FU-02 Team's Classification of Fact-checkable Opinions in NTCIR-14 QA Lab-PoliInfo Task

Ginya Nishijima[1], Masahiro Shiratori[1], Hokuto Ototake[1], Toshifumi Tanabe[1], and Kenji Yoshimura[1]

Fukuoka University, Fukuoka, Japan
{tl151314,tl151232}@cis.fukuoka-u.ac.jp,
{ototake,tanabe,yosimura}@fukuoka-u.ac.jp

Abstract. This paper reports on the achievements of Classification subtask of the NTCIR-14 QA Lab-PoliInfo task of FU-02 team. We propose a method for classifying pros and cons of a political topic, whether an utterance sentence includes fact-checkable reasons or not, and whether an utterance sentence is relevant with the topic. Our proposed method consists of three different classifiers which are based on a simple rule, keywords, and word embeddings respectively.

Team Name. FU-02

Subtasks. Classification Task (Japanese)

Keywords: Classification · Rule-based method · Word embeddings

1 Introduction

We, FU-02 team, participated in Classification subtask of the NTCIR-14 QA Lab-PoliInfo task [1]. Classiffication task aims at finding pros and cons of a political topic and presenting their fact-checkable reasons. In the subtask, a political topic such as "The Tsukiji Market should move to Toyosu" and an utterance sentence in assembly minutes are given. Participants including us classify four kinds of labels on given sentences, *Fact-checkability*, *Relevance*, *Stance* and *Class*. *Fact-checkability* means whether or not a sentence contains fact-checkable reasons. *Relevance* means whether or not a given sentence refer to a given topic. *Stance* means whether or not a speaker of the sentence agrees on the topic. A label of *Class* depends on labels of the other three class, as shown in Table 1. *Fact-checkability* and *Relevance* labels are regarded as 2-class classification, which their values can be 0 (absence) or 1 (existence). *Stance* labels are regarded as 3-class classification, which their values can be 0 (other), 1 (agree) or 2 (disagree).

We proposed a method for classifying these labels and submitted the results. Our proposed method consists of classifiers corresponding to each of the three kinds of labels, *Relevance*, *Fact-checkability* and *Stance*. *Relevance* classifier is based on a simple rule. *Fact-checkability* is classified by keywords registered in

our original dictionary. *Stance* classifier is based on word embeddings. *Class* are determined from the classification results of the other three kinds of labels. We describe the proposed method in Section 2, and evaluation results of the method in Section 3. Finally, Section 4 is the conclusion.

Table 1. Relationship between *Class* and the other three labels

Class	*Fact-checkability*	*Relevance*	*Stance*
1	1	1	1
2	1	1	2
0	All other combinations		

2 Classification Methods

In this section, we describe our methods for estimating three kinds of labels, *Relevance*, *Fact-checkability* and *Stance* respectively. For extracting words and recognizing their part-of-speech, we use MeCab morphological analyzer [2] with IPADIC dictionary in the methods.

2.1 Relevance

We consider an utterance sentence as being relevant with the topic of the utterance if the sentence includes at least one of the nouns in the topic.

2.2 Fact-checkability

We created a keyword dictionary by referring the Japanese Multiword Expression Lexicon (JMWEL)[1] [3]. Our keyword dictionary contains 32 words as shown in Table 2 that are registered in JMWEL as connection particle attribute representations and have the meaning of cause or guess. We consider an utterance sentence as fact-checkable one if the utterance sentence includes at least one of the keywords in the dictionary.

2.3 Stance

Based on original training data created by the author, we construct a *Stance* classification model using fastText text classifiers[2] [4]. The author created the training data from all utterance sentences, 2,342 sentences, of Tokyo Metropolitan Assembly minutes No.1 and No.2 in 2012. The author subjectively classified the utterance sentences in the three kinds of *Stance*: approval, opposite or the other. As a result, the breakdown of the training data 2,342 sentences was 107 for approval, 232 for opposite, and 2,003 for the other.

[1] http://jefi.info/
[2] https://fasttext.cc/

Table 2. *Fact-checkability* keywords in our dictionary.

から	より	よって	比べ	比較	が理由で
関連して	を基に	および	のに加えて	だけでなく	はもちろん
は無論	に際して	に対して	と同様に	という関係上	ではなく
の他に	の為に	を踏まえ	以外に	ので	結果
せいで	故に	と合わせて	のみならず	あげく	おかげで
の甲斐あって	のみならず				

3 Results of Formal Run

In this section, we describe the results of Formal run. In Formal run, the number of utterance sentences in the test data is 3,412. The number of topics is 14 as shown in Table 3. Since there are a large number of topics, we focus on topic "Casino" and describe the result.

Table 3. Topics

#	abbreviated name	topic sentence
1	カジノ (Casino)	カジノを含む統合型リゾートを推進するべきである
2	集団的自衛 (Self defense)	集団的自衛を認めるべきである
3	八ッ場ダム (Yanba)	八ッ場ダムの建設を進めるべきである
4	高齢者 (Elderly people)	高齢者への医療助成を増やすべきである
5	私学助成 (Private school grants)	私学助成を推進するべきである
6	中京都構想 (Medium Kyoto)	中京都構想を推進するべきである
7	オスプレイ (Osprey)	オスプレイを配備する
8	特定秘密保護法 (Secret protection)	特定秘密保護法案を進めるべきである
9	道州制 (Do-Shu-system)	道州制を導入するべきである
10	子ども医療 (Children medical expenses)	子ども医療費を無料化にするべきである
11	教員増加 (Regular faculty members)	正規の教員を増やすべきである
12	生活保護 (Welfare)	生活保護の基準額を引き下げるべきである
13	東京オリンピック (Tokyo Olympics)	東京にオリンピックを招致するべきである
14	空き家 (Vacant houses)	行政の判断で空き家を処理できるようにするべきである

Table 4 shows the confusion matrix of the *Relevance* classification result. Our method outputs that all utterance sentences are related to the topic. The recall of label 1 was 100%, and the precision was 97.2%. None of the sentences of label 0 which are slightly included can not be classified properly.

Table 4. Confusion matrix of the result of *Relevance* classification.

correct \ estimated	0	1
0	0	30
1	0	1,065

Table 5. Confusion matrix of the result of *Fact-checkability* classification.

correct \ estimated	0	1
0	383	179
1	340	193

Table 6. Confusion matrix of the result of *Stance* classification.

correct \ estimated	0	1	2
0	649	58	182
1	97	18	17
2	52	2	20

Table 5 shows the confusion matrix of the *Fact-checkability* classification result. The recall of label 1 was 36.2%. Our method cannot extract half of correct fact-checkable sentences. Additionally, since the accuracy is low (52.6%), our method based on the keyword dictionary seems to be inadequate.

Table 6 shows the confusion matrix of the *Stance* classification result. The recall rates of labels 0, 1 and 2 are 73.0%, 13.6% and 27.0% respectively. These recall rates roughly agree with the amounts of training data.

Table 7 shows the confusion matrix of the *Class* classification result. Although the accuracy achieves 83.1%, this indicates almost how much the label 0 could be classified correctly since there are many data whose label is 0 for the correct answer. The recall rates of labels 0, 1 and 2 are 90.2%, 3.1% and 0% respectively. The precision rates of labels 0, 1 and 2 are 91.7%, 4.7% and 0% respectively. The reason why the recall and precision are very low is that despite the correct classification of *Stance*, it is conceivable that the *Class* can not be correctly classified by erroneously classifying *Fact-checkablility*.

4 Conclusions

We proposed a method for classifying pros and cons of a political topic, whether an utterance sentence includes fact-checkable reasons or not, and whether an utterance sentence is relevant with the topic. As a result of the evaluation, it was found that the accuracy of *Fact-checkability* classification was particularly low, which affected the accuracy of *Class* classification. If the classification accuracy of *Fact-checkability* can be improved, the accuracy of the final *Class* output result will improve.

Table 7. Confusion matrix of the result of *Class* classification.

correct \ estimated	0	1	2
0	908	39	59
1	57	2	4
2	25	1	0

In the classification of *Relevant*, there is a problem that all labels 1 are outputted. In order to solve this problem, exclusion of example expressions such as "例として (for example)" is considered.

Additionally, when supervised machine learning is performed like our method, it is necessary to consider how to collect training data. When using training data whose labels may change depending on the subjectivity of a person, the reliability may change depending on the topic. Experiments using different training data (about 1000 sentences) shows that the performances of some topics get better and get worse. It will be necessary to think about handling of training data involving subjectivity from now on.

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. and Inui, K.: Overview of the NTCIR-14 QA Lab-PoliInfo Task. In: Proceedings of the 14th NTCIR Conference, 2019.
2. Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Vol. 4, pp.230–237, 2004.
3. Tanabe, T., Takahashi, M. and Shudo, K.: A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing. In: Computer Speech and Language, 28-6, pp.1317–1339, Elsevier, 2014.
4. Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T.: Bag of Tricks for Efficient Text Classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2, Short Papers, pp.427–431, 2017.