25	0.29	0.25	0.5	**0.5**	0.5	0	0.33
LIMSI-1	0.34	5/19	0.25	**0.5**	0.33	0.25	0	0.33	0.4	0	0.4	0.2	0.6	0.25	0.43	0.25	0.5	0.25	0.5	**0.75**	0.17
LIMSI-3	0.31	6/19	**0.5**	0.25	**0.5**	0	0.17	0.33	0.4	0.67	0.4	0.6	0	0.25	0.29	0.5	0.25	**0.5**	0.17	0.25	0.17
LIMSI-4	0.31	4/19	0.25	0.25	0.17	0.25	0.17	0.33	0.6	0	0.2	0.4	0.6	**0.75**	0.43	0	**0.75**	0	0.33	0.25	0.17
Average	0.31	-	0.24	0.29	0.25	0.31	0.38	0.31	0.20	0.37	0.32	0.31	0.30	0.42	0.30	0.26	0.50	0.36	0.26	0.36	0.33
cicnlp-8	0.3	6/19	0	0.25	0.17	**0.75**	0.33	0.67	0	0.67	**0.6**	0.2	0.4	0.25	0	0	0.5	0.25	0.17	0.25	0.67
cicnlp-2	0.29	5/19	0.25	0.25	0.17	0.25	0.5	0.33	0	0	0.4	0.2	0.2	0.5	0.43	0	0.5	**0.5**	0	0.25	0.67
NTUNLG-2	0.29	6/19	**0.5**	0	0.33	0.25	0.17	0	0	0	0.2	0.2	**0.8**	0.5	0.29	0.5	0.5	0	0.33	0.5	0.33
CoMiC-1	0.29	5/19	0.25	0.5	0.33	0.25	**0.83**	0.67	0.2	0.33	0.4	0.2	0	0	0.29	0	0.5	0.25	0.17	0.5	0
Median	0.29	-	0.25	0.25	0.17	0.25	0.47	0.33	0.1	0.33	0.4	0.2	0.3	0.5	0.29	0.25	0.5	**0.5**	0.17	0.25	0.17
cicnlp-3	0.28	7/19	0	**0.5**	0.17	0.25	0.5	0	0	0.67	0	0.4	0.2	0.5	0.14	0.25	**0.75**	**0.5**	0.17	0.5	0.17
cicnlp-5	0.28	5/19	0.25	0.25	**0.5**	0.75	0.17	0	0.2	0.33	0.4	0.2	0	0	0.14	0.5	0.25	**0.5**	0.17	0.25	0.5
cicnlp-4	0.27	5/19	0.25	0.5	0.17	0.25	0.33	0	0	0.33	0.4	0	0	0.5	0.29	0.25	**0.75**	**0.5**	0	0.25	0.5
cicnlp-6	0.26	5/19	**0.5**	0.25	0	0.25	0.5	0	0	0.33	**0.6**	0.4	0	0.5	0.29	0.25	0.25	**0.5**	0	0.25	0.17
cicnlp-1	0.26	4/19	0.25	0.25	0.17	0.25	0.5	0.33	0	0.33	0.4	0	0.2	0.5	0.43	0	0.5	**0.5**	0	0.25	0.17
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	0.25	0.25	0.25	0.25	0.25	0.25	0.25
NTUNLG-3	0.24	5/19	0.25	0	0	0.25	0.5	0	0	0	0	0.2	0.6	0.5	0.14	0.25	0.5	0.25	0.33	0.5	0.17
NTUNLG-1	0.22	3/19	0.25	0	0	0.25	0.44	0	0	0	0	0	0.48	0.5	0.2	0.25	0.5	0.25	0.44	0.5	0
cicnlp-7	0.21	3/19	0	0.25	0.17	0.25	0.17	0.33	0.2	**1**	0	0.2	0	0.5	0.14	0.25	0.5	0.25	0.17	0	0.17

to the overall c@1 score obtained, third column shows the number of tests that the systems had passed if we considered a *c@1* threshold of 0.5, and the rest of columns correspond to the c@1 value for each single test.

Under the reading perspective we say that a system passes the global exam if it passes a 50% or more tests. That is, if the system passes at least 10 reading tests. According to this requirement, only the two systems from Synapse passed the 2015 Entrance Exams task.

Although results in the QA perspective are worse than in the previous edition, results in the RC perspective are a little bit greater. In fact, the proportion of passed tests this year (84%) is higher than in the previous edition (75%).

We have also observed that the raking of system sometimes changes between the QA and the Reading perspective. For instance, system *cicnlp-3* ranked fourth in the Reading perspective, while it ranked eleventh in the QA perspective. We observed also similar changes for other systems. We think this is because participants have focused more on the Reading perspective, creating systems with low results in some tests, but good results in other tests.

We think Tables 4 and 5, which show the number of systems passing each test and the maximum score per test, offer a similar conclusion. We see in Tables 4 and 5 that the maximum scores remain similar or better this year with respect to the previous edition. Moreover, there are more systems passing a test despite the fact that we have fewer systems in this edition.

Tables 4 and 5 show also a different degree of difficulty for systems over each test. This difficulty mainly depends on the lexical gap between the text and the candidate answers. Besides, systems find also difficulties depending on different formulations of the same text, as Synapse showed last year [5] and we are studying now with different versions of the same collection.

Table 4. Number of runs (out of 18) that passed each test (from test 1 to test 10), and maximum c@1 score achieved per test.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
# Runs pass	3	12	2	5	15	10	4	8	6	6
Max. score	0.50	0.75	0.57	0.75	0.75	0.50	0.67	0.75	0.83	0.50

Table 5. Number of runs (out of 18) that passed each test (from test 11 to test 19), and maximum c@1 score achieved per test.

	T11	T12	T13	T14	T15	T16	T17	T18	T19
# Runs pass	4	3	9	5	3	6	4	4	5
Max. score	0.50	0.75	0.83	1.00	0.80	1.00	0.60	0.80	0.80

6 SUMMARY OF SYSTEMS

In this Section we offer deeper details about systems from groups that sent a reference paper about their participation.

The general architecture of participant systems includes the following components: (1) a preprocessing step for preparing texts for the next steps; (2) the retrieval of relevant passages in order to reduce the search space; (3) the creation of graph style structures from texts, questions and candidates; (4) the enrichment of structures including, for instance, background knowledge; (5) the comparison of structures as a way of finding the most probable answer; (6) the ranking of candidates with respect to the comparison score and; (7) the selection of the candidate with the best score. Most of systems follow this architecture, with slight changes at some levels or removing some steps.

The general architecture shows that systems relied on ranking methods instead of validation. In fact, only one system (*NTUNLG-1*) decided to leave unanswered some questions. Systems know that there is a correct answer, and they take the risk of returning always the candidate more similar to the text, no matter if this similarity is low. As we have pointed out above, this is not the expected behavior in the task, and we must think about new ways of promoting the desired change in such direction.

It is still not clear the impact of selecting relevant passages rather than working with the whole document. Systems working with passages do it as a way of reducing the amount of work in the following steps.

Unfortunately, participant groups did not report if they return some incorrect answers as a consequence of this selection. We think this might be an interesting study for other researchers. Anyway, it is clear that this step must be focused on recall rather than in precision.

Some participants prefer to work with a representation of texts and answers similar to graphs instead of raw text. We think these participants try to exploit the properties of such structures for representing connections between concepts, the inclusion of extra knowledge, etc.

Regarding the use of external knowledge, we think it is one of the main issues at this task. Reading comprehension texts contain a lot of implicit information that automatic systems might not be able to extract, as LIMSI reported [3]. However, the best performing systems, from Synapse, did not use any kind of external knowledge. We think current systems are still quite far from a proper way of representing, exploiting and using external knowledge in this task.

A more detailed analysis of each system showed that Synapse [2] built Clause Description Structures (CDS) structures, which are similar to graphs, of whole documents. They preferred not to include external knowledge from resources such as DBPedia or Wikipedia because they thought that the given text contained enough information for finding the correct candidate. They also removed candidates which did not match the expected answer type as a way of reducing the search space. Then, they compared CDSs from texts and candidates, measuring proximity and the number of common elements. Finally, they chose the candidate with the highest coefficient.

On the other hand, LIMSI [3] selected a set of passages in order to reduce the computation time. Next, they represented passages as graphs and enriched those graphs using external sources. They wanted to reduce the gap between the knowledge extracted from texts by humans and computers. Then, they recorded the changes required for passing from passages graphs to candidate graphs. Finally, they applied two classifiers, one for validation and the other one for rejection, using the set of changes as features. The selected candidate was the one with the highest final score according to equation *finalScore = validationScore – rejectionScore*.

CoMIC [4] retrieved also a set of relevant passages. For this purpose, they took also in consideration that passages relevant to the first questions usually appears at the beginning of the text, while passages referring to the last questions appears at the bottom of the text. Then,

they measured the similarity between relevant passages and candidates without using any intermediate graph structure. They accounted for vector-space model based measures, similarity measures using WordNet, as well as syntactic and semantic similarity measures. Finally, they applied a Ranking SVM model to obtain the final answer.

7 CONCLUSIONS

In the third edition of the task we expected a jump in performance in comparison with previous campaigns. However, we have seen similar results at the Question Answering perspective and slight improvements at the Reading Comprehension perspective. Only systems from Synapse could give more correct than incorrect answers.

We think the current task is still very hard for current technologies and it is not clear what the bottleneck is. We know that there are several issues such as (1) the semantic gap between texts, questions and answers; (2) external knowledge management; etc.

Participants tried different approaches and offered some details about the right way for obtaining progress in this task, but it is not clear what the right direction is.

Anyway, the availability of the created resources and methodology provides a benchmark able to assess real progress in the field along future years.

8 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) and the Voxpopuli project (TIN2013-47090-C3-1-P). 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, Camille Pradel and Patrick Seguela. Reading Comprehension at Entrance Exams 2015. *CLEF 2015 Working Notes,* Toulouse, 2015.
3. Martin Gleize and Brigitte Grau. LIMSI-CNRS@CLEF 2015: Tree Edit Beam Search for Multiple Choice Question Answering. *CLEF 2015 Working Notes*, Toulouse, 2015.
4. Ramon Ziai. CoMiC: Exploring Text Segmentation and Similarity in the English Entrance Exams Task. *CLEF 2015 Working Notes*, Toulouse, 2015.
5. Dominique Laurent, Baptiste Chardon, Sophie Negre and Patrick Seguela. English run of Synapse Développement at Entrance Exams 2014. CLEF 2014 Working Notes, Sheffield, 2014