NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001316

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

Koki Horikawa
Hokkaido University
Japan
ecotech13@eis.hokudai.ac.jp

Masaharu Yoshioka
Hokkaido University
Japan
yoshioka@ist.hokudai.ac.jp

ABSTRACT

The HUKB team participated in the Question Answering-2 subtask in the NTCIR-17 QA Lab-PoliInfo-4 task. Our proposed method is divided into three steps. First, we found the sentence on the beginning of the same topic as the input question from the respondent's utterances and extracted the candidate sentences. Next, we found the sentences where the respondent seemed to answer the input question directly, using BERT. Finally, we entered the selected sentences with the input question into the T5 based summarizer, and generated the answer summary. We evaluated the whole method and each process with the dataset distributed by Task Organizers.

KEYWORDS

text classification, summarization, BERT, T5

TEAM NAME

HUKB

SUBTASKS

Question Answering-2 (Japanese)

1 INTRODUCTION

The HUKB team participated in the Question Answering-2 (QA-2) subtask in the NTCIR-17 QA Lab-PoliInfo-4 task [5]. This paper describes our proposed method for the QA-2 subtask and its achievement.

Dealing with the QA-2 subtask, two processes are required. First, we need to find the answer segment corresponding to the input question from the respondent's utterances. Second, we need to summarize it to generate the answer summary as the output of this subtask. Our proposed method is to summarize the sentences selected from the answer segment where the respondent answered the input question directly.

2 RELATED WORK

2.1 QA Alignment (PoliInfo-3)

The QA Alignment subtask was conducted in the NTCIR-16 QA Lab-PoliInfo-3 task [3], and its aim was to associate each question with its answer in the minutes. Since the questions and answers in the minutes aren't associated directly, the QA-2 subtask implicitly includes the QA Alignment subtask.

Tachioka and Keyaki [8] made the paragraph of questions and answers in the minutes using the regular expressions and executed the matching algorithm on the several features of the paragraphs. Ohsugi et al. [6] first segmented the questioner's utterances and the respondent's utterances based on the rule-based approach, then applied the matching algorithm to the vectorized segments of question and answer. Igarashi et al. [2] segmented the statements of the questioner and the respondent, then matched the questions and answers based on their similarity.

2.2 Question Answering (PoliInfo-3)

The same task as the QA-2 subtask was conducted in the PoliInfo-3 task as the Question Answering subtask. Ogawa et al. [4] proposed the system to integrate the two methods based on the length of the summaries. One of the methods was to input the entire respondent's utterances to the summarizer, the other was to input the answer segment which was found using the results of the QA Alignment subtask. Tachioka and Keyaki [8] first calculated the similarity between the input question summary and the questioner's utterances, then found the answer segment using the question segment found in the previous process and made it the summarizer's input. Oshugi et al. [6] segmented the respondent's utterances by topic using the regular expressions, then used the SubTopic and QuestionSummary as the query to find the corresponding answer segment, and gave it the summarizer as the input.

3 METHODS

In the assembly, the respondent typically speaks about the direct answer to the asked question such as the policy or the approach to the issue, after speaking about the topic of the question and the current situation. In addition, observing the gold standard *Togikai-dayori*, answer summaries seem to be generated by selecting the sentences where the answer to the question was directly described and extracting their expressions.

Therefore, we proposed the method with two steps, i.e., first selecting the sentence(s) that directly answer the question, and then summarizing the selected sentence(s). This method is different from the approaches introduced in Section 2.2 by selecting limited numbers of sentence(s) for making final summary. Figure 1 shows the outline of the method. In this method, the selection of the sentence(s) has two sub-steps. One is the selection of sentences in which the respondent talks about the given *SubTopic* (Section 3.1; left half of the figure 1), and the other is the selection of the sentence(s) by checking the appropriateness as an answer to the given question from the selected sentences using BERT [1] (Section 3.2; right bottom of the figure 1). After selecting the sentence(s) that directly answer the question, we use T5 [7] for generating summary (Section 3.3; right top of the figure 1)

3.1 Extracting Answer Segment Including the Corresponding Answer

The difference between this task and the QA Alignment task is that it only identifies answer segments. In addition, since this process is a pre-process to select the appropriate sentence(s), it is preferable to extract the segment with recall oriented. Therefore, we don't

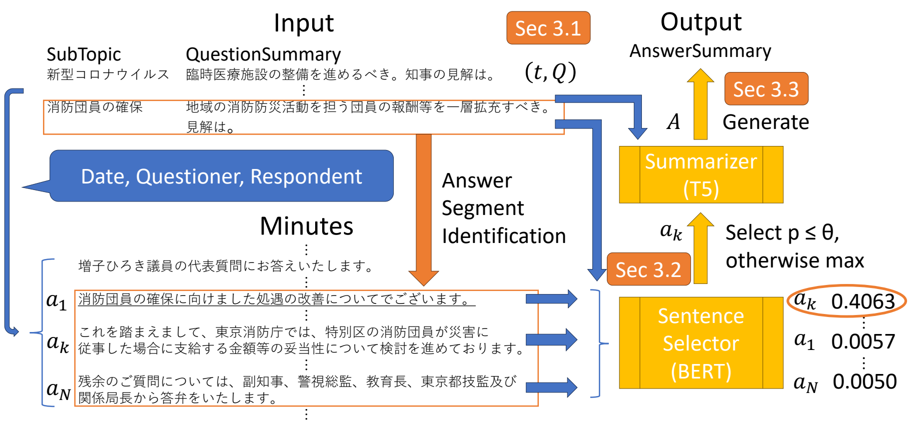
Input
SubTopic
QuestionSummary
新型コロナウイルス
臨時医療施設の整備を進めるべき。知事の見解は。
消防団員の確保
地域の消防防災活動を担う団員の報酬等を一層拡充すべき。見解は。
(t, Q)
Sec 3.1
Output
AnswerSummary
Sec 3.3
A
Generate
Summarizer (T5)
Date, Questioner, Respondent
Answer Segment Identification
Minutes
増子ひろき議員の代表質問にお答えいたします。
a1
消防団員の確保に向けました処遇の改善についてでございます。
ak
これを踏まえまして、東京消防庁では、特別区の消防団員が災害に従事した場合に支給する金額等の妥当性について検討を進めております。
aN
残余のご質問については、副知事、警視総監、教育長、東京都技監及び関係局長から答弁をいたします。
ak
Select p ≤ θ, otherwise max
Sec 3.2
Sentence Selector (BERT)
ak 0.4063
a1 0.0057
aN 0.0050

Figure 1: Outline of our proposed method for QA-2

use cue expressions that are useful for accurately identifying the boundaries of responses.

To identify the boundary, we assume that the respondent answers questions in the order in which the speaker asked them. In addition, in order to improve the recall, we split the text using *SubTopic* only by identifying the first sentence that starts to answer for the *SubTopic*.

First, we select sentences from the minutes in which the respondent answered questions asked by the questioner on that day. We would like to segment these sentences using texts of *SubTopic*. To identify the starting point of the answer segment, we use the similarity between the sentence and the text of *SubTopic*. We assume that the first sentence of the answer segment contains keywords from *SubTopic* and its first *QuestionSummary*.

Since the *target* (*SubTopic* and its first *QuestionSummary*) is shorter than the original sentence (*sent*), we use the following similarity measure to find the correspondent original sentence[1].

$$\mathrm{terms}(text) : \text{terms set that exists in the } text$$
$$\mathrm{freq}(text, w) : \text{term frequency of } w \text{ in the } text$$
$$\mathrm{sim}(target, sent) = \frac{\sum_{w \in \mathrm{terms}(sent)} \mathrm{freq}(target, w)}{\sum_{w \in \mathrm{terms}(target)} \mathrm{freq}(target, w)}$$

We select the first sentence of the *SubTopic* by using sim(*SubTopic*, *sent*) (r_t) and sim(*SubTopic* and its first *QuestionSummary*, *sent*) ($r_{t,Q}$) (Figure 2). The candidate sentences for the *SubTopic* is constructed as follows.

(1) Selection of candidates using r_t or $r_{t,Q}$
r_t and $r_{t,Q}$ are calculated for each target sentence and sentence(s) with highest score of r_t is(are) selected as candidates. When there is no sentence that contains all *SubTopic* keywords ($r_t \neq 1$), there are cases that the sentence is not the appropriate one for the *SubTopic*. In such cases, we compare the highest score of $r_{t,Q}$ with the highest score of r_t and use sentences with the highest score of $r_{t,Q}$ when the value is larger than the highest score of r_t.

(2) Selection of the first sentence
Since we assume that the respondent answers questions in the order in which the speaker asked them, the first sentence of the i-th *SubTopic* should be selected after the first sentence of the $(i-1)$-th *SubTopic* $(i > 1)$. We select the earliest sentence from the candidates that satisfy this order condition as the first sentence. The system uses a sentence before the selected sentence as the last sentence of the $(i-1)$-th *SubTopic*. However, if there is no sentence that satisfies this order condition, we treat the case as an order violation one. In this case, we select the earliest sentence from the candidates as the first sentence of the i-th *SubTopic*. The system uses a sentence before the first sentence of the $(i-1)$-th *SubTopic* as the last sentence of the i-th *SubTopic*, because there are no corresponding sentences for the i-th *SubTopic* after the $(i-1)$-th *SubTopic*. In this order violation case, we allow the *SubTopic* boundary to overlap (i.e., we don't change the boundary for the overlapped previous *SubTopic*). The last sentence of the $(i-1)$-th *SubTopic* is selected with the next *SubTopic*.

3.2 Predicting Sentences Answering Input Question Directly

We assume that the answer summary in the gold standard is generated by extracting expressions from the sentences where the respondent answers the input question directly. For example, as shown in Figure 3, there is a sentence corresponding to the answer summary. We consider that selecting such sentences by checking the appropriateness of the answer for the questions would make it possible to generate the answer summary efficiently. We considered the sentence selection task as the sentence prediction task with BERT [1], and built the model to solve this task.

Make a pseudo corpus. To train the model, we need to make a pseudo corpus from gold standard of answer summaries. First, we extracted the part of respondent's utterances which includes the corresponding answer using the procedure of 3.1. Then, for each sentence in the answer summary, we found the sentences in the

[1]We parsed texts using MeCab (https://taku910.github.io/mecab/). Also, we use verb, adjective, noun, and prefix as the parts of speech of terms.

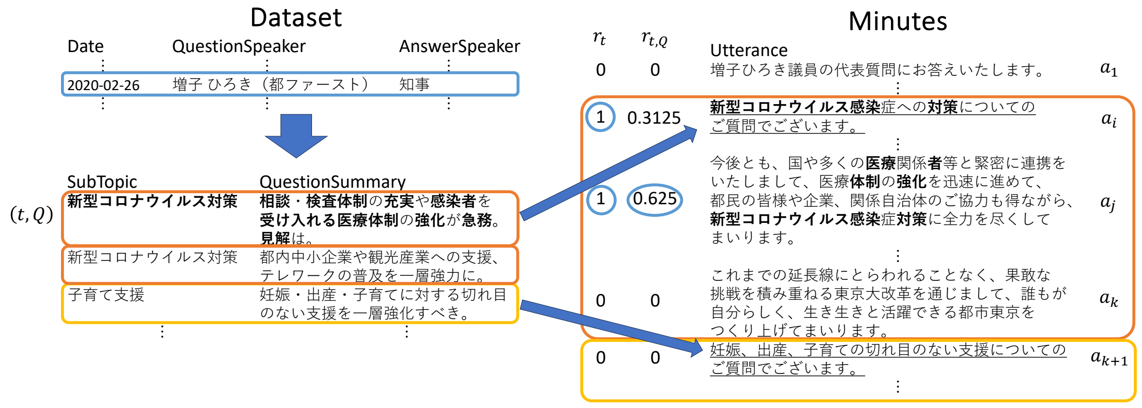
Dataset
Date
QuestionSpeaker
AnswerSpeaker
2020-02-26
増子 ひろき（都ファースト）
知事
SubTopic
QuestionSummary
(t, Q)
新型コロナウイルス対策
相談・検査体制の充実や感染者を受け入れる医療体制の強化が急務。見解は。
新型コロナウイルス対策
都内中小企業や観光産業への支援、テレワークの普及を一層強力に。
子育て支援
妊娠・出産・子育てに対する切れ目のない支援を一層強化すべき。
Minutes
r_t
$r_{t,Q}$
Utterance
0 0
増子ひろき議員の代表質問にお答えいたします。
a_1
1 0.3125
新型コロナウイルス感染症への対策についてのご質問でございます。
a_i
1 0.625
今後とも、国や多くの医療関係者等と緊密に連携をいたしまして、医療体制の強化を迅速に進めて、都民の皆様や企業、関係自治体のご協力も得ながら、新型コロナウイルス感染症対策に全力を尽くしてまいります。
a_j
0 0
これまでの延長線にとらわれることなく、果敢な挑戦を積み重ねる東京大改革を通じまして、誰もが自分らしく、生き生きと活躍できる都市東京をつくり上げてまいります。
a_k
0 0
妊娠、出産、子育ての切れ目のない支援についてのご質問でございます。
a_{k+1}

Figure 2: Outline of the process in Section 3.1

answer segment which contained the greatest number of content words of that summary sentence.[2] We considered all of them as positive examples in the pseudo corpus and others as negative examples.

Train the model. We built the model using the pseudo corpus made in the previous process. As the architecture of model, we adopted the neural model BERT, and fine-tuned the model of its architecture pre-trained on Japanese texts. When fine-turning it, we trained the model to get two inputs of the text with a space between *SubTopic* and *QuestionSummary*, and the sentence in the pseudo corpus, and output the probability that the input sentence is the positive sample.

Predict sentences with the model. For each question, we calculate confidence (appropriateness as an answer to the question) for the sentence. If the confidence is above a certain level θ, we select all sentences as candidates. If no sentence selected by θ, we use the sentence with the highest confidence as a candidate for *QuestionSummary*.

3.3 Generating the Answer Summary

We generated the answer summary using the sentences predicted by the model built in the previous section. To generate a human-readable summary, we adopted the neural language model to summarize, and T5 [7] as the model architecture. Furthermore, we used the Japanese pre-trained model of T5 to fine-tune it to generate the answer summary from answer sentences. We fine-tuned the pre-trained model to input the sentences in the positive examples of the pseudo corpus and output the answer summary in the gold standard.

4 EXPERIMENTS

4.1 Generating Datasets Using Pseudo Corpus

We made the pseudo corpus according to the procedure described in the previous chapter and generate two datasets from it to use when fine-turning the pre-trained models of BERT and T5. Tables 1 and 2 show the statistics of the two datasets for BERT and T5. In addition, each dataset was split by the ratio of 8-1-1 to make train data, validation data, and test data.

Table 1: Statistics of Dataset for BERT

Total Number of Candidate Sentences to be Selected	84,579
Total Number of Sentences in Pseudo Corpus	10,997

Table 2: Statistics of Dataset for T5

Sentence	#AnswerSummary	#SummarySource
1	5,757	5,686
2	2,176	2,026
3	106	239
4	7	45
≥5	0	50
Total	8,046	8,046

As shown in Table 2, note that the total number of sentences in the answer summaries is not exactly equal to that of positive examples in the pseudo corpus, because in section 3.2 we allowed multiple sentences to correspond to a single sentence in the answer summary.

4.2 Building Two Neural Model and Setup

We utilized cl-tohoku/bert-base-japanese-v2[3] as the pretrained BERT model for Japanese. In the same way, we use sonoisa/t5-base-japanese[4] as the T5 model pre-trained on Japanese corpus. Tables 3 and 4 show the hyperparameters to use when fine-turning the pre-trained BERT model and T5 model respectively. Each model is adopted which achieved the minimum validation loss in 10 epochs. Using the model to predict sentences to generate the answer summary, thresholds were used at 0.5, 0.8, 0.9, 1.0 for the predicted probability. If the maximum predicted probability was below the threshold, the sentence predicted with it was used as the source to generate the answer summary.

[2]In Figure 3, a_i with $r_{A_1}(a_i) = 0.9286$ as the ratio of the content words and a_j with $r_{A_2}(a_j) = 1$ are found.

[3]https://huggingface.co/cl-tohoku/bert-base-japanese-v2

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

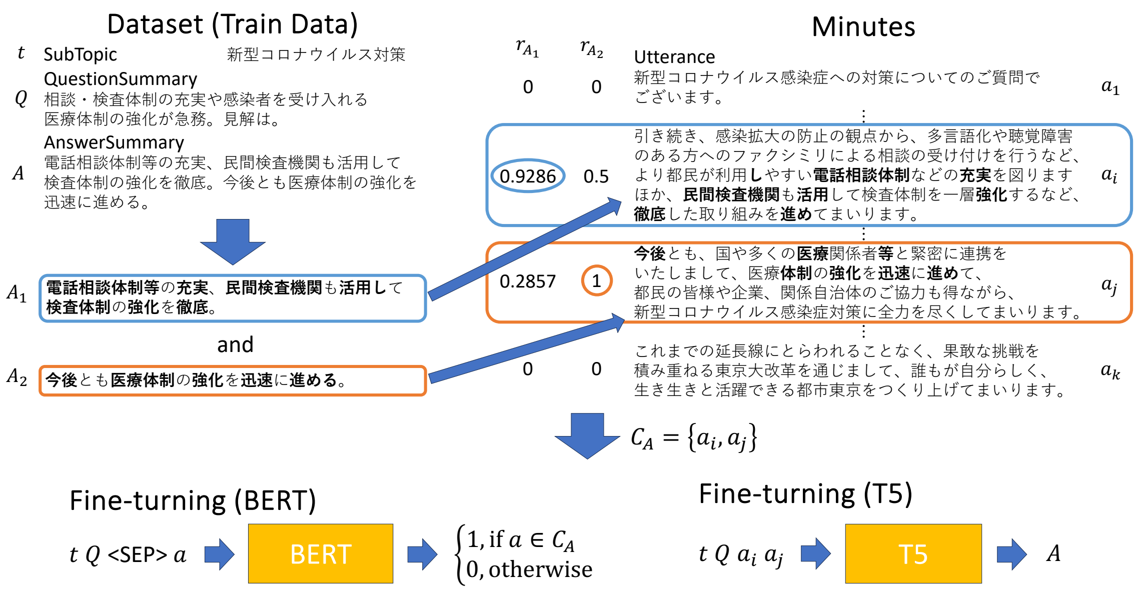
Dataset (Train Data)
t SubTopic 新型コロナウイルス対策
Q QuestionSummary 相談・検査体制の充実や感染者を受け入れる医療体制の強化が急務。見解は。
A AnswerSummary 電話相談体制等の充実、民間検査機関も活用して検査体制の強化を徹底。今後とも医療体制の強化を迅速に進める。
A1 電話相談体制等の充実、民間検査機関も活用して検査体制の強化を徹底。
and
A2 今後とも医療体制の強化を迅速に進める。
Minutes
r_A1 r_A2 Utterance
0 0 新型コロナウイルス感染症への対策についてのご質問でございます。 a1
0.9286 0.5 引き続き、感染拡大の防止の観点から、多言語化や聴覚障害のある方へのファクシミリによる相談の受け付けを行うなど、より都民が利用しやすい電話相談体制などの充実を図りますほか、民間検査機関も活用して検査体制を一層強化するなど、徹底した取り組みを進めてまいります。 ai
0.2857 1 今後とも、国や多くの医療関係者等と緊密に連携をいたしまして、医療体制の強化を迅速に進めて、都民の皆様や企業、関係自治体のご協力も得ながら、新型コロナウイルス感染症対策に全力を尽くしてまいります。 aj
0 0 これまでの延長線にとらわれることなく、果敢な挑戦を積み重ねる東京大改革を通じまして、誰もが自分らしく、生き生きと活躍できる都市東京をつくり上げてまいります。 ak
C_A = {ai, aj}
Fine-turning (BERT)
t Q <SEP> a
BERT
1, if a ∈ C_A
0, otherwise
Fine-turning (T5)
t Q ai aj
T5
A

Figure 3: Outline of fine-turning in Section 3.2 and 3.3

Table 3: Hyperparameters to fine-tune BERT

Epochs	10
Batch size	16
Learning rate	2e-5
Weight decay	0.01

Table 4: Hyperparameters to fine-tune T5

Epochs	10
Batch size	16
Learning rate	5e-5
Weight decay	0

4.3 Results

The performance of proposed method and each process was evaluated using test data in the formal run and correct data manually made from train data. Table 5 shows results of the evaluation on test data in the formal run.

5 DISCUSSIONS

5.1 Evaluation of Each Process

To check the performance of each of the modules, we evaluated each of them. Since the correct answers for the test data had not yet been published at the time of writing, we used the test data made in Section 4.1. In addition, the authors manually checked the test data and modified them.

Using them, the results of the evaluation of each module are shown in Tables 6 and 8. For Table 6, Recall (partial) is the ratio of examples that there is at least one correct sentence there. For Table 7, Exact Match is the ratio of examples that the predicted sentences match all of the correct sentences, and F1 is the mean of the F-value in each example. Also, the scores marked Preprocess Correct are calculated only on examples that the result of the preprocessing contained all correct sentences.

Table 5: Evaluation in Formal Run

Method (θ: Threshold)	ROUGE-1 F-Measure
HUKB ($\theta = 0.8$)	0.3011
HUKB ($\theta = 0.9$)	0.3008
HUKB ($\theta = 1.0$)	0.3008
HUKB ($\theta = 0.5$)	0.2949
ditlab	0.3246
omuokdlb	0.3130
TO	0.2736
IKM23	0.2724
AKBL	0.1162

Table 6: Evaluation of Answer Utterances Segmentation

Recall (partial)	0.9579

Table 7: Evaluation of Sentence Prediction

Method (θ: Threshold)	$\theta = 0.5$	$\theta = 0.8$	$\theta = 0.9$	$\theta = 1.0$
Exact Match (EM)	0.3713	0.3923	0.3948	0.3948
F1	0.5856	0.5825	0.5825	0.5825
EM (Preprocess Correct)	0.3968	0.4193	0.4220	0.4220
F1 (Preprocess Correct)	0.6125	0.6097	0.6097	0.6097

Table 8: Evaluation of Summarization

Input	ROUGE-1 F-Measure
Predicted ($\theta = 0.5$)	0.3011
Predicted ($\theta = 0.8$)	0.3027
Predicted ($\theta = 0.9$)	0.3030
Predicted ($\theta = 1.0$)	0.3030
Manual data	0.4476

Table 9: Evaluation of Pseudo Corpus

Exact Match	0.7265
F1	0.8761

5.2 Pseudo Corpus: Quality

We evaluated the pseudo corpus with the gold data made in the previous section. Table 9 shows the result of that. The meanings of evaluation scores are same as Table 7. As can be seen in Table 9, there is room for improvement in the quality of the pseudo corpus.

5.3 Utterance Segmentation: Performance

As shown in Table 6, the correct sentences are included in the candidates in over 90%. To further limit the range of the answer extraction, it seems effective to use the cue expressions with reference to the proposed methods in the PoliInfo-3.

5.4 Sentence Selection: Effectiveness and Possibility

Table 8 shows that using manually selected sentences to generate the answer summary greatly improves the ROUGE score. Therefore, the effectiveness of using the selected answer sentences to summarize is expected. However, there are problems with the method to predict such sentences.

To solve the problems, using the position of sentences in the answer segment to predict the sentences may be useful. This is because it seems that the original sentences to be summarized such as the measures and policies for the future occur in the end of texts of the answer segment.

5.5 Summarization: Performance

The performance of summarizer is 0.4476 as the ROUGE score in Table 8, when the summarizer is given the gold standard of summarized sentences as input. Note that refining the pseudo corpus is expected to improve the performance of summarizer, because the performance depends on the quality of pseudo corpus.

6 CONCLUSIONS

We participated in the QA-2 subtask in the PoliInfo-4 task and proposed the method to predict sentences to be summarized from the respondent's utterances and generate the answer summary using them. There are problems especially at the process to select sentences, and it seems effective to refine the pseudo corpus and use the positional information of sentences.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[2] Naoki Igarashi, Daiki Iwayama, Hideyuki Shibuki, and Tatsunori Mori. 2022. Forst: A Challenge to 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, Madoka Ishioroshi, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Ken ichi Yokote, Kazuma Kadowaki, Tatsunori Mori, Kenji Araki, Teruko Mitamura, and Satoshi Sekine. 2022. Overview of the NTCIR-16 QA Lab-PoliInfo-3 Task. *Proceedings of The 16th NTCIR Conference* (6 2022).

[4] Yasuhiro Ogawa, Yugo Kato, and Katsuhiko Toyama. 2022. nukl's QA System at the NTCIR-16 QA Lab-PoliInfo-3. In *Proceedings of The 16th NTCIR Conference*.

[5] 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. *Proceedings of The 17th NTCIR Conference* (12 2023). https://doi.org/10.20736/0002001326

[6] 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*.

[7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]

[8] Yuuki Tachioka and Atsushi Keyaki. 2022. ditlab at the NTCIR-16 QA Lab-PoliInfo-3. In *Proceedings of The 16th NTCIR Conference*.