Overview of the NTCIR-13 QA Lab-3 Task

Hideyuki Shibuki
Yokohama National University
shib@forest.eis.ynu.ac.jp

Kotaro Sakamoto
Yokohama National University
National Institute of Informatics
sakamoto@forest.eis.ynu.ac.jp

Madoka Ishioroshi
National Institute of Informatics
ishioroshi@nii.ac.jp

Yoshinobu Kano
Shizuoka University
kano@inf.shizuoka.ac.jp

Teruko Mitamura
Language Technology Institute, Carnegie Mellon University
teruko+@cs.cmu.edu

Tatsunori Mori
Yokohama National University
mori@forest.eis.ynu.ac.jp

Noriko Kando
National Institute of Informatics
The Graduate University for Advanced Studies (SOKENDAI)
kando@nii.ac.jp

ABSTRACT

The NTCIR-13 QA Lab-3 task aims at the real-world complex Question Answering (QA) technologies using Japanese university entrance exams and their English translation on the subject of "World history". QA Lab-3 has three end-to-end tasks for multiple-choice, term and essay questions. The essay task has three subtasks of extraction, summarization and evaluation-method. There were 85 submissions from 13 teams in total. We describe the used data, formal run results, and comparison between human marks and automatic evaluation scores for essay questions.

Categories and Subject Descriptions

H.3.4 [INFORMATION STORAGE AND RETRIEVAL]: Systems and Software - Performance evaluation (efficiency and effectiveness), Question-answering (fact retrieval) systems.

General Teams

Experimentation

Keywords

NTCIR-13, question answering, university entrance examination, world history, essay question

1. INTRODUCTION

The goal of the third QA Lab (Question Answering Lab for Entrance Exam) task at NTCIR 13 is to investigate the real-world complex Question Answering (QA) technologies as a joint effort of participants and appropriate evaluation metrics and methodologies for them. The questions were selected from two different stages - The National Center Test for University Admissions (multiple-choice questions) and secondary exams of the University of Tokyo (term and essay questions). Both Japanese and English translations of the topics (questions) were provided in the XML format that is defined in QA Lab[1].

As knowledge resources, 4 sets of high school textbook, Wikipedia and World History Ontology[3] were provided. Participants could use any other resources (need to report). Two open-source baseline QA systems and one passage retrieval systems were also provided. Tests were done in two phases (Phase-1 and -2). In each phase, three end-to-end tasks were done for multiple-choice, term and essay questions. For the essay task, besides the end-to-end task, three subtasks were done of extraction, summarization and evaluation-method.

Based on the lessons learned from NTCIR-11 and -12, the major challenges include

1) essay questions that require logical summaries along a historical theme,
2) competition with more than 3,500 students, examinees, from all over Japan (JA only),
3) questions with context,
4) answer by text as high-compress-ratio query-biased summarization,
5) advanced entity-focused passage retrieval,
6) enhance knowledge resources,
7) semantic representation and sophisticated learning,
8) appropriate evaluation measure for essay,
9) research run using the past QA Lab data/systems.

Research run investigates how much the QA technologies improved from QA Lab-1.

- Using the same training/test sets as the past QA Lab runs, comparison with the past results,

Table 1: tasks in each phase

Question	Task	Phase-1	Phase-2	Research run
Multiple-choice	End-to-end	YES	YES	YES
Term	End-to-End	YES	YES	N/A
Essay	End-to-End	YES	YES	YES
	Extraction	YES	YES	N/A
	Summarization	YES	YES	N/A
	Evaluation-method	YES	YES	N/A

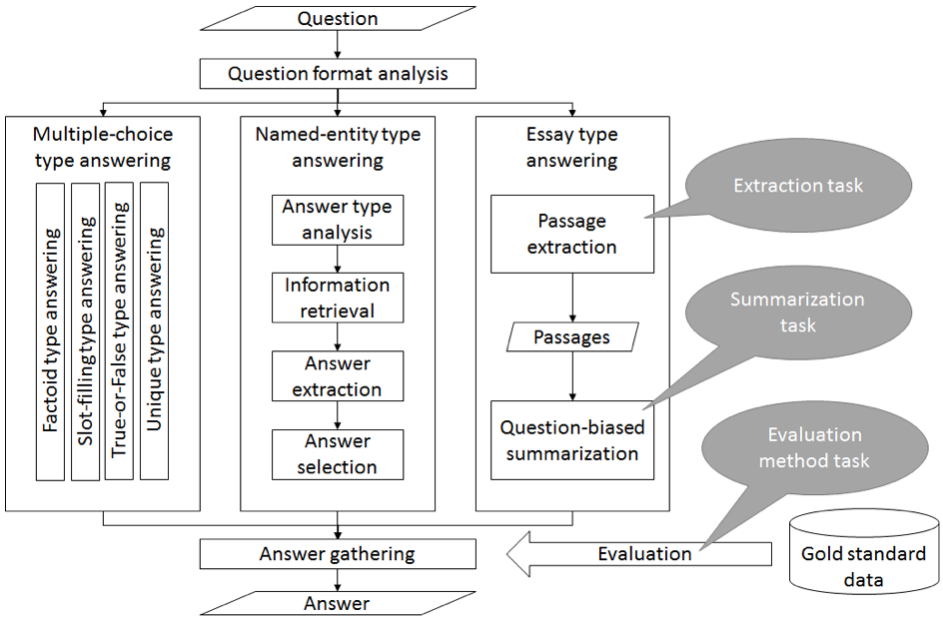

Figure 1: QA system architecture

- Using the systems participating in the past QA Lab runs, comparison with the present systems.

To tackle to them, we propose to

i) enhance question format types ontology as joint effort,

ii) define enhanced answer type,

iii) evaluate end-to-end runs as well as vertical investigation runs according to question format type âĂŞ answer-type âĂŞ knowledge needed rather than the horizontal integration planned in NTCIR-11,

iv) collect and share more knowledge resources (e.g. dictionaries, chronological tables of historical events, gazetteers, biographical dictionaries), and baseline annotated corpus. JapanâĂŹs university entrance examination is selected here, but theoretically the framework can be applicable other domains. Participation for limited-types of question or limited types of modules are possible.

2. TASK DESCRIPTION

We design the tasks as shown in Table 1. Figure 1 shows a QA system architecture mapping the tasks. For essay questions, the extraction task is the first half of the end-to-end task, and is aiming to retrieve and extract texts that should be included in essay. The summarization task is the second half, and is aiming to generate an essay by summarizing the extracted texts. The evaluation-method task is aiming to automatically evaluate essays systems generated using gold standard essays.

2.1 Topics

Table 2 shows training set and test set in each phase. Each phase has a separate training set and test set with similar difficulty. Multiple-choice questions were selected from the National Center Test in 1997, 1999, 2001, 2003, 2005, 2007, 2009, 2011, 2012, 2013 and 2014. Term and Essay questions were selected from secondary exams of the University of Tokyo in 2000 to 2014. Participants are free to participate any particular phase and either of exams.

2.2 Evaluation

For multiple-choice questions, the evaluation was done using the scores provided by National Center for University Admissions, and the accuracy. For term questions, the evaluation was done using the accuracy by exact matching with the gold standard data, which are taking account of synonym. For essay questions, the end-to-end task was evaluated by human expert marks, ROUGE method, Pyramid method and quality questions. The quality questions asked grammaticality, non-redundancy, reference clarity, fluency and 'coherence and content structure,' which were scored by four-grade human evaluation. The extraction task was

Table 2: Training set and Test set in each phase

Task	Formal run			Research run	
	Training	Phase-1	Phase-2	Training	Test
Multiple-choice	1997,1999,2001 2003,2005,2007 2009,2011	2012,2013	2014	1997,1999,2001 2003,2005,2007 2009,2011	2007,2011,2013
Term & Essay	2003,2005,2007 2009,2011	2000,2004,2008 2012,2013	2001,2002,2006 2010,2014	2000 to 2014	2002,2007,2013

evaluated by precision and recall of extracted texts including statements in Gold standard essay. The summarization task was evaluated in the same manner as the end-to-end task. The evaluation-method task was evaluated by rank correlation coefficient with human expert ranking.

2.3 Schedule

The NTCIR-13 QA Lab-3 task has been run according to the following timeline:

July 1, 2015: Training data release

Formal run Phase-1
Feb. 2, 2017: Formal run Topics release
Feb. 2 - 6, 2017: Term and Multiple-choice tasks
Feb. 9 - 13, 2017: Essay End-to-End and Essay Extraction tasks
Feb. 16 - 20, 2017: Essay Summarization task
Feb. 23 - Mar. 1, 2017: Essay Evaluation-method task

Formal run Phase-2
May 11, 2017: Formal run Topics release
May 11 - 15, 2017: Term and Multiple-choice tasks
May 18 - 22, 2017: Essay End-to-End and Essay Extraction tasks
May 25 - 29, 2017: Essay Summarization task
June 1 - 5, 2017: Essay Evaluation-method task

Research run
July 6, 2017: Research run Topics release
July 6 - 10, 2017: Essay End-to-End and Multiple-choice tasks

NTCIR-13 CONFERENCE
Sep. 1, 2017: Draft paper submission to the Task organizers
Nov. 1, 2017: Paper Submission for the Proceedings, which will be available online at the Conference.
Dec. 5 - 8, 2017: NTCIR-13 Conference

3. COLLECTION AND TOOLS

3.1 Collection

Participants are free to use any resources available with the exception of the answer sets (readily available online in Japanese). In addition, the following resources are provided, but are not required to be used.

A) Eight sets of National Center Tests

B) Five sets of Second-stage Examinations

C) Knowledge Sources (a snapshot of Wikipedia subset related to world history)

Table 3: Active participating teams

Team ID	Organization
KUAS	National Kaohsiung University of Applied Sciences
Forst	Yokohama National University
IMTKU	Tamkang University
SML	Nagoya University
KSU	Kyoto Sangyo University
SLQAL	Waseda University
CMUQA	Carnegie Mellon University
DGLab	DG Lab
tmkff	The National Center for University Entrance Examinations & Kyushu University
MTMT	Carnegie Mellon University
HagiL	Keio University

D) Right Answers

3.1.1 Sets of National Center Tests

Sets of National Center Tests, available in Japanese and English.

3.1.2 Sets of Second-stage Examinations

Sets of Second-stage Examinations of the University of Tokyo, available in Japanese and English.

3.1.3 Knowledge Sources

- Japanese high school textbooks on world history, available in Japanese.
- A snapshot of Wikipedia, available in Japanese and in English. (Participants can also use the current up-to-date version).
  - Solr Instance with Indexed Wikipedia Subset (available in English)[1]
  - NTCIR-11 QA Lab Japanese subtask: Wikipedia Data Set[2]
- World history ontology, available in Japanese.[3]

[1]https://github.com/oaqa/ntcir-qalab-cmu-baseline/wiki/Solr-Instance-with-Indexed-Wikipedia-Subset
[2]http://warehouse.ntcir.nii.ac.jp/openaccess/qalab/11QALab-ja-wikipediadata.html
[3]http://researchmap.jp/zoeai/event-ontology-EVT/

Table 4: The run number each team submitted for Phase 1, Phase 2 and Research run

Team ID	JA						EN					
	Choice	Term	Essay				Choice	Term	Essay			
			E2E	Ext	Sum	EvM			E2E	Ext	Sum	EvM
KUAS	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	1,2,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-
Forst	-,-,-	2,1,-	2,3,2	2,-,-	1,1,-	2,2,-	-,-,-	-,-,-	1,1,-	-,-,-	-,-,-	-,-,-
IMTKU	-,-,-	-,-,-	-,2,-	-,-,-	-,1,-	-,-,-	-,3,-	-,-,-	-,2,-	-,-,-	-,1,-	-,-,-
SML	-,-,-	-,1,-	1,3,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-
KSU	3,2,2	2,3,-	2,3,-	2,3,-	1,1,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-
SLQAL	1,1,1	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-
CMUQA	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,3,-	-,2,-	-,1,-	-,-,-
DGLab	-,-,-	-,-,-	-,-,1	-,-,-	-,2,-	-,2,-	-,-,-	-,-,-	-,-,1	-,-,-	-,2,-	-,2,-
tmkff	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,1,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-
MTMT	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,2,-	-,2,-	-,-,-	-,-,-
HagiL	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,1,-	-,-,-	-,-,-	-,-,-	-,-,-

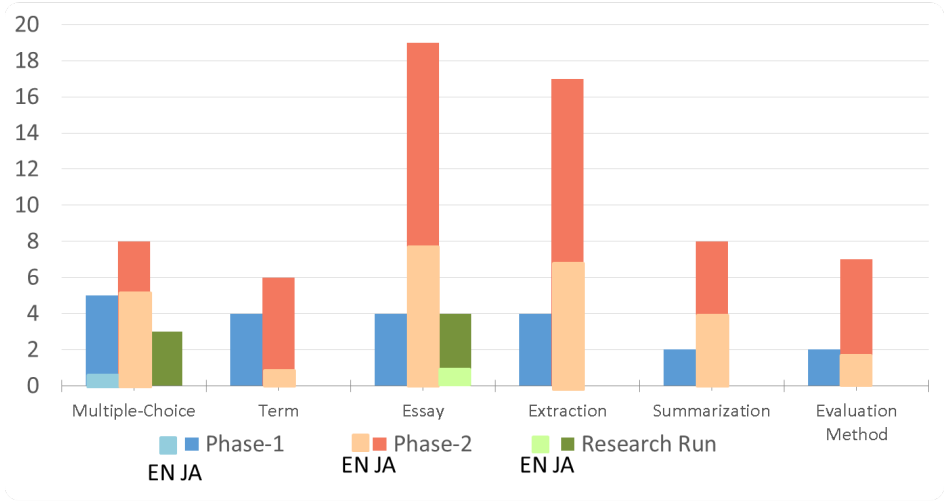

Figure 2: Total number of submissions

3.1.4 Right Answers

- Right answers for National Center Tests, available in English and Japanese.
- Right answers for Second-stage Examinations, available in English and Japanese.
- Reference essays and nuggets for Essays, available in Japanese.

3.2 Tools

- 1 baseline QA system for English, based on UIMA (CMU)[4]
- 1 baseline QA system for Japanese, based on YNU's MinerVA, CMU's Javelin and a question analysis module by Madoka Ishioroshi[5], re-constructed and implemented as UIMA components by Yoshinobu Kano[6][5]
- Scorer and Format Checker for National Center Test[6]
- Passage Retrieval Engine passache[7]

[4]https://github.com/oaqa/ntcir-qalab-cmu-baseline
[5]https://bitbucket.org/ntcirqalab/factoidqa-centerexam/
[6]https://bitbucket.org/ntcirqalab/qalabsimplescorer
[7]https://code.google.com/p/passache/

4. PARTICIPATION

18 teams were registered, and 11 teams as shown in Table 3 were participated in the end.

5. SUBMISSIONS

Table 4 and Figure 2 show the total number of submissions. Three numbers separated by comma in Table 4 show submitted numbers at Phase 1, Phase 2 and Research run respectively.

5.1 Phase 1

For Phase 1 Formal run, 24 runs from 6 teams were submitted in total as shown at the first numbers in Table 4. For Multiple-choice question task, 5 runs from 3 teams were submitted. For Term question task, 4 runs from 2 teams were submitted. For Essay question task, 6 end-to-end runs from 3 teams, 4 extraction runs from 2 teams, 2 summarization runs from 2 teams and 3 evaluation-method runs from 2 teams were submitted.

5.2 Phase 2

For Phase 2 Formal run, 56 runs from 11 teams were submitted in total as shown at the second numbers in Table 4. For Multiple-choice question task, 8 end-to-end runs from 4

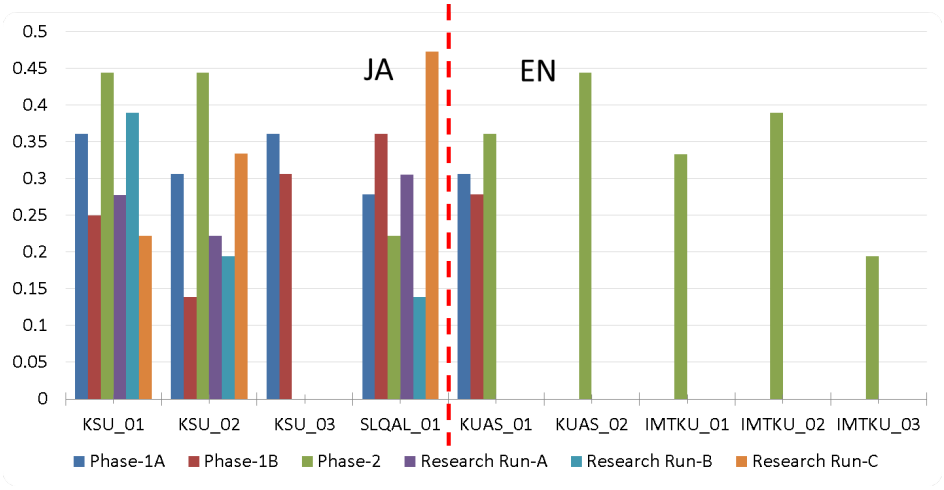

Figure 3: Correct rates in Multiple-choice question task

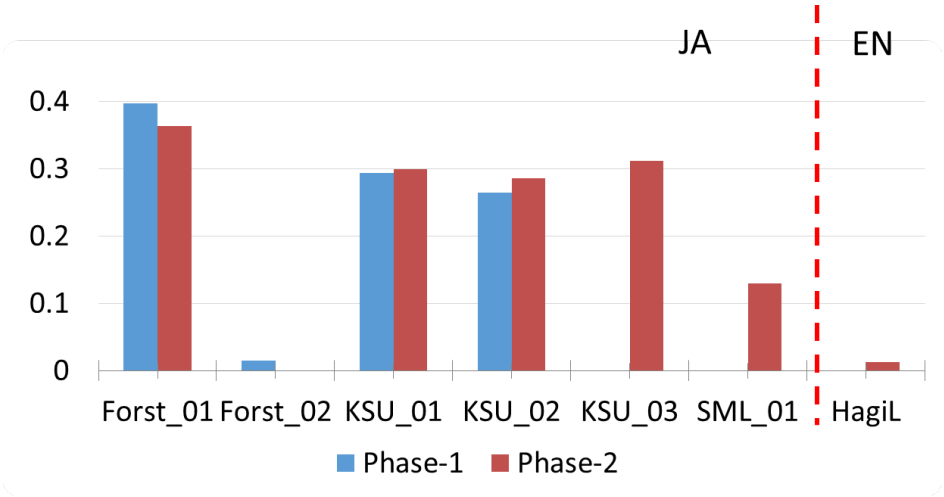

Figure 4: Correct rates in Term question task

teams were submitted. For Term question task, 6 end-to-end runs from 4 teams were submitted. For Essay question task, 19 end-to-end runs from 6 teams, 7 extraction runs from 3 teams, 9 summarization runs from 5 teams and 7 evaluation-method runs from 3 teams were submitted.

5.3 Research run

For Research run, 6 runs from 4 teams were submitted in total as shown at the third numbers in Table 4. For multiple choice questions, 3 runs from 2 teams were submitted. For Essay questions, 3 end-to-end runs from 2 teams were submitted. Note that Research run had only Multiple-choice question task and Essay question end-to-end task.

6. RESULTS

6.1 Multiple Choice Question Task

Table 8, 9 and 10 show results of the multiple-choice question task at Phase-1, -2 and Research run respectively. Figure 3 shows the correct rates in all phases. According to Figure 3, KSU achieved the best correct rate at Phase-1, KSU and KUAS were the best at Phase-2 and SLQAL was the best at Research run. The difference among the results was a little. Although the results got better than their own results at the QA Lab-2, no results could be better than the best result at the QA Lab-2.

6.2 Term Question Task

Table 11 and 12 show results of the term question task at Phase-1 and -2 respectively. Figure 4 shows correct rates in all phases. According to Figure 4, Forst achieved the best correct rates at Phase 1 and 2, and KSU was the second best.

6.3 Essay Question Task

6.3.1 End-to-end Task

Table 13, 14 and 15 show results of the essay question task at Phase-1, -2 and Research run respectively. Figure 5 and 6 show human marks, ROUGE scores and Pyramid scores at Phase-1 and -2 respectively. Figure 7 and 8 show quality question scores at Phase-1 and -2 respectively. At Phase 1 in Japanese task, Forst was the best Pyramid score and KSU

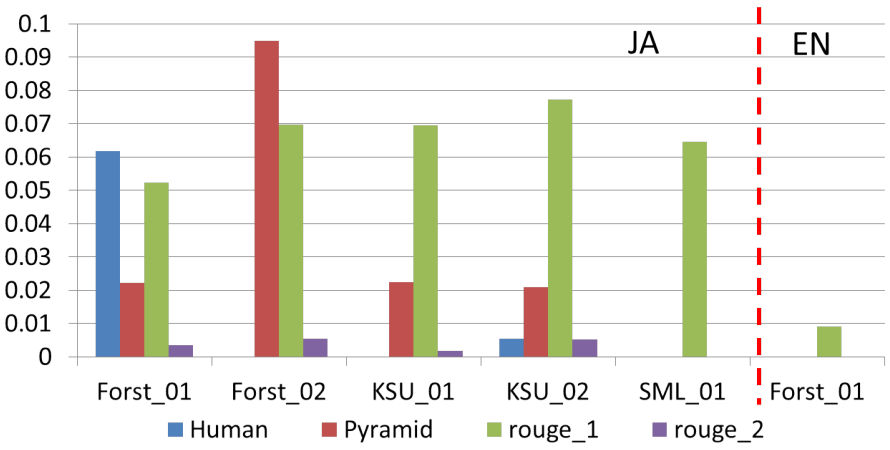

Figure 5: Human marks, ROUGE and Pyramid scores in Essay task at Phase 1

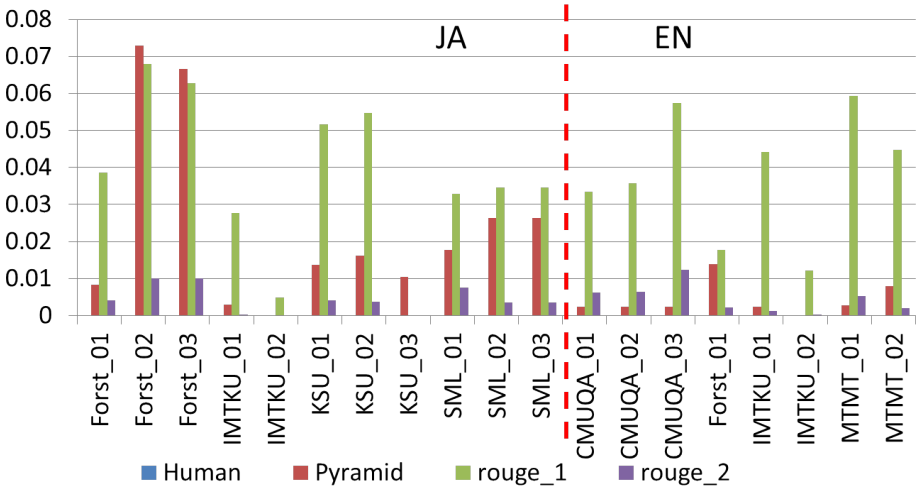

Figure 6: Human marks, ROUGE and Pyramid scores in Essay task at Phase 2

Table 5: Results of Extraction task at Phase 1 (N = 3)

TeamID	Priority	Lang	Passage Precision	Nugget Recall	Ave. of tokens
Forst	1	JA	0.267	0.019	1037.6
KSU	1	JA	0.468	0.288	1147.5
KSU	2	JA	0.251	0.100	1483.5

was the best ROUGE-1 score. At Phase 2, Forst achieved the best Pyramid and ROUGE-1 scores in Japanese task, while Forst was the best Pyramid score and MTMT was the best ROUGE-1 score in English task. In Research run, DGLab achieved the best Pyramid and ROUGE-1 scores in Japanese and English tasks. According to Figure 7 and 8, the qualities of reference clarity and 'coherence and content structure,' are low by and large. The improvement of the qualities may enhance the total improvement.

6.3.2 *Extraction Task*

Table 5 and 6 show the passage precision and the nugget recall in the extraction task at Phase-1 and -2 respectively. The passage precision is the rate of passages including at least one gold standard nugget in extracted passages of which token number is within the limit length multiplied by N. The nugget recall is the rate of nuggets included among the extracted passages in all gold standard nuggets. Table 5 and 6 show the results in the case that N is 3. The results in the case that N is 5 or 10 are shown in Table 16 to 19. At Phase 1, KSU achieved the best passage precision and the best nugget recall in Japanese task. At Phase 2, DGLab achieved the best passage precision, and KSU achieved the best nugget recall in Japanese task. IMTKU achieved the best passage precision and the best nugget recall in English task.

6.3.3 *Summarization Task*

The below of Table 13 and 14 show the results of the summarization task at Phase-1 and -2 respectively. At Phase 1, KSU achieved the best Pyramid and ROUGE-1 scores. At Phase 2 in Japanese task, DGLab was the best Pyramid score, and KSU was the best ROUGE-1 score. At Phase 2 in English task, DGLab was the best Pyramid score, and

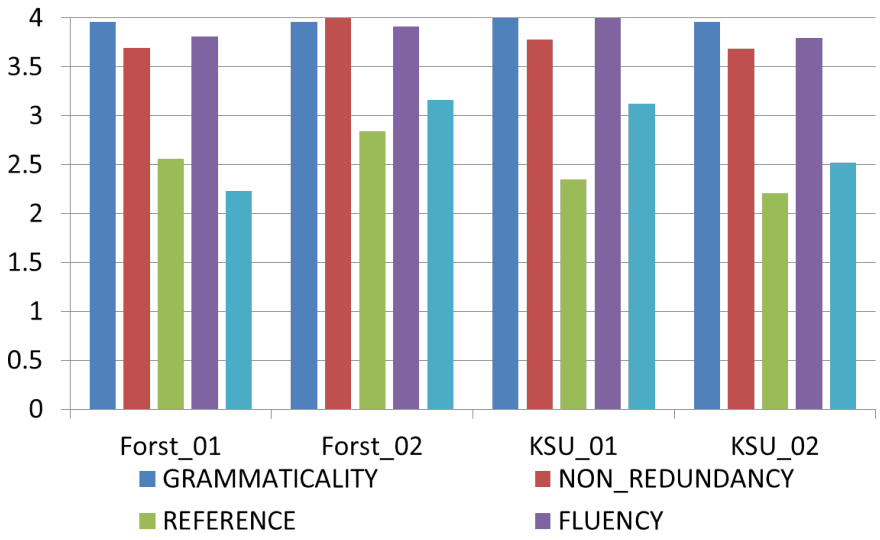

Figure 7: Quality question scores in Essay task at Phase 1

Table 6: Results of Extraction task at Phase 2 (N = 3)

TeamID	Priority	Lang	Passage Precision	Nugget Recall	Ave. of tokens
DGLab	1	JA	0.510	0.057	1875.6
DGLab	2	JA	0.479	0.044	1875.6
DGLab	3	JA	0.263	0.166	1459.2
Forst	1	JA	0.038	0.080	1578.0
Forst	2	JA	0.192	0.017	1324.4
IMTKU	1	JA	0.113	0.020	454.4
IMTKU	2	JA	0.000	0.000	336.25
KSU	1	JA	0.057	0.152	1591.8
KSU	2	JA	0.100	0.201	1592.6
KSU	3	JA	0.083	0.057	1597.6
CMUQA	1	EN	0.113	0.035	243.2
CMUQA	2	EN	0.088	0.026	274.2
DGLab	1	EN	0.087	0.035	770.4
DGLab	2	EN	0.117	0.035	770.4
IMTKU	1	EN	0.260	0.061	249.2
IMTKU	2	EN	0.234	0.058	249.2
MTMT	1	EN	0.009	0.032	797.2
MTMT	2	EN	0.014	0.019	782.4

Table 7: Results of Evaluation method task

TeamID	Priority	Lang	Spearman's Rho	Kendall's Tau-b
Phase 1				
Forst	1	JA	0.427	0.334
Forst	2	JA	0.596	0.534
Pyramid		JA	0.728	0.638
ROUGE-1		JA	0.677	0.568
ROUGE-2		JA	0.599	0.472
Phase 2				
Forst	1	JA	-0.071	-0.049
Forst	2	JA	0.404	0.360
tmkff	1	JA	0.193	0.212
DGLab	1	JA	0.200	0.167
DGLab	2	JA	0.341	0.303
DGLab	1	EN	0.333	0.286
DGLab	2	EN	-0.160	-0.067
Pyramid		JA	0.428	0.381
ROUGE-1		JA	0.620	0.588
ROUGE-2		JA	0.120	0.062
Pyramid		EN	0.086	0.073
ROUGE-1		EN	-0.263	-0.206
ROUGE-2		EN	-0.343	-0.273

CMUQA was the best ROUGE-1 score in the condition of using gold standard nuggets.

6.3.4 *Evaluation Method Task*

Table 7 shows the rank correlation coefficients with human marks in the evaluation method task[8]. For reference, the rank correlation coefficients to Pyramid scores, ROUGE -1 and -2 scores are shown in Table 7. According to Table 7, Forst achieved the best result at Phase 1 and 2, and DGLab was the second best.

7. OUTLINE OF THE SYSTEMS

[8]Because DGLab graded by deducting marks, we calculated their correlation coefficients by inversing their sign.

We briefly describe the characteristic aspects of the participating groups' systems and their contribution below.

The KUAS team tackled the multiple-choice question tasks in English. The system converted the content and Wikipedia page of the item into concept maps, and compared the similarity between the concept maps of the item and source of knowledge to determine the answer.

The Forst team tackled the term question and the essay question tasks in mainly Japanese. The system extracted named entities from question as implicit keywords, and generated essays including sentences retrieved by the implicit keywords.

The IMTKU team tackled the multiple-choice question and the essay question tasks in Japanese and English. They

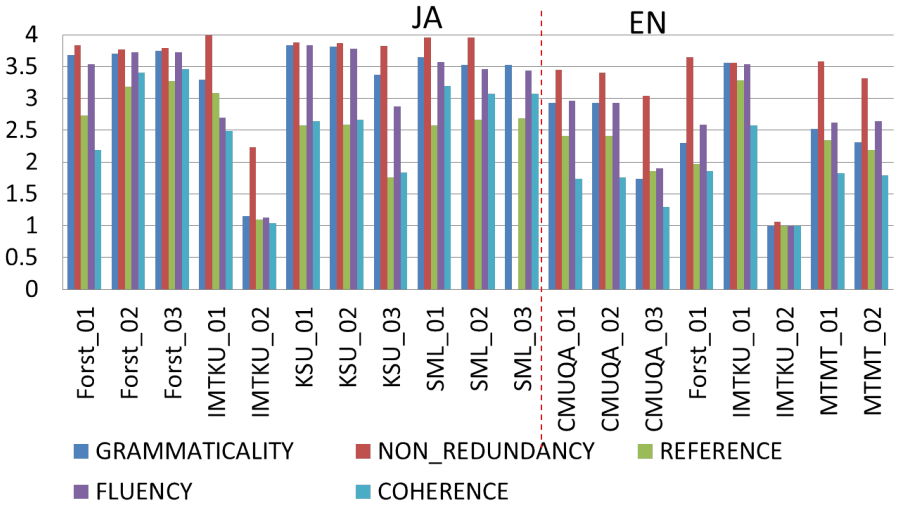

Figure 8: Quality question scores in Essay task at Phase 2

integrated various natural language processing tools and resources for each language.

The SML team tackled the term question and the essay question tasks in Japanese. They focused on simple essay questions of which length is smaller than 100 characters. The system identified a question focus using nouns in specific phrases, and compressed sentences using grammatical rules and query relevance score.

The KSU team tackled all tasks except the evaluation-method task in Japanese. For multiple-choice questions, they introduced query generation according to answer types. For term questions, they inferred answer types inference taking account of word order, and scored answer candidates based on dependency graph. For essay questions, they introduced query generation according to instruction types and simple-sentence retrieval.

The SLQAL team tackled the multiple-choice question task in Japanese. The system extracted nouns from question and choices as keywords, and estimated scores based on retrieved textbook data using the keywords.

The CMUQA team tackled the essay question tasks in English at Phase 2. The system consists of question analysis, document retrieval, sentence extraction, sentence scoring, sentence ordering and short essay generation. Wikipedia is used as knowledge source, and AMR is used as semantic representation.

The DGLab team tackled the essay question tasks in Japanese and English since Phase 2. The end-to-end system consists of condition extraction, passage retrieval, sentence selection and extractive summarization. For evaluation-method task, they used Word Mover's Distance between gold standard nuggets and essay system generated.

The tmkff team tackled the essay evaluation-method task in Japanese at Phase 2. The system evaluated essays by agreement with prepared key phrases and prediction score offered by Random Forests.

The MTMT team tackled the essay question tasks in English at Phase 2. They pointed out that the difference of available data between Japanese and English tasks, and expanded their knowledge source using English translation of Japanese data by utilizing linked open data.

The HagiL team tackled the term question task in English at Phase 2. The system extracted significant sentences by the similarities of the embeddings of the question and each sentence in retrieved documents, and extracted the answer span in the significant sentences.

8. CONCLUSIONS

We described the overview of the NTCIR-13 QA Lab-3 task. The goal is the real-world complex Question Answering (QA) technologies using Japanese university entrance exams and their English translation on the subject of "World history". We conducted 2 phases of formal runs and a research run. 11 teams submitted 86 runs in total. We described the task description, the collection, the provided tools, the participation and the results.

Acknowledgment

Our thanks to participants, National Center for University Entrance Examinations, JC Educational Institute, Inc. and the answer creaters. Part of the task organization was supported by NII's Todai Robot Project[4]

9. REFERENCES

[1] Hideyuki Shibuki, Kotaro Sakamoto, Yoshinobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y. Itakura, Di Wang, Tatsunori Mori, Noriko Kando. Overview of the NTCIR-11 QA-Lab Task. Proceedings of the 11th NTCIR Conference, 2014.

[2] Hideyuki Shibuki, Kotaro Sakamoto, Madoka Ishioroshi, Akira Fujita, Yoshinobu Kano, Teruko Mitamura, Tatsunori Mori, Noriko Kando. Overview of the NTCIR-12 QA Lab-2 Task. Proceedings of the 12th NTCIR Conference, 2016.

[3] Ai Kawazoe, Yusuke Miyao, Takuya Matsuzaki, Hikaru Yokono, Noriko Arai. World History Ontology for Reasoning Truth/Falsehood of Sentences: Event Classification to Fill in the Gaps between Knowledge Resources and Natural Language Texts. In Nakano,

Yukiko, Satoh, Ken, Bekki, Daisuke (Eds.), New Frontiers in Artificial Intelligence (JSAI-isAI 2013 Workshops), Lecture Notes in Computer Science 8417, pp.42–50, 2014.

[4] http://21robot.org/

[5] Madoka Ishioroshi, Yoshinobu Kano, Noriko Kando. A study of multiple choice problem solver using question answering system. IPSJ NL-215 research report. 2014. (in Japanese)

[6] Yoshinobu Kano, 2014. Materials delivered at the Hands-on Tutorial for UIMA and the QA Lab baseline systems

[7] Tatsunori Mori: Japanese question-answering system using A* search and its improvement. ACM Trans. Asian Lang. Inf. Process. 4(3): 280–304 (2005)

[8] Shima, H., Lao, N., Nyberg, E., Mitamura, T. (2008). Complex Cross-lingual Question Answering as Sequential Classification and Multi-Document Summarization Task. In NTCIR-7 Workshop.

[9] https://github.com/oaqa/ntcir-qalab-cmu-baseline

[10] https://code.google.com/p/passache/

[11] Yoshinobu Kano. Kachako: a Hybrid-Cloud Unstructured Information Platform for Full Automation of Service Composition, Scalable Deployment and Evaluation. In the 1st International Workshop on Analytics Services on the Cloud (ASC), the 10th International Conference on Services Oriented Computing (ICSOC 2012). Shanghai, China, November 12nd 2012.

[12] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the ACL-04 workshop 8, 2004.

[13] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In Proceedings of HLT/NAACL 2004, 2004.

APPENDIX

We describe the detail results in Table 8 to 29.

Table 8: Detail results of Multiple-Choice questions in Phase-1

End-to-End Run									
TeamID	Language	Priority	# of ques	# of correct	# of incorrect	# of N/A	Correct rate	Total score	Average score
KSU	JA	1	72	22	49	1	0.306	60	0.300
KSU	JA	2	72	16	55	1	0.222	45	0.225
KSU	JA	3	72	24	47	1	0.333	66	0.330
KUAS	EN	1	72	21	51	0	0.292	55	0.275
SLQAL	JA	1	72	23	49	0	0.319	65	0.325

Table 9: Detail results of Multiple-Choice questions in Phase-2

End-to-End Run									
TeamID	Language	Priority	# of ques	# of correct	# of incorrect	# of N/A	Correct rate	Total score	Average score
IMTKU	EN	1	36	12	24	0	0.333	34	0.340
IMTKU	EN	2	36	14	22	0	0.389	40	0.400
IMTKU	EN	3	36	7	29	0	0.194	18	0.180
KSU	JA	1	36	16	20	0	0.444	45	0.450
KSU	JA	2	36	16	20	0	0.444	44	0.440
KUAS	EN	1	36	13	19	4	0.361	37	0.370
KUAS	EN	2	36	16	16	4	0.444	44	0.440
SLQAL	JA	1	36	8	28	0	0.222	23	0.230

Table 10: Detail results of Multiple-Choice questions in Research run

End-to-End Run									
TeamD	Language	Priority	# of ques	# of correct	# of incorrect	# of N/A	Correct rate	Total score	Average score
KSU	JA	1	108	32	76	0	0.296	87	0.290
KSU	JA	2	108	27	81	0	0.250	70	0.233
SLQAL	JA	1	108	33	75	0	0.306	89	0.297

Table 11: Detail results of Term questions in Phase-1

End-to-End Run							
TeamID	Language	Priority	# of ques	# of correct	# of incorrect	# of N/A	Correct rate
Forst	JA	1	68	27	41	0	0.397
Forst	JA	2	68	1	1	66	0.015
KSU	JA	1	68	20	48	0	0.294
KSU	JA	2	68	18	50	0	0.265

Table 12: Detail results of Term questions in Phase-2

End-to-End Run							
TeamID	Language	Priority	# of ques	# of correct	# of incorrect	# of N/A	Correct rate
Forst	JA	1	77	21	45	2	0.273
KSU	JA	1	77	20	48	0	0.260
KSU	JA	2	77	17	50	0	0.221
KSU	JA	3	77	20	48	0	0.260
SML	JA	1	77	8	58	2	0.104
HagiL	EN	1	77	1	67	0	0.013

Table 13: Detail results of Essay questions in Phase-1

End-to-end Run													
TeamID	Priority	Lang.	#of ques	#of N/A	content score				quality score				
					Human	NUGGET	rouge_1	rouge_2	GRAMMATICALITY	NON_REDUNDANCY	REFERENCE	FLUENCY	COHERENCE
Forst	1	JA	26	1	0.011	0.0221	0.0523	0.00351	3.96	3.69	2.56	3.81	2.23
Forst	2	JA	22	5		0.095	0.0698	0.00536	3.95	4	2.84	3.91	3.16
Forst	3	JA	24	3	0.0339	0.219	0.0887	0.00953	4	3.9	3.15	3.39	3.27
KSU	1	JA	16	11	0	0.0224	0.0695	0.00178	4	3.78	2.34	4	3.13
KSU	2	JA	24	3	0.00097	0.0209	0.0772	0.00533	3.96	3.69	2.21	3.79	2.52
SML	1	JA	22	5			0.0646	0					
Forst	1	EN	22	5			0.00921	0					
Summrization Run													
TeamID	Priority	source	Lang.	#of ques	#of N/A	content score				quality score				
						Human	NUGGET	rouge_1	rouge_2	GRAMMATICALITY	NON_REDUNDANCY	REFERENCE	FLUENCY	COHERENCE
Forst	1	Exp	JA	5	0	0	0.00356	0.01	0.00118	4	3.6	2.5	4	2
Forst		GSN+ExP	JA	5	0	0	0.00356	0	0	4	3.6	2.5	4	2
Forst		GSN	JA	5	0	0	0.00698	0	0	4	3.8	3.5	4	3
KSU		ExP	JA	4	1	0	0.00991	0.0223	0.00182	4	3.13	2	3.75	2
KSU		GSN+ExP	JA	4	1	0	0.00991	0.0223	0.00182	4	3.13	2	3.75	2
KSU		GSN	JA	5	0	0.0587	0.0527	0.0659	0.0279	4	3.8	2.8	4	3.5

Table 14: Detail results of Essay questions in Phase-2

End-to-end Run												
TeamID	Priority	Lang.	#of ques	#of N/A	content score				quality score			
					Human	NUGGET	rouge_1	rouge_2	GRAMMATICALITY	NON_REDUNDANCY	REFERENCE	FLUENCY	COHERENCE
Forst	1	JA	27	0	0	0.00829	0.0385	0.0042	3.68	3.84	2.73	3.53	2.19
Forst	2	JA	21	6		0.073	0.068	0.0101	3.7	3.77	3.18	3.73	3.41
Forst	3	JA	21	6		0.0666	0.0627	0.0101	3.75	3.8	3.27	3.73	3.45
IMTKU	1	JA	21	6	0	0.00295	0.0277	0.000318	3.29	3.98	3.09	2.7	2.49
IMTKU	2	JA	19	8		0	0.00491	0	1.15	2.23	1.1	1.12	1.04
KSU	1	JA	26	1	0	0.0137	0.0517	0.00404	3.83	3.88	2.58	3.84	2.64
KSU	2	JA	26	1		0.0161	0.0548	0.00374	3.81	3.87	2.58	3.78	2.66
KSU	3	JA	27	0		0.0105	0	0	3.37	3.82	1.76	2.88	1.83
SML	1	JA	21	6		0.0178	0.0328	0.00751	3.64	3.95	2.57	3.57	3.19
SML	2	JA	22	5		0.0264	0.0346	0.00345	3.52	3.95	2.66	3.45	3.07
SML	3	JA	22	5		0.0264	0.0346	0.00345	3.52	0	2.68	3.43	3.07
CMUQA	1	EN	27	0	0	0.00241	0.0334	0.0063	2.93	3.44	2.41	2.96	1.73
CMUQA	2	EN	27	0		0.00241	0.0358	0.00635	2.93	3.41	2.41	2.93	1.76
CMUQA	3	EN	27	0		0.00241	0.0573	0.0124	1.74	3.04	1.85	1.9	1.3
Forst	1	EN	17	10		0.014	0.0177	0.0021	2.29	3.65	1.97	2.59	1.85
IMTKU	1	EN	16	11	0	0.00234	0.0441	0.00124	3.56	3.56	3.29	3.53	2.58
IMTKU	2	EN	15	12	0		0.0121	7.86E-05	1	1.07	1	1	1
MTMT	1	EN	26	1		0.00282	0.0593	0.00523	2.52	3.58	2.34	2.62	1.83
MTMT	2	EN	24	3	0	0.00787	0.0448	0.0019	2.31	3.31	2.19	2.65	1.79
Summrization Run													
TeamID	Priority	source	Lang.	#of ques	#of N/A	content score				quality score			
						Human	NUGGET	rouge_1	rouge_2	GRAMMATICALITY	NON_REDUNDANCY	REFERENCE	FLUENCY	COHERENCE
DGLab		Exp	JA	5	0	0	0.00641	0.0246	0.00169	4	2.87	3.5	3.7	2.47
DGLab	1	GSN+ExP	JA	5	0		0.0414	0.0603	0.0305	3.93	3.03	3.3	3.4	2.67
DGLab		GSN	JA	5	0		0.0464	0.0617	0.0317	4	3.03	3.3	3.4	2.67
DGLab		Exp	JA	5	0		0.0129	0.0229	0.000782	3.8	2.77	3.07	3.5	2.4
DGLab	2	GSN+ExP	JA	5	0		0.0468	0.0627	0.0299	3.93	3.03	3.1	3.4	2.57
DGLab		GSN	JA	5	0		0.0475	0.0627	0.0299	4	3.03	3.1	3.47	2.7
Forst		Exp	JA	5	0		0.00143	0.00797	0.000175	3.47	3.93	2.63	3.07	2.37
Forst	1	GSN+ExP	JA	5	0		0.00143	0	0	3.47	3.93	2.63	3.07	2.37
Forst		GSN	JA	5	0		0.00737	0	0	4	4	3.1	4	3.1
IMTKU	1	Exp	JA	5	0		0.00295	0	0	3.13	3.87	3.03	2.57	2.63
KSU		Exp	JA	5	0		0.0074	0.0252	0.00214	3.9	3.3	2.8	4	2.4
KSU	1	GSN+ExP	JA	5	0		0.00521	0.0264	0.00359	3.57	3.47	2.2	3.13	2.23
KSU		GSN	JA	3	2		0.0269	0.0682	0.0354	3.93	3.9	3.9	3.37	3.47
CMUQA	1	GSN	EN	5	0		0.0198	0.0708	0.0338	4	3.3	2.9	4	2.5
DGLab		Exp	EN	5	0	0	0.00335	0.0255	0.00249	2.1	2.5	2.6	1.5	1.5
DGLab	1	GSN+ExP	EN	5	0		0.0254	0.0635	0.0305	4	2.4	2.7	3.3	2.5
DGLab		GSN	EN	5	0		0.026	0.0636	0.0308	4	2.5	3	3.5	2.63
DGLab		Exp	EN	5	0		0.00321	0.026	0.00246	2.4	2.7	2.6	1.6	2
DGLab	2	GSN+ExP	EN	5	0		0.0288	0.066	0.0329	4	2.4	2.6	3.3	2.5
DGLab		GSN	EN	5	0		0.0292	0.0661	0.0329	4	2.8	3	3.4	2.5
IMTKU	1	Exp	EN	5	0		0.00262	0	0	3.8	3.4	3.2	3.5	2.4

Table 15: Detail results of Essay questions in Research run

End-to-end Run												
TeamID	Priority	Lang.	#of ques	#of N/A	content score			quality score				
					NUGGET	rouge_1	rouge_2	GRAMMATICALITY	NON_REDUNDANCY	REFERENCE	FLUENCY	COHERENCE
DG_Lab	1	JA	16	3	0.0278	0.0394	0.00505	3.78	3.91	3.31	4	2.84
DG_Lab	1	EN	19	0	0.0529	0.0247	0.00112	2.79	3.24	2.39	2.79	2.08
Forst	2	JA	16	3	0.0239	0.0203	0.00492	3.91	4	3.56	3.97	3.16
Forst	3	JA	16	3	0.0239	0.0197	0.00492	3.91	4	3.56	3.97	3.06

Table 16: Results of Extraction task at Phase 1 (N = 5)

TeamID	Priority	Lang	Passage Precision	Nugget Recall	Ave. of tokens
Forst	1	JA	0.667	0.162	1968.8
KSU	1	JA	0.517	0.319	1670.0
KSU	2	JA	0.398	0.151	2502.5

Table 17: Results of Extraction task at Phase 1 (N = 10)

TeamID	Priority	Lang	Passage Precision	Nugget Recall	Ave. of tokens
Forst	1	JA	0.667	0.162	2337.8
KSU	1	JA	0.517	0.319	1670.0
KSU	2	JA	0.398	0.151	2510.25

Table 18: Results of Extraction task at Phase 2 (N = 5)

TeamID	Priority	Lang	Passage Precision	Nugget Recall	Ave. of tokens
DGLab	1	JA	0.510	0.057	1875.6
DGLab	2	JA	0.479	0.044	1875.6
DGLab	3	JA	0.375	0.206	2534.6
Forst	1	JA	0.058	0.113	2656.0
Forst	2	JA	0.192	0.017	1905.4
IMTKU	1	JA	0.113	0.020	454.4
IMTKU	2	JA	0.000	0.000	336.25
KSU	1	JA	0.064	0.164	2652.6
KSU	2	JA	0.163	0.243	2680.8
KSU	3	JA	0.109	0.081	2683.8
CMUQA	1	EN	0.113	0.035	243.2
CMUQA	2	EN	0.088	0.026	274.2
DGLab	1	EN	0.087	0.035	1029.2
DGLab	2	EN	0.117	0.035	1029.2
IMTKU	1	EN	0.260	0.061	249.2
IMTKU	2	EN	0.234	0.058	249.2
MTMT	1	EN	0.016	0.041	1336.4
MTMT	2	EN	0.017	0.030	1325.6

Table 19: Results of Extraction task at Phase 2 (N = 10)

TeamID	Priority	Lang	Passage Precision	Nugget Recall	Ave. of tokens
DGLab	1	JA	0.771	0.130	4493.6
DGLab	2	JA	0.740	0.117	4493.6
DGLab	3	JA	0.600	0.281	4283.0
Forst	1	JA	0.086	0.155	5269.2
Forst	2	JA	0.364	0.033	5054.4
IMTKU	1	JA	0.113	0.020	454.4
IMTKU	2	JA	0.000	0.000	336.25
KSU	1	JA	0.086	0.203	5251.2
KSU	2	JA	0.184	0.270	4979.4
KSU	3	JA	0.195	0.126	5441.6
CMUQA	1	EN	0.113	0.035	243.2
CMUQA	2	EN	0.088	0.026	274.2
DGLab	1	EN	0.132	0.038	2366.4
DGLab	2	EN	0.162	0.038	2366.4
IMTKU	1	EN	0.260	0.061	249.2
IMTKU	2	EN	0.234	0.058	249.2
MTMT	1	EN	0.032	0.081	2546.4
MTMT	2	EN	0.050	0.077	2599.8

Table 20: Submissions of Japanese 2000's question in Essay Evaluation Method task in Phase-1

System Essay	Forst1	Forst2
Forst_e2e_01	3.60E1	5.00E0
Forst_e2e_03	2.90E1	3.00E0
Forst_summarization_ExP+GSN_01	2.00E0	0.00E0
Forst_summarization_ExP_01	2.00E0	0.00E0
Forst_summarization_GSN_01	4.00E0	7.00E0
KSU_e2e_01	2.90E1	0.00E0
KSU_e2e_02	2.90E1	0.00E0
KSU_summarization_ExP+GSN_01	2.60E1	0.00E0
KSU_summarization_ExP_01	2.60E1	0.00E0
KSU_summarization_GSN_01	2.00E1	1.60E1

Table 21: Submissions of Japanese 2004's question in Essay Evaluation Method task in Phase-1

System Essay	Forst1	Forst2
Forst_e2e_01	3.60E1	8.00E0
Forst_e2e_03	4.70E1	1.50E1
Forst_summarization_ExP+GSN_01	1.30E1	1.00E0
Forst_summarization_ExP_01	1.30E1	1.00E0
Forst_summarization_GSN_01	1.60E1	4.00E0
KSU_e2e_01	3.50E1	3.00E0
KSU_e2e_02	4.60E1	5.00E0
KSU_summarization_ExP+GSN_01	4.30E1	3.00E0
KSU_summarization_ExP_01	4.30E1	3.00E0
KSU_summarization_GSN_01	3.10E1	1.30E1

Table 22: Submissions of Japanese 2008's question in Essay Evaluation Method task in Phase-1

System Essay	Forst1	Forst2
Forst_e2e_01	5.00E1	8.00E0
Forst_e2e_03	3.80E1	1.00E0
Forst_summarization_ExP+GSN_01	7.00E0	0.00E0
Forst_summarization_ExP_01	7.00E0	0.00E0
Forst_summarization_GSN_01	8.00E0	2.00E0
KSU_e2e_01	4.10E1	1.00E0
KSU_e2e_02	3.30E1	4.00E0
KSU_summarization_ExP+GSN_01	3.30E1	4.00E0
KSU_summarization_ExP_01	3.30E1	4.00E0
KSU_summarization_GSN_01	3.20E1	1.10E1

Table 23: Submissions of Japanese 2012's question in Essay Evaluation Method task in Phase-1

System Essay	Forst1	Forst2
Forst_e2e_01	3.70E1	8.00E0
Forst_e2e_03	2.10E1	1.00E0
Forst_summarization_ExP+GSN_01	9.00E0	0.00E0
Forst_summarization_ExP_01	9.00E0	0.00E0
Forst_summarization_GSN_01	5.00E0	2.00E0
KSU_e2e_01	3.90E1	1.00E0
KSU_e2e_02	4.10E1	4.00E0
KSU_summarization_ExP+GSN_01	4.20E1	4.00E0
KSU_summarization_ExP_01	4.20E1	4.00E0
KSU_summarization_GSN_01	2.90E1	1.10E1

Table 24: Submissions of Japanese 2013's question in Essay Evaluation Method task in Phase-1

System Essay	Forst1	Forst2
Forst_e2e_01	2.70E1	4.00E0
Forst_e2e_03	2.40E1	4.00E0
Forst_summarization_ExP+GSN_01	9.00E0	0.00E0
Forst_summarization_ExP_01	9.00E0	0.00E0
Forst_summarization_GSN_01	2.00E0	0.00E0
KSU_summarization_GSN_01	2.90E1	1.30E1

Table 25: Submissions of Japanese 2001's question in Essay Evaluation Method task in Phase-2

System Essay	DGLab1	DGLab2	Forst1	Forst2	tmkff1
DGLab_summarization_ExP_01	6.91E-1	6.77E-1	3.10E1	2.00E0	0.00E0
Forst_e2e_01	8.86E-1	6.14E-1	3.90E1	3.00E0	0.00E0
IMTKU_e2e_01	1.12E0	1.85E0	6.20E1	4.00E0	0.00E0
KSU_e2e_01	1.01E0	7.09E-1	4.40E1	3.00E0	2.00E0

Table 26: Submissions of Japanese 2002's question in Essay Evaluation Method task in Phase-2

System Essay	DGLab1	DGLab2	Forst1	Forst2	tmkff1
DGLab_summarization_ExP_01	7.33E-1	7.28E-1	3.50E1	0.00E0	0.00E0
Forst_e2e_01	9.60E-1	6.70E-1	3.90E1	0.00E0	0.00E0
IMTKU_e2e_01	1.13E0	1.72E0	4.70E1	1.00E0	0.00E0
KSU_e2e_01	1.08E0	7.54E-1	4.20E1	1.00E0	0.00E0

Table 27: Submissions of Japanese 2006's question in Essay Evaluation Method task in Phase-2

System Essay	DGLab1	DGLab2	Forst1	Forst2	tmkff1
DGLab_summarization_ExP_01	6.58E-1	6.27E-1	1.80E1	2.00E0	0.00E0
Forst_e2e_01	9.31E-1	6.31E-1	2.40E1	1.00E0	0.00E0
IMTKU_e2e_01	1.06E0	1.63E0	3.10E1	1.00E0	0.00E0
KSU_e2e_01	9.85E-1	6.57E-1	4.10E1	4.00E0	3.00E0

Table 28: Submissions of Japanese 2010's question in Essay Evaluation Method task in Phase-2

System Essay	DGLab1	DGLab2	Forst1	Forst2	tmkff1
DGLab_summarization_ExP_01	7.34E-1	7.22E-1	3.80E1	1.00E0	5.00E0
Forst_e2e_01	9.90E-1	7.05E-1	4.30E1	4.00E0	0.00E0
IMTKU_e2e_01	1.13E0	1.64E0	5.90E1	2.00E0	0.00E0
KSU_e2e_01	9.72E-1	6.45E-1	4.80E1	1.00E0	4.00E0

Table 29: Submissions of Japanese 2014's question in Essay Evaluation Method task in Phase-2

System Essay	DGLab1	DGLab2	Forst1	Forst2	tmkff1
DGLab_summarization_ExP_01	6.77E-1	6.47E-1	4.50E1	5.00E0	0.00E0
Forst_e2e_01	9.07E-1	6.29E-1	5.30E1	4.00E0	4.00E0
IMTKU_e2e_01	1.11E0	1.72E0	6.70E1	1.00E0	0.00E0
KSU_e2e_01	9.44E-1	6.69E-1	6.00E1	7.00E0	4.50E0

Table 30: Submissions of English 2001's question in Essay Evaluation Method task in Phase-2

System Essay	DGLab1	DGLab2
CMUQA_e2e_01	8.59E-1	8.46E-1
DGLab_summarization_ExP_01	1.29E0	1.10E0
IMTKU_e2e_01	1.25E0	1.00E0
MTMT_e2e_01	1.36E0	9.71E-1

Table 31: Submissions of English 2002's question in Essay Evaluation Method task in Phase-2

System Essay	DGLab1	DGLab2
CMUQA_e2e_01	1.11E0	2.42E0
DGLab_summarization_ExP_01	1.38E0	1.12E0
IMTKU_e2e_01	1.23E0	8.86E-1
MTMT_e2e_01	1.31E0	9.23E-1

Table 32: Submissions of English 2006's question in Essay Evaluation Method task in Phase-2

System Essay	DGLab1	DGLab2
CMUQA_e2e_01	8.12E-1	7.98E-1
DGLab_summarization_ExP_01	1.22E0	8.89E-1
IMTKU_e2e_01	1.18E0	8.14E-1
MTMT_e2e_01	1.27E0	9.19E-1

Table 33: Submissions of English 2010's question in Essay Evaluation Method task in Phase-2

System Essay	DGLab1	DGLab2
CMUQA_e2e_01	9.30E-1	9.24E-1
DGLab_summarization_ExP_01	1.30E0	9.32E-1
IMTKU_e2e_01	1.22E0	9.49E-1
MTMT_e2e_01	1.32E0	9.59E-1

Table 34: Submissions of English 2014's question in Essay Evaluation Method task in Phase-2

System Essay	DGLab1	DGLab2
CMUQA_e2e_01	9.56E-1	9.44E-1
DGLab_summarization_ExP_01	1.30E0	9.72E-1
IMTKU_e2e_01	1.19E0	1.03E0
MTMT_e2e_01	1.31E0	9.69E-1