nukl’s QA System at the NTCIR-16 QA Lab-PoliInfo-3

Yasuhiro Ogawa  
Nagoya University  
Japan  
yasuhiro@is.nagoya-u.ac.jp

Yugo Kato  
Nagoya University  
Japan  
kato.yugo.c4@s.mail.nagoya-u.ac.jp

Katsuhiko Toyama  
Nagoya University  
Japan  
toyama@is.nagoya-u.ac.jp

ABSTRACT

Our nukl team participated in the NTCIR-16 QA Lab-PoliInfo-3’s question answering (QA) subtask. This paper describes the QA system for Japanese assembly member speeches using T5. We generated answer summaries using two input types: the answerer’s entire utterance and the answer text corresponding to the input question. We made two T5 models for each input type and determined the final output according to the length of the answerer’s utterance. Our system achieved the highest score in both automatic and human evaluations in this subtask.

KEYWORDS

question answering, summarization, T5

TEAM NAME

nukl

SUBTASKS

Question Answering (Japanese)

1 INTRODUCTION

NTCIR-16’s QA Lab-PoliInfo-3 [2] (Question Answering Lab for Political Information 3) dealt with political information and set out four subtasks: Question Answering (QA) alignment, question answering (QA), fact verification, and budget argument mining. Our team participated in the QA subtask.

We previously participated in NTCIR-14’s QA Lab-PoliInfo and, during its summarization task, developed a new summarization system: Progressive Ensemble Random Forest (PERF) [9]. Our system achieved the best performance in the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score evaluation. We also participated in the dialog summarization subtask in NTCIR-15’s QA Lab-PoliInfo-2, where we applied PERF and achieved good performance, but did not outperform the system using deep learning [8].

The QA subtask in QA Lab-PoliInfo-3 is formally a QA task, but it requires an answer summarizing the answerer’s utterance rather than a simple answer phrase. Thus, we considered this subtask as a type of summarization task. However, rather than apply PERF, we used T5 [11] based on deep learning.

We applied two methods to the task: one directly using T5 and the other using T5 with a QA alignment result by another system. Finally, we proposed a system that integrates the two methods, which achieved the best results in automatic and manual evaluations.

This paper is organized as follows. In Section 2, we discuss related works. Next, we describe our proposed methods in Section 3 and their experiments in Section 4. We provide discussion in Section 5 and, finally, we conclude in Section 6.

2 RELATED WORKS

This section briefly discusses past works on QA and summarization.

2.1 Related Works on Question Answering

QA has been actively studied. Deep learning studies are prevalent and many studies use pre-trained models. For example, a study using the pre-trained model T5 [11] achieved state-of-the-art in the QA task SQuAD [12].

In tasks such as SQuAD, QA systems search relatively short text to answer questions. On the other hand, in PoliInfo-3’s QA subtask, QA systems need to search long text, including multiple topics and multiple answerers’ answers. Therefore, it is uncertain whether a conventional system such as T5 can be directly applied to this QA subtask.

2.2 Related Works on Summarization

This PoliInfo-3’s QA subtask is called a QA task, but it requires an answer summarizing the answerer’s utterance rather than a simple answer phrase. In that sense, it can be said to be a kind of summarization task.


The summarization subtask in PoliInfo was an ordinal summarization task. Although one speaker’s utterance includes multiple questions or answers, the input in this subtask is only one question or answer text.

The summarization subtask in PoliInfo-2 was different from PoliInfo and is called dialog summarization. Its purpose is to summarize a transcript based on the dialogue structure, which consists of an assembly member’s question and a prefectural governor’s or superintendent’s answer. When the speaker’s utterance includes multiple questions or answers, we need to find the most relevant text to the input subtopic and summarize it. This task requires summarizing both a question and its answer.

The PoliInfo-3’s QA subtask gives us a question’s summary and requires us to output its answer. The input question’s summary is more useful than a subtopic in the PoliInfo-2’s subtask, so we can use another approach to find an appropriate text from the answerer’s utterance that contains multiple answers.

3 PROPOSED METHODS

Since the T5 model achieved a good summarization result, we use it to summarize the answer text. Thus, the problem we tackled next is how to find the answer text area from the input answerer’s utterance.

As described in the overview paper [2], when an input question is given, its answerer’s name is also provided, making it easy
to find the answerer’s entire utterance. Of course, this utterance contains several answers, so we need to find an appropriate text aligned to the input question. We propose two approaches to this problem and ultimately choose one depending on the length of the answerer’s utterance.

We describe the two approaches in Sections 3.1 and 3.2, respectively, and illustrate how to choose the appropriate one in Section 3.3.

3.1 Method 1: Input the Entire Utterance

The first method is to input the answerer’s entire utterance into T5, which will make T5 find an appropriate text for summarization.

We concatenate the input question, its subtopic, and the answerer’s entire utterance using a comma (,) as a separator. Note that subtopics are given with questions. The input question is a summary of the actual question utterance. Since this summary assumes a subtopic, the keywords in the subtopic are often omitted. Therefore, we considered it would not be easy to find an appropriate text from the answerer’s utterance only with the summarized question, so we provided a subtopic with the input.

Then, we tokenized the concatenated text by SentencePiece [5] and input it into T5. However, an entire utterance can be long and sometimes exceeds the input limit of T5, so we selected the maximum number of last sentences from the utterance within the limit. We chose the last sentences because, in assembly, answerers often first touch on the topic of the question, then talk about the current situation, and finally talk about solutions or future measures. Thus, the last sentences are usually essential.

Figure 1 shows the details, where the input is the entire utterances of a governor’s answer on September 26, 2001, and contains 123 sentences. The limit is 1,024 and the ’sum of the number of tokens’ indicates the sum from the last sentence. The sum of the last 22 sentences is less than 1,024, but that of the last 23 sentences is more than 1,024, so we used the last 22 sentences for the T5 input.

3.2 Method 2: Input Aligned Text

The second method uses the result of the QA alignment subtask in PoliInfo-3, where we leave the alignment between questions and answers to the QA alignment system and make T5 only summarize. The QA alignment system divides questioners’ utterances into some questions and answerers’ utterances into some answers, respectively, in the assembly minutes. Then, it aligns questions with their corresponding answers. This result gives us the appropriate answer text for the question. However, an input question in the QA subtask is not a separate question but its summary. Figure 2 shows an example.

We therefore found the original question text for the input question by calculating their similarity. We used word matching considering duplication as follows: First, we did morphological analysis of the input question by MeCab [6] and picked up content words whose part-of-speech is noun, verb, adjectival, quasi-adjectival, adverb, or adnominal adjective. We also did morphological analysis of the original question and picked up content words.

We used the SentencePiece tokenizer for T5 but, because it does not offer parts-of-speech, we used MeCab to calculate the similarity.
4 EXPERIMENTS

This section describes our experimental setting and the formal run results.

4.1 Experimental Setting

We used the GPU environment of Google Colaboratory and a pre-trained T5 model with published Japanese data\(^2\). We used pre-trained SentencePiece [5] with Japanese data as a tokenizer since the T5Tokenizer was built based on SentencePiece. Table 1 shows the experimental settings.

<table>
<thead>
<tr>
<th>Subtopic</th>
<th>Method 2 (without human evaluation)</th>
<th>Proposed (ROUGE-1-F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>System</td>
<td></td>
</tr>
<tr>
<td>316</td>
<td>Method 2</td>
<td>0.2787</td>
</tr>
<tr>
<td>266</td>
<td>Method 1</td>
<td>0.2823</td>
</tr>
<tr>
<td>311</td>
<td>Proposed (θ = 1,000)</td>
<td>0.3013</td>
</tr>
<tr>
<td>313</td>
<td>Proposed (θ = 2,500)</td>
<td>0.3129</td>
</tr>
<tr>
<td>310</td>
<td>Proposed (θ = 2,000)</td>
<td>0.3132</td>
</tr>
</tbody>
</table>

\(^2\)https://huggingface.co/sonoisa/t5-base-japanese

Method 2 requires a QA alignment result; we used the ID 235 result submitted by the AKBL team [10], which achieved the best score.

Our proposed method uses threshold θ, where we chose 1,000, 2,000, and 2,500 because we set the maximum input length as 1,024 tokens. We surmised that the number of tokens may be less than 1,024 if the length of the input text is less than 2,500. We also tried Methods 1 and 2 for comparison.

The number of training data in Method 1 is 7,627 tuples, consisting of an input question, its subtopic, its answerer’s entire utterances, and its correct answer. The training data in Method 1 is all data from 2001 to 2019 provided by Task Organizer [2]. The number of training data in Method 2 is 2,171 tuples, consisting of an input question, appropriate answer text, and its correct answer. We consider the gold data of the QA alignment subtask as the appropriate answer text. Task Organizer provided the data only from 2011 to 2016. Thus, the training data in Method 2 is smaller than that in Method 1.

The number of test data is 416, as described in the overview paper [2].

4.2 Experimental Results

In the PoliInfo-3 QA subtask, there are two types of evaluation. One is automatic evaluation using the ROUGE-1 F-measure [7] and the other is the human evaluation of four people.

Table 2 shows the automatic evaluation result. ID 166 indicates the baseline result submitted by Task Organizer. IDs 288 and 190 indicate the highest score by other teams.

The proposed method (θ = 2,000) achieved the highest score in this automatic evaluation and we submitted its output to the human evaluation.

Method 1 was inferior to the proposed method. This is because an answerer, especially a governor, sometimes answers many questions, but Method 1 only uses the last sentences, as shown in Figure 1. We will discuss the input length in the next section.

Method 2 was inferior to the proposed method and Method 1. We think this was caused by the training data size and will discuss it in the next section.

Table 3 shows the human evaluation results, where ID 310 indicates the result of the proposed method (θ = 2,000). All other results are described in the overview paper [2]. The proposed method also achieved the best result among the participants.
5 DISCUSSION

In this section, we discuss our experimental results. We first check the output of the proposed method in Section 5.1 and then investigate the distributions of the input utterances in Section 5.2. In Section 5.3, we describe which method is used in the proposed method. Finally, we carry out an additional experiment using the gold data of the QA alignment subtask in Section 5.4.

5.1 Example Outputs

Figure 3 shows some examples of the proposed method’s outputs. We provided the English translation. The proposed outputs are good answers in Examples 1 and 2. In Example 3, the output resembles the gold standard but the year is wrong, where “27 年” indicates “Heisei 27 (2015)” and “2 年” indicates “Reiwa 2 (2020).” This mistake might cause fake news, but it is not easy to correct. Neural summarization systems or neural translation systems might output the expression, but not in the original. In addition, in this case, the year was indicated by “来年 (next year)” in the original text, so we need to determine the specific year using non-textual information.

5.2 Distribution of Input Utterances

Since the maximum input length limit for T5 is 1,024 tokens in our experiments, we selected the last sentences as the input for some long utterances. We investigated the distribution of the sentence length of the training and test data as shown in Figures 4 and 5.

Figure 4(a) shows the distribution of the utterance length of the training data. The maximum limitation of T5 was also applied in the training process, so we shortened the training data. Figure 4(b) shows the distribution. Sentences over 1,024 tokens were shortened and included in the histogram into 800-1000 or 1000-1200.

Figures 4(c) and 4(d) show the same data in Method 2. Notice that the number of training data in Method 2 is smaller than that in Method 1 as described in Section 4.1. Since the input in Method 2 is selected text from the speaker’s entire utterances, its length is shorter than that in Method 1 and most are below 1024 tokens.

Figure 5 shows the distribution of the sentence length of the test data. While the original data in Method 1 includes some long sentences, that in Method 2 has no long sentences. These results imply that 1024 tokens are enough for Method 2.

Figure 5(e) shows the distribution in the proposed method, where Method 1 applied shorter sentences and Method 2 applied longer sentences.

5.3 Choice of Method 1 or Method 2

Our proposed method chooses the summary from the results of Methods 1 and 2 by the length of the answerer’s utterance. Table 4 shows which method was chosen in the test data consisting of 416 sentences. Although we chose the methods by the character length in the formal run, we should choose them by the token length. Thus, we investigated the test data’s token length and, fortunately, the result was the same as that of 2,000 characters, as shown in the last row in the Table 4.

We also investigated whether Method 1 or Method 2 was applied to the 100 sentences evaluated by the four people, as shown in Table 5. Notice that using Method 1 indicates that the answerer’s entire utterance is shorter than 2,000 characters. Table 5 illustrates that Method 1 produced a better result than Method 2, which implies that T5 can find appropriate text for summarization without the result of the QA alignment subtask for short utterances.

5.4 Using Correct Alignment

In the formal run, we used the AKBL team’s QA alignment result, which included some mistakes. After the formal run, the gold standard data of the QA alignment task was opened, so we used it for our methods, as shown in Table 6.

The gold data improved both the proposed method and Method 2. Although Method 2 with the gold data used the correct input data, it is inferior to the proposed method. This is because the training data size in Method 1 is larger than that in Method 2. The proposed method used Method 1 for shorter utterances, which produced better results than Method 2. This suggests that more training data for Method 2 may improve our system.

6 CONCLUSIONS

This paper described our QA system for the NTCIR-16 QA Lab-PollInfo-3. We used T5 to summarize the answer text, where the input was chosen from an answerer’s entire utterance or corresponding text to the input question by the length of the utterance. We achieved the best performance in both the ROUGE-1 F-measure evaluation and the human evaluation.

We consider that increasing the correct answer data of the QA alignment will improve performance. In addition, we will solve the problem of determining a correct value for notations such as “this year.”

REFERENCES

[1] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamatsu, Maoika Ishikosho, Teruko Mitamura, Masaharu Yoshioka, Tomoyoshi Akiba, Yushiro Ogawa, Minoru Sasaki, Kenichi Yokote, Tatsunori...
<table>
<thead>
<tr>
<th>ID</th>
<th>Input Question</th>
<th>System Output</th>
<th>HE</th>
<th>Gold Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>① カジノの誘致はしないと明言すべき。いつもまでに明確にするのかも含め、見解は。(You should state that we will not attract casinos. What are your views, including when you will clarify?)</td>
<td>国の動向を注視し、引き続き総合的に検討する必要がある。(It is necessary to pay close attention to the government’s trends and continue to consider them comprehensively.)</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>受け入れ施設の確保や今後の整備も含め果敢な行動を。(Take bold action, including securing reception facilities and future maintenance.)</td>
<td>感染の疑いのある患者の検査を迅速に実施。感染の疑いのある方の経過観察施設の機能等に関する調査検討を行う。(Promptly test patients with suspected infection. We will investigate and examine the functions of follow-up facilities for those suspected of being infected.)</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>緑地整備を進めるべき。(You should proceed with green space maintenance.)</td>
<td>27年度早期に整備方針改定し、区市町との連携を更に深めながら整備を加速。(We revised the maintenance policy in early 2015 and accelerated maintenance while further deepening cooperation with wards, cities, and towns.)</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>4</td>
<td>コロナで経済的格差が明確に、国と連携し生活底上げを。(The economic disparity becomes clear in the corona crisis. Work with the government to improve the standard of living.)</td>
<td>区市町村と連携し、各学校が現状に即した指導計画への再構築を行う。(We will work with the municipalities to restructure each school into a teaching plan in line with the current situation.)</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

HE† indicates the result of overall evaluations by four people using a three-grade system: A, B, and C.

Figure 3: Example of the Proposed Method Outputs and Their Human Evaluations

Table 4: Selected Method

<table>
<thead>
<tr>
<th>Threshold θ</th>
<th>Method 1</th>
<th>Method 2</th>
<th>Ratio†</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000 characters</td>
<td>184</td>
<td>232</td>
<td>0.44</td>
</tr>
<tr>
<td>2,000 characters</td>
<td>277</td>
<td>139</td>
<td>0.67</td>
</tr>
<tr>
<td>2,500 characters</td>
<td>296</td>
<td>120</td>
<td>0.71</td>
</tr>
<tr>
<td>1,024 tokens</td>
<td>277</td>
<td>139</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Ratio† indicates the percentage of test data to which Method 1 was applied.

Table 5: Comparison of Methods 1 and 2

<table>
<thead>
<tr>
<th>Method</th>
<th># of answers</th>
<th>Correspondence</th>
<th>Content</th>
<th>Well-formed</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A   B   C</td>
<td>A   B   C</td>
<td>A   B   C</td>
<td>A   B   C</td>
</tr>
<tr>
<td>Method 1 (&lt;2,000) (ratio %)</td>
<td>75</td>
<td>278 28 4</td>
<td>111 165 24</td>
<td>282 18 0</td>
<td>120 157 23</td>
</tr>
<tr>
<td>Method 2 (≥2,000) (ratio %)</td>
<td>25</td>
<td>85 7 8</td>
<td>27 46 27</td>
<td>99 1 0</td>
<td>28 46 26</td>
</tr>
</tbody>
</table>
(a) Length of Original Training Data in Method 1  
(b) Length of Shortened Training Data in Method 1 
(c) Length of Original Training Data in Method 2  
(d) Length of Shortened Training Data in Method 2 

Figure 4: Distribution of Length of Training Data

(a) Original Length in Method 1  
(b) Shortened Length in Method 1 
(c) Original Length in Method 2  
(d) Shortened Length in Method 2 
(e) Original Length in Proposed  
(f) Shortened Length in Proposed 

Figure 5: Distributions of Length of Test Data
Table 6: Scores with Correct Alignment Data

<table>
<thead>
<tr>
<th>system</th>
<th>ROUGE-1-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed ($\theta = 2,000$) with gold data</td>
<td>0.3333</td>
</tr>
<tr>
<td>Proposed ($\theta = 2,000$)</td>
<td>0.3132</td>
</tr>
<tr>
<td>Method 2 with gold data</td>
<td>0.3049</td>
</tr>
<tr>
<td>Method 1</td>
<td>0.2823</td>
</tr>
<tr>
<td>Method 2</td>
<td>0.2787</td>
</tr>
</tbody>
</table>


