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.

PoliInfo [3, 4] and PoliInfo-2 [1] offered summarization subtasks for the Japanese assembly minutes.

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][1] and picked up content words whose part-of-speech is noun, verb, adjective, quasi-adjective, adverb, or adnominal adjective. We also did morphological analysis of the original question and picked up content words.

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

Tokenized Sentence	# of Tokens	Sum of Tokens
次いで 、 住宅 政策 の 改革 について ... が必要 だと思います 。	35	1052
今回は 、 都営 住宅 の 抜 本 的 改革 ... 手 ぬ る い と 。	128	1017
民間 に 相談 して 、 もっと 大きな 容積 率 ... や っち まえ と 。	18	889
それ で 、 それを 国 が が たが たい ... 構 わない から や れ と 。	32	871
役人 は び く び く する けれど も 、 ... から や れ と 。	26	839
民間 に 相談 して 、 どこ まで だった ら ... を持って こい と 。	32	813
それを 行 っちゃ うこと で 国 は ... ざるを得ない でしょう 。	28	781
次いで 、 今後の 福祉 改革 の ... することで あります 。	51	753
これまで 、 福祉 改革 推進 プラン に基づき ... まい り ました 。	38	702
今後 も こうした 理念 を 、 高齢者 、 障害者 、 ... おり ます 。	45	664
先 般 開 所 しました 駅前 の 、 <unk> ... ところ だと思います 。	21	619
次いで 、 都 立 病院 改革 会議 の報告 ... ところ だと思います 。	28	598
それが 非常に 過剰 に オーバー ラップ ... と思って おり ます 。	85	570
今後 、 報告 内容を 十分に 尊重 しま ... と思って おり ます 。	22	485
この ためにも 、 年 内 を 目 途 に マスター ... だと思います 。	38	463
該当する 地域の 方 々 は 、 この 病院 の性格 ... だと思います 。	85	425
今 度の 報告 も そういう 視点 ... と思って おり ます 。	12	340
最後に 、 新たな 都 立 大学 のイメージ ... だい ており ます 。	46	328
構成 委員 が非常に 熱心な 余り 、 すぐ ... 大学 に しようと 。	75	282
ただ 、 石原 さん 、 新しい 大学 ... 考えて いただき たい 。	48	207
西 澤 先生 は 、 そこで 非常に 該 博 な ... をつくり たい と 。	28	159
一方 、 アメリカ のように 、 ビジネス ... 期待 しており ます 。	118	131
その他の 質問 については 、 教育 長 及び ... 答弁 いたします 。	13	13

Figure 1: Example of Text Shortening

Second, we counted up the number of content words in both the input and the original questions. If a word occurred in the original question twice, we counted it only once. However, if a word occurred in the input question twice, we counted it twice because duplicate occurrence in the input question is important. We consider this number a similarity and find the most similar one from the questioner's questions.

Third, we found the corresponding answer to the question using the result of the QA alignment subtask.

Finally, we concatenated the input question and the answer as we did in Method 1 and input it to T5. Notice that we did not use a subtopic in Method 2; we used it as a clue to find the appropriate text in Method 1, but the QA alignment system finds the appropriate text so we do not need it. If the concatenation was longer than the limit, we shortened it as in Method 1.

3.3 Proposed Method: Mixed

Method 1 and Method 2 each have their drawbacks. In the case of Method 1, if the answerer's utterance is long, its beginning is deleted, so the part that should be summarized may be missing from the input. In the example shown in Figure 1, only 22 sentences out of 123 sentences are used and the rest are not. In the case of Method 2, if the result of the QA alignment subtask is wrong, it results in the wrong answer.

Therefore, we considered selecting both methods according to the length of the answerer's utterance; we call this method the proposed method. The parameter θ indicates the threshold of the

Subtopic

産業振興

Input Question (Summary)

中小企業・小規模企業振興条例の理念に基づき、活力ある地域社会をつくり雇用の創出を。

Original Question (from the Minutes)

東京都**中小企業・小規模企業振興条例**についてお伺いいたします。事業所数において都内企業の九九%を占める中小企業の成長は、東京都の成長と発展の根幹であります。経済のグローバル化、ICT技術の進展、生産年齢人口の減少など、都内中小企業を取り巻く環境が大きく変化する中では、都内の中小企業振興に関する基本的な考え方を、都民の代表である都議会の意思も反映された条例として制定することは極めて重要です。また、先般公表されました森記念財団都市戦略研究所による世界の都市総合ランキングにおいては、東京のスタートアップ環境、つまり、新規創業環境の弱さが指摘されております。この課題を克服するためには、都内における産業の集積を生かし、大手企業、研究機関、創業支援機関など、さまざまな関係者が連携し、新たなイノベーションやユニコーンと呼ばれるベンチャー企業を生み出す環境整備を進める必要があります。また、中小企業、小規模企業は、都内経済を支えるとともに、都民の暮らしも支えております。都内在住の事業者や従業員は、地域のまちづくりに欠かすことのできない人材でもあります。都として、**条例に掲げる理念に基づき、**中小企業、小規模企業の業績向上や、ものづくり、事業を継承する支援を進めることで、にぎわいと**活力のある地域社会をつくり、雇用の創出**にも積極的に取り組むべきと考えますが、知事の見解を伺います。

Bold words indicate that they appear in the input question.

Figure 2: Example of Input Question and Its Original

Table 1: Experimental Setting

Maximum input length	1,024 tokens
Maximum output length	64 tokens
Number of epochs	4
Batch size	2

number of characters. If the answerer's utterance is longer than θ, we used Method 2; if not, we used Method 1.

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.

[2]https://huggingface.co/sonoisa/t5-base-japanese

Table 2: Scores in the Formal Run (ROUGE F-measure)

ID	System	ROUGE-1-F
310	Proposed ($\theta = 2,000$)	**0.3132**
313	Proposed ($\theta = 2,500$)	0.3129
311	Proposed ($\theta = 1,000$)	0.3051
266	Method 1	0.2823
316	Method 2	0.2787
288	ditlab	0.3013
190	AKBL	0.2306
166	TO	0.0767

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 ($\theta = 2,000$) 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 ($\theta = 2,000$). All other results are described in the overview paper [2]. The proposed method also achieved the best result among the participants.

Table 3: Scores in the QA Subtask in the Formal Run (Human Evaluation Results)

ID	Team	Correspondence				Content				Well-formed				Overall			
		A	B	C	Score	A	B	C	Score	A	B	C	Score	A	B	C	Score
Gold		377	20	3	774	208	170	22	586	391	8	1	790	217	164	19	598
310	**nukl**	363	25	12	**751**	138	211	51	**487**	381	19	0	781	148	203	49	**499**
288	ditlab	348	33	19	729	138	200	62	476	379	17	4	775	142	200	58	484
269	ditlab	346	31	23	723	129	209	62	467	384	16	0	**784**	136	207	57	479
190	AKBL	320	42	38	682	104	196	100	404	381	6	13	768	103	203	94	409
166	TO	83	77	240	243	4	58	338	66	99	33	268	231	4	36	360	44

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-PoliInfo-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 Takamaru, Madoka Ishioroshi, Teruko Mitamura, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Tatsunori

ID	Input Question	System Output	HE†	Gold Standard
1	カジノの誘致はしないと明言すべき。いつ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?)	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.)	A A A A	国は法に基づき基本方針を公表するとしており、今後の動向を注視し総合的に検討。(The government has stated that it will announce its basic policy based on the law and carefully consider future trends.)
2	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.)	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.)	B B A A	国や医療関係者と緊密に連携し検査・医療体制を強化。感染の疑いのある方の経過観察を行う一時滞在施設の調査・検討を行う。(We will strengthen the inspection and medical system in close cooperation with the government and medical personnel. We will investigate and examine temporary accommodation facilities to follow up on suspected infections.)
3	緑地整備を進めるべき。(You should proceed with green space maintenance.)	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.)	B B B C	2 年度早期に整備方針を改定，区市町との連携を更に深めながら整備を加速させ、ゆとりと潤いのある東京の実現を図る。(We will revise the maintenance policy as early as 2020 and accelerate the maintenance while further deepening the cooperation with the wards, cities, and towns, and aim to realize a comfortable and charming Tokyo.)
4	コロナで経済的格差が鮮明に。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.)	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.)	C C C C	生活資金の無利子貸し付け等を講じている、支援を国の取組含め検索できるサイトを立ち上げ、情報が届く仕組みも整えている。(We have set up a site where you can search for support, including governmental efforts, such as interest-free lending of living funds, and have a system to receive information.)

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

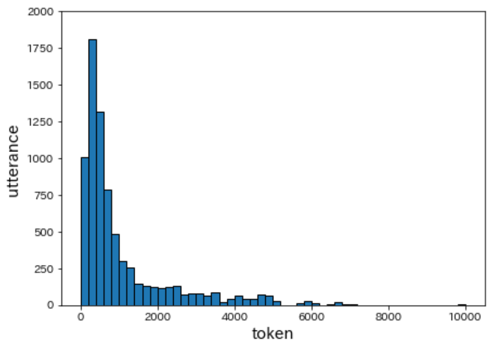
Table 4: Selected Method

Threshold θ	Method 1	Method 2	Ratio†
1,000 characters	184	232	0.44
2,000 characters	277	139	0.67
2,500 characters	296	120	0.71
1,024 tokens	277	139	0.67

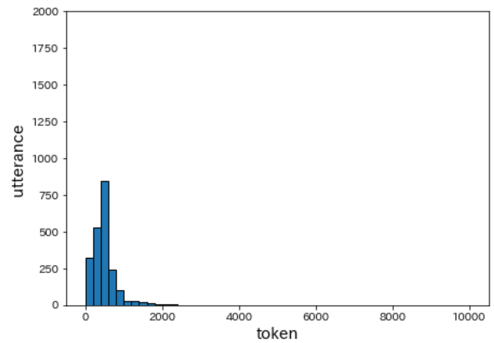
Ratio† indicates the percentage of test data to which Method 1 was applied.

Table 5: Comparison of Methods 1 and 2

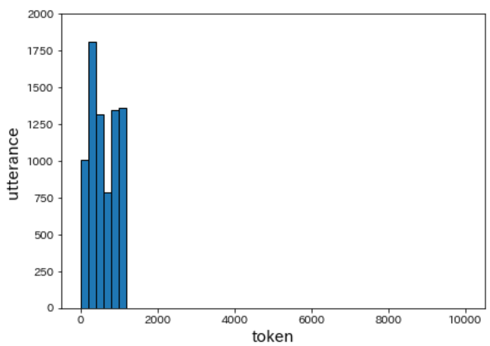
Method	# of answers	Correspondence			Content			Well-formed			Overall		
		A	B	C	A	B	C	A	B	C	A	B	C
Method 1 (< 2,000)	75	278	28	4	111	165	24	282	18	0	120	157	23
(ratio %)		92.7	6.0	1.3	37.0	55.0	8.0	94.0	6.0	0.0	40.0	52.3	7.7
Method 2 (≥ 2,000)	25	85	7	8	27	46	27	99	1	0	28	46	26
(ratio %)		85.0	7.0	8.0	27.0	46.0	27.0	99.0	1.0	0.0	28.0	46.0	26.0



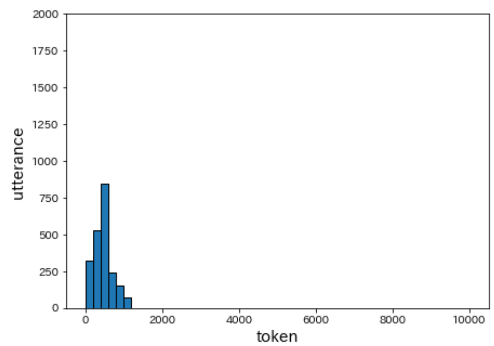
(a) Length of Original Training Data in Method 1



(c) Length of Original Training Data in Method 2

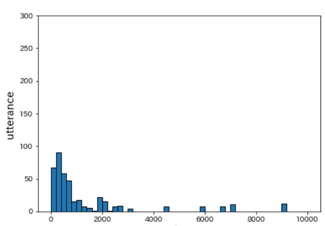


(b) Length of Shortened Training Data in Method 1

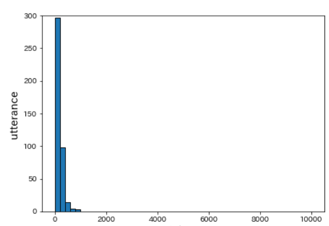


(d) Length of Shortened Training Data in Method 2

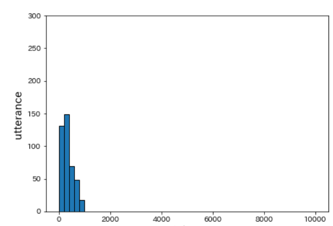
Figure 4: Distribution of Length of Training Data



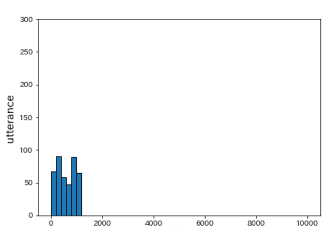
(a) Original Length in Method 1



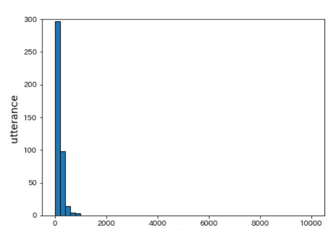
(c) Original Length in Method 2



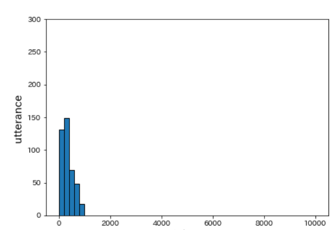
(e) Original Length in Proposed



(b) Shortened Length in Method 1



(d) Shortened Length in Method 2



(f) Shortened Length in Proposed

Figure 5: Distributions of Length of Test Data

Table 6: Scores with Correct Alignment Data

system	ROUGE-1-F
Proposed ($\theta = 2,000$) with gold data	0.3333
Proposed ($\theta = 2,000$)	0.3132
Method 2 with gold data	0.3049
Method 1	0.2823
Method 2	0.2787

Mori, Kenji Araki, Satoshi Sekine, and Noriko Kando. 2020. Overview of the NTCIR-15 QA Lab-PoliInfo-2 Task. In *Proceedings of The 15th NTCIR Conference*.

[2] 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 *Proceedings of The 16th NTCIR Conference*.

[3] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura, Noriko Kando, Tatsunori Mori, Harumichi Yuasa, Satoshi Sekine, and Kentaro Inui. 2019. Final Report of the NTCIR-14 QA Lab-PoliInfo Task. In *NII Conference on Testbeds and Community for Information Access Research*. Springer, 122–135. https://doi.org/10.1007/978-3-030-36805-0_10

[4] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura, Noriko Kando, Tatsunori Mori, Harumichi Yuasa, Satoshi Sekine, and Kentaro Inui. 2019. Overview of the NTCIR-14 QA Lab-PoliInfo Task. In *Proceedings of the 14th NTCIR Conference*.

[5] Taku Kudo and John Richardson. 2018. Sentencepiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Brussels, Belgium, 66–71. https://doi.org/10.18653/v1/D18-2012

[6] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain, 230–237. https://aclanthology.org/W04-3230

[7] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013

[8] Yasuhiro Ogawa, Yuta Ikari, Takahiro Komamizu, and Katsuhiko Toyama. 2020. NUKL at the NTCIR-15 QA Lab-PoliInfo-2 Task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*.

[9] Yasuhiro Ogawa, Michiaki Satou, Takahiro Komamizu, and Katsuhiko Toyama. 2019. nagoy team's summarization system at the NTCIR-14 QA Lab-Poliinfo. In *NII Conference on Testbeds and Community for Information Access Research*. Springer, 110–121. https://doi.org/10.1007/978-3-030-36805-0_9

[10] Ryoto Ohsugi, Teruya Kawai, Yuki Gato, Tomoyosi Akiba, and Shigeru Masuyama. 2022. AKBL at the NTCIR-16 QA Lab-PoliInfo-3 Task. In *Proceedings of The 16th NTCIR Conference*.

[11] 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:1910.10683

[12] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv:1606.05250