Attempt to Develop An Approach Based on BERT for Task of NTCIR-16 QA Lab-Poliinfo-3 Budget Argument Mining

Akio Kobayashi*
akio.kobayashi@naro.go.jp
National Agriculture and Food Research Organization
Tsukuba, Ibaraki, Japan

Hiroki Sakaji*
sakaji@sys.t.u-tokyo.ac.jp
The University of Tokyo
Bunkyo-Ku, Tokyo, Japan

ABSTRACT

The SMLAB team participated in the budget argument mining subtask of the NTCIR-16 QALab Poli-info Task. This paper reports our approach to solving this task and discusses the official results. As an reult, our model underperformed other approaches from other teams drastically.

KEYWORDS

Information Retrieval, BERT, Local Assembly Analisys, Multi-Task Learning

TEAM NAME

SMLAB

SUBTASKS

Budget Argument mining

1 INTRODUCTION

The SMLAB team participated in the budget argument mining subtask of the NTCIR-16 QALab Poli-info Task[2]. This paper reports our approach to solving the problem and discusses the official results.

To tackle with the budget argument mining task, participants need to link money expressions in each untterance in the assembly held at the government to the budget item of that government. Moreover, participants need to classify a money expression into 7 “Argument Class”.

Since these two subtasks, the budget argument mining task seemed to be a challenging one, thus we attempt to employ the SOTA transformer-based model. As a result, our approach did not get good results as expected due to lack of pre-processing of the data.

2 RELATED WORK

As related works, Tamamaru et al. constructed a corpus of Japanese local assembly minutes[5]. Sakaji et al. proposed a method for extracting volitional utterances from Japanese Local Political Corpus[4]. Ototake et al. develop a Web-based system for visualizing local politics in Japan [3]. Kimura et al. propose an approach to uniquely identify the speakers by hand[6].

On the other hand, we attempt to identify “relatedID” and “argumentClass” that are part of budget argument mining tasks.

3 OUR APPROACH

In this section, we introduce our budget argument mining approach. We employ a pre-trained model for the budget argument mining task because pre-trained models such as transformer-based models achieved SOTA in several tasks recently. In NTCIR-16 poliinfo budget argument mining task consists of two subtasks, identifying “relatedID” and “argumentClass” of money exprssions. The task of identifying the “relatedID” is to identify which budget the input money expression is related to. For example, “約2億8700万円(about 287 million yen)” is a money expression that appears in the utterance of assembly of Otaru City. The second task of idetification of “argumentClass” is to classify input money expression into 7 classes, “Premise : 過去・決定事項,” “Premise : 未来（現在以降）・見積,” “Premise : その他（例示・訂正事項など）,” “Claim : 意見・提案・質問,” “Claim : その他,” “金額表現ではない,” and “その他.” For example, the money expression “約2億8700万円(about 287 million yen)” is classified into the class:“Premise : 未来（現在以降）・見積.”

In the minutes data for this task has structured for each utterance in each meeting. Also the utterance structured by its property such as uttered person or the money expressions mentioned above. In the supervised data, the money expression is linked to its budget data and classified into “argumentClass.”

To address these two subtasks, we assumed that the documents of utterance at the assembly and the budget summary of the same government were similar by comparing some examples extracted from the supervised data.

Therefore, we attempt to identify “relatedID” and “argumentClass” using bidirectional encoder representations from transformers (BERT)[1] and multi-task learning. In this research, we use BERT based on Wikipedia[1]. By we conduct fine-tuning of BERT, we identify “relatedID” and “argumentClass” of input money expressions.

We create a model for identification of “relatedID” and “argumentClass” using BERT. We illustrate our model in Figure 1. Our model has two inputs, input three sentences, and the description of the target budget. In this task, we have to search budget ID related to input money expressions. Therefore, we decided to find budget ID using cosine similarity based on vectors of BERT.

From Figure 1, input three sentences consist of the sentence including the money expression and before and after the sentence including the money expression. Our model is multi-task learning that consists of identification of “relatedID” and “argumentClass.” Therfore, our model has two loss functions. One of the loss functions is cosine similarity loss related to the identification of “relatedID.”

*Both authors contributed equally to this research.

[1]https://github.com/cl-tohoku/bert-japanese

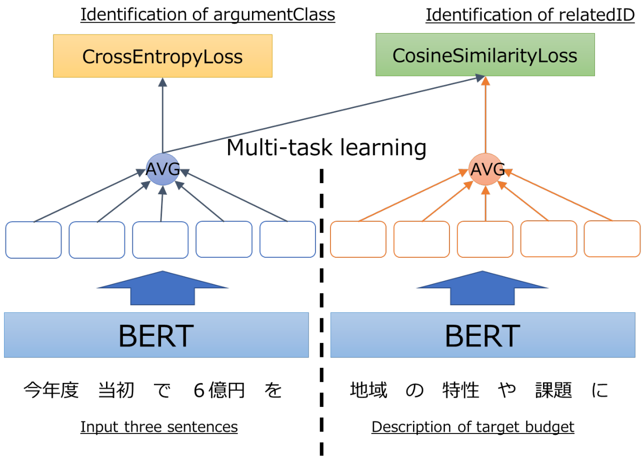

Figure 1: Our training architecture.

Table 1: Best results for Each Participant

Rank	Team Name	Score	Argument Class	Related ID
1	JRIRD	0.5106	0.5827	0.6170
2	OUC	0.4468	0.5712	0.6596
3	fuys	0.2340	0.5692	0.3404
4	rVRAIN	0.1702	0.5654	0.6170
5	takelab	0.0426	0.3942	0.0638
6	**SMLAB**	**0**	**0.3827**	**0**

Another one is cross-entropy loss related to the identification of "argument Class."

We describe the flow of your model training. First, our model obtains vectors of BERT from inputs. Then, we obtain two types of the average of vectors, vectors based on the input three sentences and the description of the target budget. We calculates cosine similarity loss using obtained vectors and labeled data. Additionally, using labeled data and obtained average vectors based on input three sentences, we calculate the cross-entropy loss.

4 EXPERIMENTS

Table 1 is a official score of the formal run. In the table 1, the "Argument Class" and "Related ID" are a accuracy of each subtask, and "Score" is overrall score that is calculated from these two scores.

4.1 Results & Discussion

As a result, our method was a worst because we could not find any budget (Related ID) for all money expressions. Thus our assumption that sentences appearing near money expressions will contain a similar expressions to the budget summary was declined. This means we need any kind of knowledge (e.g. a dictionary for paraphrase or any kind of knowledge base for government) to handle the utterance at assembly and the budget document at the same time.

5 CONCLUSIONS

We attempt to identify "relatedID" and "argumentClass" using BERT. Our model consists of multi-task learning, cosine similarity loss, and cross-entropy loss. As an experiment reult, our model underperformed other approaches from other teams drastically. We consider that the result is caused by our incorrect assumption.

As future work, we attempt to analyze the labeled data in detail because our assumption was incorrect. Additionally, we try to create knowledge graphs from local assembly data. Moreover, we attempt to develop a new model using created knowledge graphs for identification of "relatedID" and "argumentClass."

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.
[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. *Proceedings of The 16th NTCIR Conference* (6 2022).
[3] Hokuto Ototake, Hiroki Sakaji, Keiichi Takamaru, Akio Kobayashi, Yuzu Uchida, and Yasutomo Kimura. 2018. Web-based system for Japanese local political documents. *International Journal of Web Information Systems* 14, 3 (2018), 357–371.
[4] Hiroki Sakaji, Yasutomo Kimura, Kiyoshi Izumi, and Hiroyasu Matsushima. 2019. Extraction of Volitional Utterances from Japanese Local Political Corpus. In *2019 International Conference on Data Mining Workshops (ICDMW)*. 24–29.
[5] Keiichi Takamaru, Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura, and Noriko Kando. 2020. Extraction of the Argument Structure of Tokyo Metropolitan Assembly Minutes: Segmentation of Question-and-Answer Sets. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, 2064–2068.
[6] Yuzu Uchida Yasutomo Kimura and Keiichi Takamaru. 2018. Speaker Identification for Japanese Prefectural Assembly Minutes. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).