ditlab at the NTCIR-17 QA Lab-PoliInfo-4 Task

Yuuki Tachioka
Denso IT Laboratory
Japan
tachioka.yuki@core.d-itlab.co.jp

ABSTRACT

The ditlab team participated in the Question Answering 2 subtask of the QA Lab-Poliinfo-4. First, we modified a QA Alignment system that has been developed for PoliInfo-3 QA Alignment subtask in order to make paragraphs composed of the related answer sentences. BM25 vectors were constructed for each paragraph of all answers and the target answers were selected by the question summaries and subtopics based on the cosine similarity. Second, a Text-to-Text Transfer Transformer (T5) was used to summarize the associated answer. For making fine-tuning data of T5, all data were used and the data selection based on the ROUGE scores was used.

KEYWORDS

BM25, Gale-Shapley algorithm, T5

TEAM NAME

ditlab

SUBTASKS

Question Answering 2 task (Japanese)

1 INTRODUCTION

The ditlab team participated in the Question Answering 2 tasks of the NTCIR-17 QA Lab-PoliInfo-4 task [4]. We use heuristic rules that were originally developed for the QA Alignment subtask of the NTCIR-16 QA Lab-PoliInfo-3 task [3, 7] to make paragraphs composed of related sentences for question and answer and we use BM25 [6] to calculate similarities between question summaries and subtopics and answer paragraphs. In this task, government minutes of the Tokyo Metropolitan Assembly are composed of sentences. Each sentence has a Q/A/O tag. For preparation of Question Answering, it is necessary to make paragraphs and associate a question with its answer. We do not use questions themselves. QA Alignment is performed in three steps (shown in Fig. 2). First step finds the corresponding part from the entire minutes by date and questioner ID. Second step combines multiple related sentences with “A” tags to form a paragraph. Third step matches question summaries and subtopics and answer paragraphs based on the similarities between them. We also developed a QA system that utilizes the results of the QA Alignment. In the fourth step, the associated answers are input to Text-to-Text Transfer Transformer (T5)[5] model in order to generate answer summary. The fifth step is optional. Training data were refined on the basis of ROUGE F-measures by comparing the generated answer summaries with correct summaries.

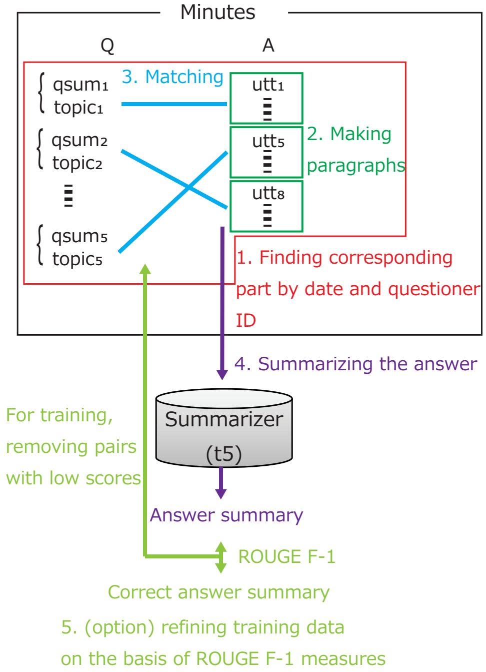

Figure 1: Finding answers by using question summaries and topics and summarizing the answers.

2 QA ALIGNMENT

2.1 Heuristic rules to make paragraphs

We can accurately combine sentences to make paragraphs by regular expressions that are optimized for the minutes because questions and answers in the Diet have a fixed format. Heuristic rules use fixed phrases at the beginning or the ending of the sentence that start paragraphs and fixed phrase at the ending of the sentence that terminates paragraphs [7].

2.2 BM25 for matching

After we prepared word n-grams by morphological analysis, processed by MeCab[1], BM25 models [6] were constructed on the morphemes excluding tokens, auxiliary verbs, and post-positional particles. The effectiveness of BM25 was shown in our PoliInfo-3 paper [7]. BM25 outperformed Bidirectional Encoder Representations from Transformers (BERT) [1] and Wikipedia2Vec [8]. BM25 values are high-dimensional sparse vectors that only have BM25 values at the existing morpheme. Cosine similarities between sparse vectors were used.

[1] https://taku910.github.io/mecab/

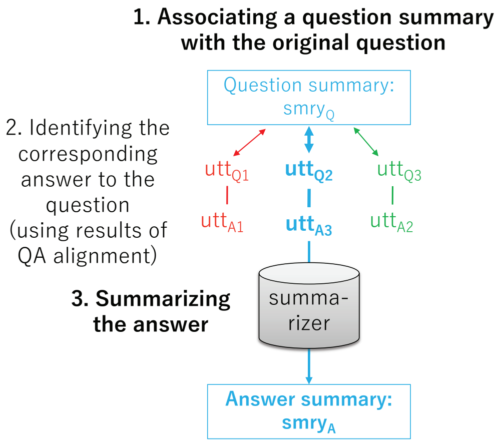
1. Associating a question summary with the original question
Question summary: smryQ
2. Identifying the corresponding answer to the question (using results of QA alignment)
uttQ1
uttQ2
uttQ3
uttA1
uttA3
uttA2
3. Summarizing the answer
summa-rizer
Answer summary: smryA

Figure 2: Linking questions and answers and summarizing the answers.

2.3 Matching algorithm

We used the hospital and resident[2] [2] matching algorithm, which is the most basic algorithm, for matching question summaries and subtopics and answer paragraphs.

3 QUESTION AND ANSWERING

3.1 Alignment

We took the approach of generating an answer summary direcltly from the original answer in minutes. In the PoliInfo-3, as shown in Fig. we have tried to find the original question asked in the Tokyo Metropolitan Assembly from a question summary as input, and then used the results of QA Alignment to find the original answer. However, this approach requires linking twice, i.e., linking between question summary and question and linking between question and answer. If one of the two linking phases fails, the correct answer cannot be obtained. In addition, the specification of corresponding questions is not necessarily needed for this task. In the PoliInfo-4, we directly associated question summary and subtopics with answers to reduce the influence of mistakes of alignment between question and answer.

3.2 Summarizing the answer

We utilized a commonly used transfer learning model, T5, as the model of the summarizer. The pre-trained model of the summarizer was **sonoisa/t5-base-japanese** trained with a 100 GB Japanese corpus. We fine-tuned the model using answer–answer summary pairs extracted from the training data.

3.3 Refinement of training data

There are some mistakes of the alignment of the question summary and the answer in the training data. To reduce this influence, we refined the training data on the basis of the ROUGE F-measure scores. For training data, once answer summaries were generated by using the T5 model that was fine-tuned on all training data and they were compared with the correct summaries in terms of ROUGE F-1 scores. If alignment was wrong, the generated answer summaries were different from the correct summaries. Thus, the pairs under the threshold were removed from the training data and again T5 models were fine-tuned on the refined data.

Table 1: ROUGE-1 F-measure results.

	ROUGE-1 F-measure
baseline (organizer)	0.2736
baseline with QA alignment (1)	0.3246
+refinement of training data (2)	0.3246

4 EXPERIMENTS

4.1 Experimental setup

We evaluated the performance by using "formal run" of the NTCIR-17 QA Lab-PoliInfo-4. BM25 models were trained using the distributed data[3] for "Himawari" derived from the minutes of the plenary session and budget committee of the national Diet. T5 models were used and fine-tuning was done according to the baseline code that was provided by the task organizers. The main difference between the provided baseline and our proposed method is that the baseline method input whole answers to T5 when generating answer summaries but our method only input the corresponding part selected from the original answers.

4.2 Results and discussion

Table 1 shows the ROUGE-1 F-measure. The selection of corresponding part significantly improved the performance. Refinement did not improve the performance in terms of ROUGE-1 F-measure, but the generated summaries were different.

Table 2 shows the examples of generated summaries by system (1) and (2). Although the ROUGE-1 F-measure scores are the same, the generated summaries are quite different.

Table 3 shows the human evaluation scores. The selection improved the all scores from the baseline. Without selection, the content was wrong but our content is as correct as the gold. With refinement, the correspondence and content scores were improved but the well-formed scores degraded.

5 CONCLUSION

In order to associate question summaries with answers, we refined heuristic rules that make a paragraph. We generated an answer summary from the original answer in minutes. In our method, we directly find the corresponding answer to the question summaries and subtopics using similarity calculation. The answer was input to T5 in order to generate answer summaries. When training T5, we refine training data on the basis of ROUGE F-measures. Experimental results showed that this method significantly improved the performance from the baseline that input whole answers to T5 and refinement improve the correspondence and content scores.

[2] https://pypi.org/project/matching/

[3] https://csd.ninjal.ac.jp/lrc/index.php

Table 2: Examples of generated answer summaries.

gold	コロナを乗り越え持続可能な回復を実現し積極果敢に施策展開する予算を編成した。都民ファーストの視点に立ち未来を切り開く。
(1)	緊急雇用対策や東京版ニューディール作戦、行政のデジタル化等施策を着実に実行し、都民ファーストの視点で未来を切り拓く。
(2)	コロナを乗り越え持続可能な回復、サステーナブルリカバリーの実現に向け、積極果敢に施策を展開し、都民ファーストの視点に立ち、東京の明るい未来を切り拓く。

Table 3: Human evaluation results.

	Correspondence				Content				Well-formed				Overall			
	A	B	C	Score	A	B	C	Score	A	B	C	Score	A	B	C	Score
baseline (organizer)	86	8	6	180	35	49	16	119	89	6	5	184	48	29	23	125
baseline with QA alignment (1)	94	5	1	193	46	48	6	140	92	8	0	**192**	67	22	11	**156**
+refinement of training data (2)	94	6	0	**194**	47	48	5	**142**	84	12	4	180	65	25	10	155
Gold	93	6	1	192	47	47	6	141	96	3	1	195	69	28	3	166

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Procceings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. 4171–4186.

[2] D Gale and L Shapley. 1962. College Admissions and the Stability of Marriage. *Americam Mathematical Monthly* 92 (1962), 261–268.

[3] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Kazuma Kadowaki, Tatsunori Mori, Kenji Araki, Teruko Mitamura, and Satoshi Sekine. 2022. Overview of the NTCIR-16 QA Lab-PoliInfo-3 Task. In *Procceings of The 16th NTCIR Conference on Evaluation of Information Access Technologies.* 156–174.

[4] Yasuhiro Ogawa, Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Kazuma Kadowaki, Tomoyoshi Akiba, Minoru Sasaki, Akio Kobayashi, Masaharu Yoshioka, Tatsunori Mori, Kenji Araki, Satoshi Sekine, and Teruko Mitamura. 2023. Overview of the NTCIR-17 QA Lab-PoliInfo-4 Task. In *Procceings of The 17th NTCIR Conference on Evaluation of Information Access Technologies.*

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv.

[6] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Journal Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.

[7] Yuuki Tachioka and Atsushi Keyaki. 2022. ditlab at the NTCIR-16 QA Lab-PoliInfo-3. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.* 207–210.

[8] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. In *Procceings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Association for Computational Linguistics, 23–30.