AKBL at NTCIR-14 QA Lab-PoliInfo Task

Kazuki Terazawa[1], Daiki Shirato[1], Tomoyoshi Akiba[1], and Shigeru Masuyama[1]†

[1] Toyohashi University of Technology, Japan

Abstract. The local council's proceedings are useful as research materials and as a basis for selecting one among candidates during elections, but they are large in volume. For this reason, it would be helpful for many people to be able to identify the source from a summary published on the Web and automatically summarize the utterance.Therefore, in this paper, using the dataset of the Tokyo Metropolitan Conference proceedings provided by NTCIR-14 QA Lab-PoliInfo, we worked on the automatic specification of the original utterance extent (Segmentation Task) and the summary of the utterance (Summarization Task) .In the Segmentation Task, the extraction extent is specified by the number of lines in the conference proceedings, and the result is evaluated by its Precision, Recall, and F-measure. In the Summarization Task, the ROUGE and human evaluation are performed. In addition, it is difficult for individuals to conduct fact-checking to cope with fake news that is disseminated to a large amount of information on the Internet. Therefore, we use the dataset of the Tokyo Metropolitan Congress Proceedings, which is also provided by NTCIR-14 QA Lab-PoliInfo, and work on estimation of fact-checkability against utterances and classification of utterances when it is possible to check facts (Classification Task). In the Classification Task, labels support, against and other are assigned to an utterance and the results are evaluated by its Precision, Recall, F-measure. We also made a decision on the relevance of the target policy.

Keywords: NTCIR14 ・QALab・PoliInfo・segmentation・summarization・classification

Team Name. akbl

Subtasks. Segmentation task (Japanese), Summarization task (Japanese), Classification Task (Japanese)[1]

1. Introduction

In the local council, various discussions are being conducted to decide policies for the future, and conference proceedings in which all utterances in the council are recorded, are published on the web.The proceedings are useful as materials for

† Currently, with Tokyo University of Science

studying local politics and economics, and also serve as reference for election candidates. However, since the total number of utterances in 47 prefectures is more than 2 million in a year, we may not have sufficient time to read every single utterance. Although some municipalities have published a summary of their utterances on the Web, it is considered that summarizing such utterances manually is quite costly because one utterance is long.In addition, since the amount of information is insufficient only by the summarized utterance, it may be necessary to identify the original utterance from the summary.So, we work on the specification of the original utterance extent (Segmentation Task) and the summary of the utterance (Summarization Task) using the Tokyo Metropolitan Conference proceedings as an example.

Today, various types of information are spread across borders on the Internet, which naturally includes fake news, but it is difficult for individuals to verify facts on all the information they receive. Therefore, we will work on the estimation of fact verifiability with respect to the utterances of the Tokyo Metropolitan Conference proceedings, the classification of the utterances in case of fact verification (support or against), and determination of the relevance to the target policy (Classification Task).

In working on this study, we will use the data provided by NTCIR-14 QA Lab-PoliInfo. Details of the data are described in [1].

2. Task Description

In this study, we worked on three tasks: Segmentation Task, Summarization Task, and Classification Task.

2.1 Segmentation Task

This task uses the summarized utterances to identify the extent of the original utterances in the conference proceedings.

Use the following procedure to specify the extent of the original utterance of its summary.

(1)Divide the data of conference proceedings into each utterance.
(2)Use date information to exclude data other than that day.
(3)Extraction extent candidates are created based on sentence head expressions and sentence end expressions (if no such expression is found, one utterance is taken as an extraction extent candidate).
(4)Calculate the TF-IDF value of the target word (Nouns of Summary and Subtopic) for each extraction extent candidate, and further multiply the number of the target words present in the extraction extent candidate, and output the extent in which the obtained value is the largest.

・Calculation method of TF-IDF value.

TF = The number of occurrences of target nouns / The number of nouns in target utterance

IDF = $\log_{10}$ (The number of utterances / The number of utterances in which the target noun appeared)
TF-IDF = TF*IDF

・Sentence head expressions

(i) Sentence head expressions of question

・・・[次に、(tsugi ni,: next), 伺います(ukagaimasu: will ask you), お伺いいたします(oukagai itashimasu: will ask you), 質問をさせていただきます(shitsumon wo sasete itadakimasu: will ask a question), 質問いたします(shitsumon itashimasu: will ask a question),質問をいたします(shitsumon wo itashimasu: will ask a question), 質問します(shitsumon shimasu: will ask a question),続いての質問は(tsuduiteno shitsumon wa: the next question is)]

(ii)Sentence head expressions of answer

・・・[お答えいたします(okotae itashimasu: will answer), お答えをいたします(okotae wo itashimasu: will answer), 次いで、(tsuide,: next), 次に、(tsugi ni,: next), まず、(mazu,: first), 他方で、(tahou de,: on the other hand), 最後に(saigo ni: finally)]

・Sentence end expressions

(i)Sentence end expressions of question

・・・[伺います(ukagaimasu: will ask you), お伺いいたします(oukagai itashimasu: will ask you), 伺いたい(ukagaitai: want to ask), 質問を終わります(shitsumon wo owarimasu: finish the question), 見解を求め(kenkai wo motomemasu: ask for a view), 質問といたします(shitsumon to itashimasu: will ask a question)]

(ii)Sentence end expressions of answer

・・・[おります(orimasu: I am), まいります(mairimasu: I will)]

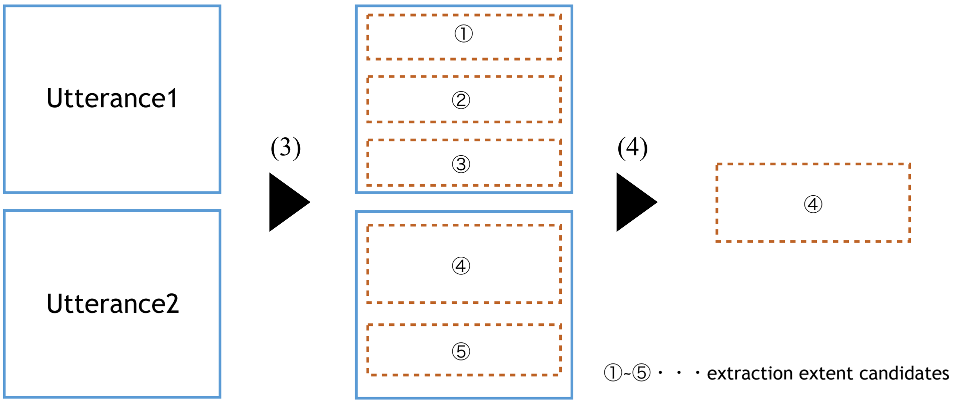

Figure1. Segmentation Task method

2.2 Summarization Task

In this task, we summarize utterances within a fixed number of characters.

Use the following procedure to summarize utterance.

(1)Treat a sentence as a question sentence if there is a line containing clue expressions (i), otherwise handle it as an answer sentence (Clue expressions and cutting expressions to be illustrated later used in question sentences are (i), and those used in answer sentences are (ii)).

(2)Extract lines containing clue expressions or subtopic nouns, and divide the lines up to the lines containing each clue expression as one topic.

(3)If the number of characters exceeds the predetermined limit, calculate the TF-IDF value for each line of each topic, and do the following work until it is within the number of characters.

(3-1)Delete after the cutting expression of the line with the cutting expression

(3-2)Remove clauses that contain adverbs, adjectives and influential.

(3-3)Delete the lines with the lowest TF-IDF value in order from the topic above (however, leave at least one line in each topic).

If the limit number of characters even after the above processing, delete the top topic and perform the above work on the remaining topics.

・Clue expressions

(i)Clue expressions of a question

・・・[伺います(ukagaimasu: will ask you), お伺いいたします(oukagai itashimasu: will ask you, お伺いします(oukagaishimasu: will ask you), 見解を(kenkai wo: view), 所見を(syoken wo: findings), 質問いたします(shitsumon itashimasu: will ask a question), 質問を終わります(shitsumon wo owarimasu: finish the question), お答えください(okotae kudasai: please answer), 答弁を求めます(touben wo motomemasu: please answer), 提案します(teian shimasu: propose), いかがですか(ikaga desuka: how is it)]

(ii)Clue expressions of an answer

・・・[まいります(mairimasu: I will), 思っております(omotte orimasu: I think)]

・Cutting expressions

(i)Cutting expressions of a question

・・・[が必要(ga hitsuyou: ~ is necessary), べき(beki: should)]

(ii)Cutting expressions of an answer

・・・[いきたい(ikitai: want to), べき(beki: should)]

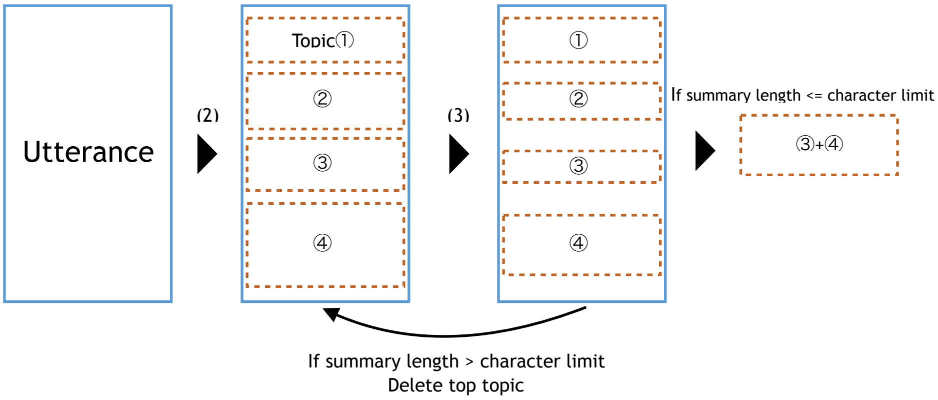
Utterance
(2)
Topic①
②
③
④
(3)
①
②
③
④
If summary length <= character limit
③+④
If summary length > character limit
Delete top topic

Figure2. Summarization Task method

2.3 Classification Task

In this task, we categorize an utterance (one sentence) whether it is 'support' or ‘against’ with the policy. Also, utterances that have no basis for discrimination are classified as other.

Use the following procedure to classify utterance.

(1)Based on the matching rate of the labeling results among the annotators, an annotator graph is created by using the annotator as a node and the matching rate of the labeling between the annotators as the weight of the edge connecting the annotators, and perform graph-based clustering. For clustering, we use the Newman algorithm that performs clustering from the bottom in order to define Modularity (Eq.1), which is an index indicating the goodness of the clustering result of the graph, and aiming to maximize it.

$$Q = \frac{1}{2M} \sum_{vw} (A_{vw} - \frac{k_v k_w}{2M}) \delta(C_v C_w) \quad ・・・\text{Eq.1}$$

M is the total number of edges, A_{vw} is the vw component of the adjacency matrix of the graph, and k_v and k_w are the orders of the nodes v and w. Also, $\delta\ (C_v,\ C_w)$ is 1 when clusters C_v and C_w are the same cluster, and 0 otherwise. The Newman algorithm is performed by the following procedure.

(1-1)Assign all nodes to different clusters.

(1-2)Calculate Modularity when merging two clusters for all combinations of clusters.

(1-3)Merge clusters with the highest modularity combination.

(1-4)Repeat steps 1-2 and 1-3 until the no clusters can be merged.

(1-5)The merge result with the highest Modularity is the clustering result.

(2)Of the annotator clusters obtained by clustering, a cluster having the highest labeled matching rate among the clusters is set as cluster X. The data labeled by annotators in the cluster X are used for training and modeling with a two-layer classifier consisting of LSTM layer and output layer.The procedure is as follows.

(2-1)Create a word dictionary of utterances labeled by the cluster X annotator.

(2-2)The utterance is divided into words, converted into a one-hot vector based on the word dictionary, and input to the LSTM through the embedded layer.At this time, the LSTM gets 200 hidden states for each input word.

(2-3)Make an estimate of fact-checkability with the total bonding layer and the softmax layer. The training label of the classifier uses a probabilistic label obtained from the percentage of annotators determined to be fact-checkable, and Kullback-Leibler Divergence[2] is used as a loss function, which represents the difference between the probability labels and the probability distribution of the prediction results.Also, initialize hidden state and cell of LSTM with 0, and use Adam for parameter optimization.

3 Results

The results of this study are shown in this section. Detailed evaluation methods for each task are described in [1].

3.1 Segmentation Task

In this task, evaluate the result by Precision, Recall, F-measure. The results are shown in Table 1.

Table 1 Results of Segmentation Task

Recall	0.768
Precision	0.538
F-measure	0.633

Looking at the results, it is considered that many cases have been extracted to unnecessary extents because Recall is higher than Precision.

3.2 Summarization Task

In this task, results were evaluated by human and ROUGE. The results of human evaluation are shown in Table 2 and the results of ROUGE evaluation are shown in Table 3.

Table 2 Results of Summarization Task(human evaluation)

	all-topic	single-topic	multi-topic
content(X=0)	0.722	0.708	0.739
content(X=2)	1.005	1.009	1.000
formed	1.833	1.844	1.821
total	0.826	0.849	0.799

Table 3 Results of Summarization Task(ROUGE evaluation)

	Recall							F-measure						
	N1	N2	N3	N4	L	SU4	W1.2	N1	N2	N3	N4	L	SU4	W1.2
Surface Form	0.400	0.173	0.113	0.076	0.345	0.189	0.157	0.361	0.156	0.102	0.068	0.310	0.167	0.185
Stem	0.415	0.184	0.122	0.083	0.357	0.203	0.164	0.375	0.165	0.110	0.074	0.322	0.179	0.195
Content Word	0.256	0.113	0.065	0.034	0.247	0.124	0.148	0.224	0.098	0.056	0.031	0.216	0.100	0.158

Tables 2 and 3 show that although the human evaluation 'formed' is high to some extent but the 'content' is overall low, and the ROUGE results are not too high. Therefore, although many of the summaries are grammatically correct, the content is not appropriate.

3.3 Classification Task

The clustering results are shown in Fig. 3 and the labeled matching rates of each cluster (except for the cluster with only annotator T) are shown in Table 4.

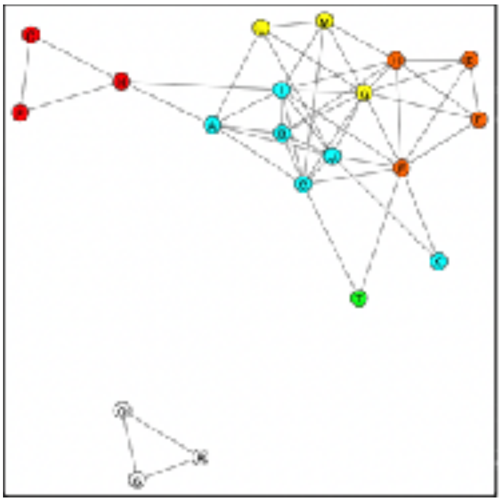

Fig.3 Results of clustering

Table 4 Average labeled matching rates

Annotaters	Labeled matching rate
L,M,G	0.921
H,D,E,F	0.782
A,B,C,I,J,K	0.573
Q,S,R	0.461
O,P,N	0.445

Since Table 4 shows that clusters of annotators L, M, and G have the highest labeled matching rate, 4611 utterances labeled by the annotators of this cluster are used for training of the classifier. Based on the convergence condition of the loss at the time of learning, the learning rate is set to 0.0002, and classification results using a classifier obtained by repeating 100 epoch learning are shown in Table 5. Evaluate the accuracy rate of relevance and for each of the support, against, and other labels using Recall, Precision, and F-measure.

Table 5 Results of Classification Task

	Support	Against	Other	Relevance
Recall	0.118	0.034	0.983	0.923
Precision	0.344	0.097	0.939	
F-measure	0.176	0.050	0.960	

It can be seen from Table 5 that the results for the support and against labels are low and the results for the other labels are high.It is considered that the reason is that most data have other labels. Therefore, in order to create a more accurate classifier, it may be necessary to prepare data with less deviation in the number of data of each label.

4. Conclusion

In this study, we tried to extract the extent of the utterance source based on the summary of the utterance in the Tokyo Metropolitan Conference proceedings(Segmentation Task), to summarize the utterance(Summarization Task), to estimate fact-checkability for a topic of an utterance and to classify it (Classification Task). Since the Segmentation Task results in low Precision, and there are many cases where the extraction extent is too wide, it may be necessary to make further improvements to narrow the extent. Regarding the Summarization Task, although human evaluations show that the summary is grammatically correct to some extent, the result is not accompanied by much content. It is thought that it is so even if it sees a point with a low ROUGE score.For this reason, we think that improvement should be made focusing on capturing important information that is essential to the summary.The Classification Task is highly biased in data, making it difficult to create an accurate classifier. Therefore, it is considered necessary to prepare data with less bias.

References

[1]Kimura, Y., Shibuki, H., Ototake, H., Uchida, Y., Takamaru, K.,Sakamoto, K., Ishioroshi, M., Mitamura, T., Kando, N., Mori, T.,Yuasa, H., Sekine, S., Inui, K.: Overview of the NTCIR-14 QA Lab-PoliInfo Task. In: Proceedings of the 14th NTCIR Conference, 2019.

[2]Lin, Chin-Yew. Rouge: A package for automatic evaluation of summaries. Text summarization branches out: Proceedings of the ACL- 04 workshop. Vol. 8. 2004.

[3]S. Kullback and R. A. Leibler. On information and sufficiency. Ann. Math. Statist. Vol. 22. 1951.