Forst: Question Answering System for Term and Essay Questions at NTCIR-13 QA Lab-3 Task

Kotaro Sakamoto
Yokohama National University
National Institute of Informatics
ksakamoto@nii.ac.jp

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

Yuta Fukuhara
Yokohama National University
yuta_f@forest.eis.ynu.ac.jp

Akihiro Iizuka
Yokohama National University
iizuka-a@forest.eis.ynu.ac.jp

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

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

We participated in all phases of the term question task and the essay question task in Japanese. We described changes since the QA Lab-2 and methods for the evaluation method subtask. Although the changes did not bring the major improvement, using 'implicit keywords' extracted from question texts makes the results better. The evaluation method using gold standard nuggets achieved the best results.

Team Name

Forst

Subtask

Japanese

Keywords

question answering, essay questions, term questions, university entrance examination, world history

1. INTRODUCTION

Question answering is widely regarded as an advancement in information retrieval. However, QA systems are not as popular as search engines in the real world. In order to apply QA systems to real-world problems we tackle the QA-Lab task dealing with Japanese university entrance exams of world history. Japanese university entrance exams have the following two stages: The National Center Test (multiple choice-type questions) and second-stage exams (essay questions and term questions).

The NTCIR-13 QA Lab-3 set the following three tasks: multiple-choice question task, term question task and essay question task. At the QAL Lab-3 a new subtask, evaluation-method subtask, is added to the essay question task. We tackled the term question task and the essay question task including the evaluation-method subtask, and reported the results and the consideration in this paper.

Basically, our systems for the term question task and the essay question end-to-end subtask are successors of our systems at the QA Lab-2. Therefore, we described main changes since the QA Lab-2.

2. RESOURCE

We used the following data as the knowledge source.

- four world history textbooks which QA Lab organizers supplied
- world history glossary (6,081 entry words)
- Q&A collection (4,324 Q&A pairs)
- world history event ontology [2][1]
- Japanese thesaurus (about 300,000 entry words)

3. TERM QUESTION ANSWERING SYSTEM

Figure 1 shows the term type question answering pipeline, which is the same as the QA Lab-2 Forst system.

Changes since the QA Lab-2 are as follows:

- The current system became able to answer questions that together asks different type things such as a title of work and the author,
- The later keywords appear in a question, the more emphasized they are in answer selection,
- Extending dictionary for named entities of world history,
- Adding decision rules for question types,

[1]http://researchmap.jp/zoeai/event-ontology-EVT/

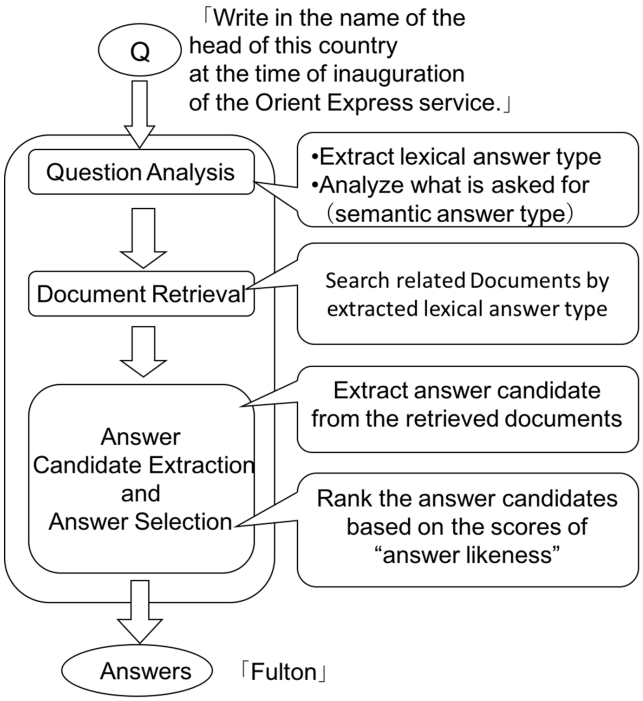

Figure 1: Term type answering pipeline

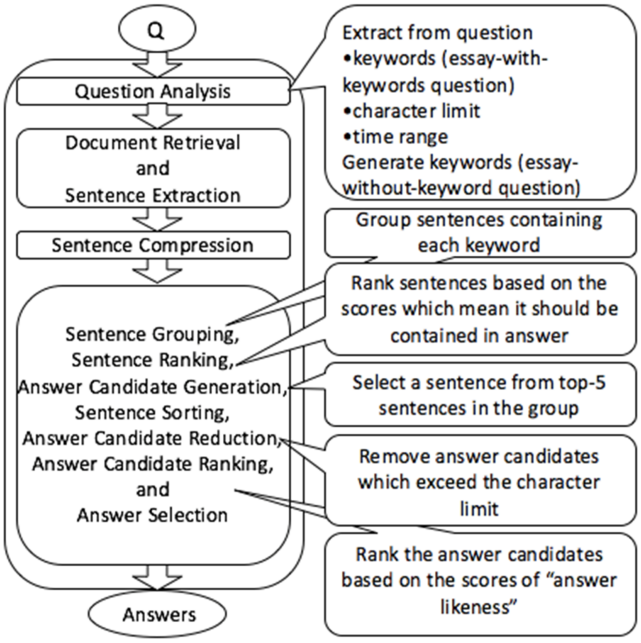

Figure 2: Essay type answering pipeline

- If the same answer is extracted from different documents, the scores are merged (majority decision).

4. ESSAY QUESTION ANSWERING SYSTEM

About the extraction and the summarization subtasks, there are no changes since the QA Lab-2. Therefore, we described the systems for the end-to-end and the evaluation-method subtasks.

4.1 For End-to-end Task

We developed four types of end-to-end systems. The first system (Priority 1 at Phase-1 and -2) is almost the same as one at QA Lab-2, as shown in Figure 2. The change is to add sentences from top in the MMR ranking [3] to answer when the answer is shorter than the length limitation. The flow of the second system (Priority 2 at Phase-1 and -2) is as follows:

- Extracting nouns from a given question,
- Selecting keywords using the world history named entity dictionary,
- Retrieving passages using the passage retrieval module of the term question answering system,
- Making essays by combining passages within the length limitation,
- Ranking essays by using keywords in essay.

The concept of the second system is use of 'implicit keywords' that are question focuses but not stated positively. For finding such implicit keywords, we assumed that they were represented by named entities in questions. The third system (Priority 3 at Phase-1) is based on the similarity using Q&A corpus, which is our own collection of Q&A pairs we collected from the past exams and so on. The system selects a Q&A pair of which question text is the most similar to given question text, and outputs the answer text of the pair if the answer text includes keywords. The third system is an example-based question answering, and focuses on simple essay questions of which length is smaller than 100 characters. The fourth system (Priority 3 at Phase-2) is the extend of the second system, which takes account of the word co-occurrence frequency between question and answer texts in the Q&A pairs.

We also developed a system for English end-to-end subtask. The system is the same as the Japanese first system. However, the knowledge source is translated by Google translation, and Stanford CoreNLP [4] is used for parsing.

4.2 For Evaluation Method Task

We developed two types of evaluation method systems. The first system (Priority 1 at Phase-1 and -2) is based on world history terms. Using the world history named entity dictionary, the system counts the number of terms in a essay, and the count is simply regarded as the essay score. The second system (Priority 2 at Phase-1 and -2) is based on a given set of gold standard nuggets. The system segments an essay into sentences by punctuation, and counts the number of nuggets that are matched with any one of the sentences. If more than one words in a nugget are included in a sentence, the nugget is matched with the sentence. The matching number is regarded as the essay score.

5. EXPERIMENTS

Table 1 shows the results of the term question task. In comparison with the QA Lab-2 results, the QA Lab-3 changes did not bring the improvement of the correct rates. The idea of majority decision did not always work well, sometimes

Table 1: Results in the term question task

	Priority	# of Ques	# of Correct	# of Incorrect	# of N/A	Correct Rate
Phase-1	1	68	27	41	0	0.397
	2	68	1	1	66	0.015
Phase-2	1	77	21	45	2	0.273

Table 2: Results in the essay question task

	Priority	Lang	# of Ans	# of N/A	Content				Quality				
					Human	Nugget	ROUGE-1	ROUGE-2	QQ1	QQ2	QQ3	QQ4	QQ5
End-to-end subtask													
Phase-1	1	JA	26	1	0.011	0.0221	0.0523	0.00351	3.96	3.69	2.56	3.81	2.23
	2	JA	22	5	-	0.095	0.0698	0.00536	3.95	4.00	2.84	3.91	3.16
	3	JA	24	3	0.0339	0.219	0.0887	0.00953	4.00	3.90	3.15	3.39	3.27
	1	EN	22	5	-	-	0.0092	0.00000	-	-	-	-	-
Phase-2	1	JA	27	0	0	0.00829	0.0385	0.00420	3.68	3.84	2.73	3.53	2.19
	2	JA	21	6	-	0.0730	0.0680	0.0101	3.70	3.77	3.18	3.73	3.41
	3	JA	21	6	-	0.0666	0.0627	0.0101	3.75	3.80	3.27	3.73	3.45
	1	EN	17	10	-	0.0140	0.0177	0.0021	2.29	3.65	1.97	2.59	1.85
Research	2	JA	16	3	-	0.0239	0.0203	0.00492	3.91	4.00	3.56	3.97	3.06
	3	JA	19	0	-	0.0239	0.0197	0.00492	3.91	4.00	3.56	3.97	3.06
Summarization subtask													
Phase-1	ExP	JA	5	0	0	0.00356	0.0100	0.00118	4.00	3.60	2.50	4.00	2.00
	GSN+ExP	JA	5	0	0	0.00356	0.0000	0.00000	4.00	3.60	2.50	4.00	2.00
	GSN	JA	5	0	0	0.00698	0.0000	0.00000	4.00	3.80	3.50	4.00	3.00
Phase-2	ExP	JA	5	0	-	0.00143	0.00797	0.000175	3.47	3.93	2.63	3.07	2.37
	GSN+ExP	JA	5	0	-	0.00143	0.00000	0.000000	3.47	3.93	2.63	3.07	2.37
	GSN	JA	5	0	-	0.00737	0.00000	0.000000	4.00	4.00	3.10	4.00	3.10

Table 3: Results in the evaluation-method task

	Priority	Spearman's Rho	Kendall's Tau-b
Phase-1	1	0.427	0.334
	2	0.596	0.534
Phase-2	1	-0.071	-0.049
	2	0.404	0.360

brought a wrong answer even if the first-ranked answer was correct.

Table 2 shows the results of the essay question task. The second system was better than the first system. This means that using the implicit keywords was effective. The third system achieved the best score, although the score obtained by answering only simple essay questions. In world history questions, example-based question answering is effective if there are enough amount of Q&A pairs. However, in the case of complex essay questions, there are little similar questions. The fourth system of which scoring module took account of co-occurrence in the Q&A pairs was not better than the second system that is the base system of the fourth system. Developing a hybrid system with example-based is future work.

Table 3 shows the results of the evaluation method task. The second system using the given gold standard nuggets was better than the first system using named entities, and obtained the best result among all submissions to the QA Lab-3. However, it could not exceed the Pyramid and the ROUGE scores. We will further research on the difference from human marks.

6. CONCLUSION

We participated in all phases of the term question task and the essay question task in Japanese. We described changes since the QA Lab-2 and methods for the evaluation method subtask. Although the changes did not bring the major improvement, using 'implicit keywords' extracted from question texts makes the results better. The evaluation method using gold standard nuggets achieved the best results.

7. REFERENCES

[1] Hideyuki Shibuki, Kotaro Sakamoto, Madoka Ishioroshi, Yoshinobu Kano, Teruko Mitamura, Tatsunori Mori, Noriko Kando. Overview of the NTCIR-13 QA Lab-3 Task. Proceedings of the 13th NTCIR Conference, 2017.

[2] 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.

[3] Jaime Carbonell, Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. SIGIR '98 Proceedings of the 21st annual international ACM SIGIR conference on

Research and development in information retrieval, pp. 335–336, 1998.

[4] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60, 2014.