TMCIT at the NTCIR-14 QALab-PoliInfo Task

Tatsuya Ogasawara[1] and Takeru Yokoi[1]

Tokyo Metropolitan College of Industrial Technology, 1-10-40, Higashi-oi, Shinagawa, Tokyo, Japan
s19010@g.metro-cit.ac.jp, takeru@metro-cit.ac.jp

Abstract. In this research, we classified the utterances of assemblymen according to three viewpoints: whether utterance is 1)relevant to the policy(Relevance Classification), 2)fact-checkable(Fact-checkability Classification), and 3)one of agreement, disagreement or neutral(Stance Classification). In the Relevance Classification experiment, classification was performed by the method using cosine similarity. Also, in the Fact-checkability Classification experiment, classification was performed by a decision tree classifier. The number of specific words such as evidence expression, named entity, etc., were used as the training features. Then, in the Stance Classification experiment, classification was performed by a support vector machine. The vector of polarity values for the word appearing in each utterance was used as the training feature for the support vector machine. The polarity values were decided by the emotion polarity dictionary.
As a result, the accuracy was approximately 80% in each classification result. However, in the minority class of each classification experiment, the scores of precision and recall were low.
In order to improve the scores, in the Relevance Classification and the Fact-checkability Classification, the preparation of more higher quality training data is necessary. In the Fact-checkability Classification, the evidence expression was not successfully extracted. Therefore, extraction of evidence expression based on a more complicated rule base than that used in this experiment is a future task. Also, in the Stance Classification, it is a challenge to construct features that capture polarity reversal or expression, or incorporate appropriate knowledge for the domain of data usage.

Keywords: NTCIR-14 · QA Lab · PoliInfo · TMCIT · Classification Task(Japanese)

Team Name. TMCIT

Subtasks. Classification Task (Japanes)

1 Introduction

The purpose of this research is to classify that utterances of assemblymen included in the regional assembly minutes into the following three classes: 1) agreement utterances with fact-checkable evidence, 2) disagreement utterances with fact-checkable evidence and 3)other utterances. For this purpose, we have classified utterances according to three viewpoints: whether the utterance is 1)relevant to the policy(Relevance Classification), 2)fact-checkable(Fact-checkability Classification) and 3)one of agreement, disagreement or neutral(Stance Classification).

In the Relevance Classification, we have classified utterances by cosine similarity based on the Tf-Idf value of each utterance. In the Fact-checkable Classification, we have classified utterances using the decision tree constructed based on evidence expressions and named entities. In the Stance Classification, we have classified utterances using the support vector machine. It was trained based on words with the emotion polarity value included in each utterance.

In this paper, we will describe the related works in Section 2 and explain each classification method in Section 3. In addition, we will explain the experimental data and the results of the experiment.In this research, the accuracy was finally obtained around 80% in each classification result. However, in the minority class of each classification experiment, the scores of precision and recall were low. Therefore, we will discuss the high and low in accuracy, precision, and recall based on the experimental results.

2 Related Work

Akamine et al. have proposed the WISDOM[1]. WISDOM is the information analysis system which collects web information from various viewpoints by classifying web pages in various ways, and presents the overall image of them. This system analyzes the web pages from the viewpoint of "information sender", "evaluation information (such as opinion) on topics", "page appearance", etc. Then, it helps users verify the reliability of web information by classifying web pages from the viewpoint "by sender class", "by affirmative denial" and "by advertisement amount".

In this research, our goal is to extract utterances with fact-chackable evidence by classifying the regional assembly minutes from the viewpoint of "relevance", "fact-checkability" and "stance". The research by Akamine et al. targeted web pages. The web page is a document which consists of multiple sentences. This research targets one sentence in the regional assembly minutes.

3 Classification Methods

In this section, we explain methods of the Relevance Classification in Section 3.1, the Fact-checkability Classification in Section 3.2, and the Stance Classification in Section 3.3, respectively.

3.1 Relevance Classification

Utterances on various policies exist in the regional assembly minutes. In this section, the method of classifying utterances according to whether a utterance is related to a certain policy is explained.

In this research, the utterances with labels were used as training data. If the utterance is related to a policy, the label 1 was assigned, and otherwise, 0 was assigned. In addition, N utterances without labels were used as test data. Then, the Relevance Classification was performed by the cosine similarity between training utterances and test utterances.

Utterances with the label 1 of training data are combined into one sentence, only nouns and adjectives are extracted from them, and one word set is constructed. In the same way, one word set is also constructed for the label 0. In the test data, also only nouns and adjectives of each utterance are extracted, and N word sets are constructed. MeCab[3], which is one of Japanes morphological analysis systems, is used to extract nouns and adjectives. The mecab-ipadic-NEologd[5] is used as a dictionary of MeCab. It updates data daily from language resources on the Web. The same system and dictionary are used for morphological analysis from then on.

Tf-Idf values are derived for the elements of those word sets, respectively, and vectorization is performed. Here, word vectors of the label 1 and $0(\boldsymbol{R_1}, \boldsymbol{R_0})$ and N unlabeled word vectors $\boldsymbol{R^{(i)}}$ were prepared.

The similarity of N unlabeled word vectors $\boldsymbol{R^{(i)}}$ to the word vector $\boldsymbol{R_1}$ of the label 1 and also $\boldsymbol{R_0}$ of the label 0 is calculated using the cosine similarity as shown in Equation (1). Here, $\boldsymbol{C_1}$ is the cosine similarity vector which calculated the cosine similarity of each unlabeled word vector $\boldsymbol{R}$, with respect to the word vector $\boldsymbol{R_1}$ of the label 1. Likewise, $\boldsymbol{C_0}$ is the cosine similarity vector of $\boldsymbol{R}$ for $\boldsymbol{R_0}$. Here, $|\boldsymbol{R}|$ denotes the number of elements of the $\boldsymbol{R}$. The classification of utterances without labels was performed by comparing values of the i-th cosine similarity, $\boldsymbol{C_1^{(i)}}$ and $\boldsymbol{C_0^{(i)}}$ as following Equation (4). In Equation (4), cosine similarity $\boldsymbol{C_1^{(i)}}$ and $\boldsymbol{C_0^{(i)}}$ were divided by their average value $average(\boldsymbol{C_1})$ or $average(\boldsymbol{C_0})$, respectively. As a result, the comparison was performed after taking into account the scale of each cosine similarity vector.

$$cos(\boldsymbol{R_x}, \boldsymbol{R_y}) = \frac{\sum_{j=1}^{|\boldsymbol{R}|} R_x^{(j)} R_y^{(j)}}{\sqrt{\sum_{j=1}^{|\boldsymbol{R}|} {R_x^{(j)}}^2} \cdot \sqrt{\sum_{j=1}^{|\boldsymbol{R}|} {R_y^{(j)}}^2}} \tag{1}$$

$$\boldsymbol{C_1} = cos(\boldsymbol{R_1}, \boldsymbol{R}) \tag{2}$$

$$\boldsymbol{C_0} = cos(\boldsymbol{R_0}, \boldsymbol{R}) \tag{3}$$

$$label_R^{(i)} = \begin{cases} 1\ if \frac{\boldsymbol{C_1^{(i)}}}{average(\boldsymbol{C_1})} > \frac{\boldsymbol{C_0^{(i)}}}{average(\boldsymbol{C_0})}, \\ 0\ otherwise. \end{cases} \tag{4}$$

3.2 Fact-checkability Classification

In this section, a method of classifying utterances according to whether a certain utterance has fact-checkable evidence is explained. In this research, a decision tree was constructed using training data with the label of 1 if a certain utterance has fact-checkable evidence, otherwise labeled 0. Using the decision tree, the Fact-checkability Classification was performed on the test data without label.

The following six types of expressions were used as features for training data of decision tree training.

Evidence Number of words used to indicate evidence such as "for"("tame" in Japanese),"so"("node" in Japanese), "because"("dakara" in Japanese), "therefore"("shitagatte" in Japanese).

Numeral Number of half-width and double-byte numeric character strings and numbers of Chinese numeric character(zero, one, ..., nine, ten, hundred, ten thousand, one hundred million, trillion, etc.) strings.

Time Number of words representing the era such as Heisei, Showa, Meiji and other dates and periods such as days, months, years, days of the week, and other time expressions.

Money Number of words representing units of currency such as yen, dollar, euro, and other money expressions.

Percentage Number of words representing percentages such as percentage, %, multiply and other percentage expressions.

Named Entity Number of words of each person's name (family name, first name, last name), region (country, others), organization, and other named entities. Such words are extracted by MeCab.

In the parameter tuning of the decision tree classifier, the tree depth (max_depth) was set from 2 to 20, and the minimum value of branch destination (min_samples_split) was set from 2 to 30 by the grid search. The optimal hyperparameters were searched by 5-fold cross validation. The Gini coefficient was used as an indication of impurity of each node.

3.3 Stance Classification

In this section, a method of classifying a utterance according to whether it is in a position of agreement, disagreement or neutral is explained. In this research, the label 0 is assigned if the utterance is neutral, the label 1 is assigned if the utterance is in a position of agreement, and the label 2 is assigned if it is in the position of disagreement. We used the utterances with such labels as the training data for the support vector machine. Using the trained support vector machine, the Stance Classification was performed for the unlabeled utterances.

The features used for the support vector machine are explained. First, a feature vector whose size is the total utterance number N multiplied by the number of word types N_w. Here, for word types, only words contained in the emotion polarity dictionary were targeted out of all words of the training data and the test data. In other words, a feature vector is like a bag-of-words. The value of each element of the feature vector is assigned with the emotion polarity value of the word included in each utterance. Also, the sum of the polarity values of each utterance is added at the end of each row of the vector. The dimension of feature vectors is as many as word types ($N_w + 1$) and extends over several thousands of dimension. Therefore, the dimension reduction has been carried out up to 150 dimensions by principal component analysis. RBF was used as a kernel function of the support vector machine. In the parameter tuning, the cost parameter C was set to 1, 10, 100, and 1000, and the gamma value was set to 0.001 and 0.0001 by the grid search. The optimal hyperparameter was searched by 5-fold cross validation.

4 Classification Experiment

In this section, the data used in the evaluation experiment and experimental results are shown.

4.1 Data Used in the Experiment

In this research, the regional assembly minutes corpus were provided by Kimura et al[2]. In these data, “task statement” in which the policy is described, “utterance (only sentence)” and four class labels are included. There are 14 kinds of policies in total. In addition, these data are labeled by 3 or 5 annotators. In this experiment, the training data was prepared with the label given by the most annotators as the correct answer for the Relevance Classification so that the recall is higher based on preliminary experiments. In the training data in the Fact-checkability Classification, and the Stance Classification, when a minority label was given even by one person, the label was regarded as the correct answer.

Also, Table 1 summarizes the Kappa values, which are indicators representing the degree of concordance of labeling of annotators for each classification. Further, the label ratios in the training data are summarized in Table 2. In the Relevance Classification, the label 0 means no relevance, and the label 1 means relevant to the policy. In the Fact-checkability Classification, the label 0 means fact check impossibility, and the label 1 means fact check possibility. In the Stance Classification, the label 0 means neutral, the label 1 means agreement, and the label 2 means disagreement. Then, in the Final Classification, the label 1 means the position of agreement with fact-checkable evidence, the label 2 means the position of disagreement with fact-checkable evidence, and the label 0 means otherwise.

Table 1. Average of Kappa value in 14 kinds of policies

	Relevance	Fact-checkability	Stance
Kappa	**0.114**	**0.280**	**0.348**

In the emotion polarity dictionary, the lists of words and their semantic orientations generated by the Takamura et al.[4] were used. In this dictionary, about 55,000 words assigned real values of -1 to $+1$ are included. In this dictionary, the negative expression is assigned the polarity value closed to -1, and the positive expression is assigned the polarity value closed to $+1$.

Table 2. The label ratio in label-by-label

Class	Relevance		Fact-checkability		Stance			Final		
	0	1	0	1	0	1	2	0	1	2
Ratio	**0.13**	0.87	0.65	**0.35**	0.8	**0.12**	**0.08**	0.94	**0.04**	**0.02**

4.2 Experimental Result

In this section, the results of the evaluation experiment are shown in Tables 3 to 10. The evaluation was carried out with the evaluation method Nn which makes the label given by $n(0 \leq n \leq 3)$ or more people the correct answer. For example, when a label given by two or more people was taken as a correct answer, it is described as an evaluation method $N2$. Table 10 shows the results obtained by integrating the results of the Relevance Classification, the Fact-checkability Classification, and the Stance Classification and classifying them into agreement utterance with fact-checkable evidence, disagreement utterance with fact-checkable evidence, and the other.

In Tables 3 to 6, confusion matrices indicating output labels for each correct answer data are shown.

In Tables 7 to 9, the score of the evaluation index for each label is shown as the classification experiment result. For the evaluation index, accuracy, precision for each label, and recall for each class were used.

Table 3. Confusion matrix of the Relevance Classification

Correct		N1			N2		N3		
		0	1	0,1	0	1	0	1	-
Output	0	**22**	217	601	**264**	576	**22**	219	599
	1	3	**2,117**	452	44	**2,528**	3	**2,122**	447

Table 4. Confusion matrix of the Fact-checkability Classification

Correct		N1			N2			N3		
		0	1	0,1	0	1	0,1	0	1	-
Output	0	**894**	55	928	**1,592**	268	17	**930**	67	880
	1	243	**323**	969	814	**712**	9	249	**337**	949

5 Discussion

In this section, discussions will be made on the results obtained in each classification experiment in Section 4, such as the high and low of the precision , and the recall.

Table 5. Confusion matrix of the Stance Classification(N1)

Correct		N1						
		0	1	2	0,1	0,2	1,2	0,1,2
Output	0	**1,806**	72	66	439	399	4	6
	1	164	**81**	1	259	17	0	1
	2	31	0	**15**	9	41	1	0

Table 6. Confusion matrix of the Stance Classification(N2,N3)

Correct		N2						N3			
		0	1	2	0,1	0,2	-	0	1	2	-
Output	0	**2,408**	181	184	4	13	2	**1,834**	80	70	808
	1	343	**161**	6	0	12	1	178	**89**	2	254
	2	63	2	**32**	0	0	0	31	0	**15**	51

Table 7. The Relevance Classification experiment result

		N1	N2	N3	Average
Accuracy		0.936	0.818	0.906	0.887
Precision	Label 0	**0.742**	**0.314**	**0.091**	**0.382**
	Label 1	0.999	0.983	0.999	0.994
Recall	Label 0	0.578	0.857	0.880	0.772
	Label 1	0.758	0.814	0.906	0.826

Table 8. The Fact-checkability Classification experiment result

		N1	N2	N3	Average
Accuracy		0.913	0.683	0.800	0.799
Precision	Label 0	0.971	0.857	0.933	0.920
	Label 1	0.842	**0.470**	0.575	0.629
Recall	Label 0	0.601	0.662	0.789	0.684
	Label 1	0.568	0.717	0.834	0.706

Table 9. The Stance Classification experiment result

		N1	N2	N3	Average
Accuracy		0.893	0.771	0.843	0.836
Precision	Label 0	0.947	0.869	0.924	0.913
	Label 1	0.65	0.331	0.331	**0.437**
	Label 2	0.577	0.330	0.326	**0.411**
Recall	Label 0	0.835	0.853	0.898	0.862
	Label 1	0.395	0.469	0.527	**0.464**
	Label 2	0.104	0.142	0.172	**0.139**

Table 10. The Finally Classification experiment result

		N1	N2	N3	Average
Accuracy		0.946	0.908	0.936	0.93
	Label 0	0.996	0.980	0.994	0.99
Precision	Label 1	0.387	0.081	0.053	**0.174**
	Label 2	0.350	0.100	0.000	**0.150**
	Label 0	0.921	0.924	0.941	0.929
Recall	Label 1	0.250	0.333	0.471	**0.351**
	Label 2	0.063	0.118	0.000	**0.060**

5.1 Relevance Classification

Focusing on Table 2, it can be observed that the label ratio of relevance is only about 1/4 of the label 1, with the label 0 being about 13% as a whole. Therefore, by making the training data prepared with the label given by the most annotators as the correct answer for the Relevance Classification, the recall is higher based on preliminary experiments. As a result, the accuracy is 88.7%, the recall is 77.2% for the label 0, the precision is 99.4%, and the recall is 82.6% for the label 1, that are a relatively high score. However, the precision is a considerably low score of 38.2% for the label 0. Here, in the evaluation index Nn, as the number n increases, the number of labels regarded as correct answers decreases. For this reason, it is considered that the precision for the label 0, which is a minority label, decreases remarkably as n increases.

Focusing on Table 3, in the evaluation method $N1$, the data considered to be correct is 25 utterances for the label 0 and 2,334 utterances for the label 1. There are 1,053 utterances that are considered as the correct answer regardless of the label 0 or 1 output. In the evaluation method $N2$, the label 0 is 308 utterances and the label 1 is 3,103 utterances, which are regarded as correct answers. In addition, there is no data regarded as a correct answer no matter which is output. As can be observed from the above, in $N2$, compared with $N1$, the precision decreases unless prediction is more accurate. In the evaluation method $N3$, the label 0 is 25 utterances and the label 1 is 2,341 utterances, which are regarded as correct answers. There is no data regarded as a correct answer no matter which is output, and there are 1,046 utterances which are not regarded as correct answers in the first place. Therefore, in $N3$, it is considered that the precision is further reduced as compared with $N1$, and $N2$,.

Next, focusing on Table 3, it is 840 utterances that the label 0 is output. However, the label 0 is 1,078 utterances, 308 utterances, and 25 utterances in evaluation methods $N1$, $N2$ and $N3$, respectively. From this, it can be observed that in the evaluation of $N2$ or more, the label 0 is excessively output. As described above, it is considered to be due to the use of training data that the recall increases. Besides that, from Table 1, it is considered that the precision and the recall are reduced because the concordance rate of annotator's labeling is as low as about 0.114 on average. From the above, preparation of better quality training data is one of the tasks in the future.

Next, the results of extraction of 20 words in the order of descending Tf-Idf values in the word vector of the label 0 and 1 in the policy ”We should promote integrated resorts including casinos” are shown as follows.

The Label 0 Ten, hundred, chairman, two, six, eight, seven, Daio Paper, illegality, 10 billion yen, casino, four, person, car, major companies, store, paper company, time, arrest, loss (in Japanese 十, 百, 会長, 二, 六, 八, 七, 大王製紙, 不正, 100 億円, カジノ, 四, 人, 車, 大手, 店, 製紙会社, 時, 逮捕, 損失, respectively).

The Label 1 Casino, things, ir, attract, of, bringing in, to, integrated resort, country, facility, inside, governor, sightseeing, this, resort, consideration, Singapore, such, integration, citizens of a prefectural (in Japanese カジノ, こと, ｉｒ, 誘致, の, 導入, よう, 統合型リゾート, 国, 施設, 中, 知事, 観光, これ, リゾート, 検討, シンガポール, 等, 統合, 県民, respectively).

From this result, it is found that the Tf-Idf value of the numerals such as “ten”, “hundred”, and “two” is high on the label 0. Also, many words related to the so-called ”Daio Paper Incident” such as ”Chairman”, ”Daio Paper”, ”Illegal”, ”10 billion yen”, ”Arrest”, etc., are extracted. However, these do not directly relate to the policy. From that, it can be said that important words are extracted correctly. In the example of the label 1, words directly related to policy such as “casino”, “integrated resort”, “attract”, “bringing in”, etc., can be extracted as words of high importance. From the above, it turned out that it is effective to use the word vector of Tf-Idf value.

5.2 Fact-checkability Classification

In Fact-checkability Classification, classification was performed using decision trees. Focusing on Table 8, scores of accuracy, precision and recall of nearly 70% were obtained, respectively. However, the precision for the label 1 was approximately 47% in the evaluation method $N2$, which was a low score. Since the label 1 was assigned to more utterances than those originally with the label 1, it is considered that the precision is low.

Next, among the constructed decision trees, some of the classification processes up to nodes that have relatively well classified samples with a sample count of 50 or more and impurity less than 0.1 are shown below.

Example of processes classified as fact-checkable: 1

1. Number of named entity>1.5,gini$=0.453$,samples$=1,165$,value$=[404,761]$
2. Number of numeral>0.5,gini$=0.311$,samples$=534$,value$=[103,431]$
3. Number of time expression>0.5,gini$=0.175$,samples$=238$,value$=[23,215]$
4. gini$=0.085$,samples$=135$,value$=[6,129]$

Example of processes classified as fact-checkable: 2

1. Number of numeral>0.5,gini$=0.382$,samples$=664$,value$=[171,493]$
2. Number of time expression>1.5,gini$=0.258$,samples$=381$,value$=[58,323]$
3. gini$=0.088$,samples$=87$,value$=[4,83]$

Example of processes classified as fact-checkable: 1

1. Number of numeral≤ 1.5,gini$= 0.278$,samples$= 1,174$,value$= [978, 196]$
2. Number of named entity≤ 2.5,gini$= 0.121$,samples$= 864$,value$= [808, 56]$
3. gini$= 0.041$,samples$= 481$,value$= [471, 10]$

Example of processes classified as fact-checkable: 2

1. Number of time expression≤ 1.5,gini$= 0.329$,samples$= 241$,value$= [191, 50]$
2. Number of numeral≤ 0.5,gini$= 0.204$,samples$= 199$,value$= [176, 23]$
3. gini$= 0.019$,samples$= 103$,value$= [102, 1]$

Here, the expression of inequality sign is a conditional expression for the next node, "gini" is an index indicating impurity, "samples" is the number of samples at that node, "value" is the number of data whose left value denotes not fact-checkable and the value on the right denotes fact-checkable. Focusing on these processes, it seems that three kinds of numbers of named entity, numeral, and time expression, contribute greatly to the Fact-checkability Classification. In addition, other than the above, we found that the number of words represent money, country, place, etc. contributes as the feature quantity in the classification process up to well classified nodes. The number of words which represent evidence, percentage, person, organization, etc. does not correspond to the above, and it turned out that those features did not contribute much.

It is conceivable that the reason why the number of evidence expressions does not contribute much to the Fact-checkability Classification, is that words such as "tame", "node", "dakara", and "shitagatte" could not be correctly extracted. For example, "tame" appearing in the following two sentences is a evidence expression that expresses the cause or reason in the first sentence, and in the second sentence is the objective expression that expresses the profit and the target. However, when morphologically analyzed in either case, they are extracted as morphemes given the same part-of-speech. Therefore, it is considered that it is necessary to extract evidential expressions with a more complicated rule base.

- Regarding casinos, discussions of the necessity of redevelopment in country law has been long going in the city council, **as the topic violates gambling laws**(in Japanese, **"tobaku zai ni fureru tame"**).
- So, I would like to propose attracting casinos **in order to** pull tourists from overseas and **lead to economic growth**(in Japanese, "**keizai seicho ni tsunageru tame ni**").

5.3 Stance Classification

Experimental results of the Stance Classification resulted in the lowest score among the three classification experiments. Focusing on Table 9, precision and recall of the label 1 are approximately 44% and 46%, and for the label 0, the precision and recall are approximately 41% and 14%, which are considerably low scores. This is because from Table 2, the ratios of the label 1 and 2 are low, so it is conceivable that precision and recall are low.

Based on the above, since majority labels have a great influence on training, it is thought that improvement is achieved by weighting each label. Therefore, for each label, a weight of w_{label} as shown in Equation (5) was added and a re-experiment was conducted. As a result, it was evaluated as shown in Table 1. Here, $N_{samples}$ denotes the number of training data, N_{labels} denotes the number of label types, and N_{label} denotes the corresponding label.

$$w_{label} = \frac{N_{samples}}{N_{labels} \times N_{label}} \tag{5}$$

From the re-experimental result of weighting(Table 11), recall of the label 2 yielded a dramatic improvement in score of approximately 20% to 30%. However, the score still remains below 50%. Furthermore, as a result of weighting, the accuracy, precision, and recall of each label showed decreases of 11% to 4%. From this fact, although the weighting can improve the recall of the label 2, it is considered that it has not reached a fundamental solution.

Table 11. Re-Experimental Result of Stance Classification

		N1	N2	N3	Average
Accuracy		0.840	0.707	0.776	**0.774**
	Label 0	0.956	0.884	0.937	0.926
Precision	Label 1	0.605	0.290	0.291	**0.395**
	Label 2	0.461	0.257	0.187	**0.302**
	Label 0	0.729	0.750	0.808	**0.762**
Recall	Label 1	0.434	0.485	0.556	0.492
	Label 2	0.327	0.436	0.460	**0.408**

Next, focusing on the features used in training. Then, in the Stance Classification experiment, classification was performed by a support vector machine that trained. The vector of polarity values for the word appearing in each utterance was used as the training feature. The polarity values were decided by the emotion polarity dictionary. However, with such a simple feature, we think that it was not correctly classified or detected. Specifically, the types of approval or disapproval are quite varied, and it is considered that a polarity reversal or expression occurs depending on a specific word pair or word set. In addition, it is necessary to extract the polarity according to the policy, since a reversal of approval / disapproval can also occur for each policy. Based on the above, it is a challenge to construct features that capture a polarity reversal or expression, or incorporate appropriate knowledge for the domain of data usage.

6 Conclusion

In this research, we classified the utterances of assemblymen according to three viewpoints: whether utterance is 1)relevant to the policy(Relevance Classification), 2)fact-checkable(Fact-checkability Classification), and 3)one of agreement, disagreement or neutral(Stance Classification). In the Relevance Classification experiment, classification was performed by the method using cosine similarity. Also, in the Fact-checkability Classification experiment, classification was performed by a decision tree classifier that trained by the number of words such as evidence expression, named entity, etc., as features. Then, in the Stance Classification experiment, classification was performed by a support vector machine that trained. The vector of polarity values for the word appearing in each sentence was used as the training feature. The polarity values were decided by the emotion polarity dictionary.

As a result, the accuracy was approximately 80% in each classification result. However, in the minority class of each classification experiment, the scores of precision and recall were low. Specifically, the precision score for irrelevant utterances in the Relevance Classification was approximately 38%. Also, the precision for the fact-checkable utterance in the Fact-checkability Classification was approximately 63%, which is low. Moreover, in the Stance Classification, the agreement and disagreement scores were extremely low, with results of precision of approximately 44% and 41% and recall of 46% and 14%, respectively.

In order to improve the above issues, in the Relevance Classification, the preparation of higher quality training data is necessary. Also, the Fact-checkability Classification had the same problem as the Relevance Classification. In the Fact-checkability Classification, the evidence expression was not successfully extracted. Therefore, extraction of evidence expression based on a more complicated rule base than that used in this experiment is a future task. In the Stance Classification, the polarity was not captured well because the feature was simple. Therefore, it was a challenge to construct features that capture a polarity reversal or expression, or incorporate appropriate knowledge for the domain of data usage.

References

1. Susumu Akamine, Daisuke Kawahara, Yoshikiyo Kato, Tetsuji Nakagawa, Kentaro Inui, Sadao Kurohashi, and Yutaka. Kidawara. : “WISDOM: A web information credibility analysis systematic”. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 1–4, Suntec, Singapore, August 2009. Association for Computational Linguistics.
2. Shibuki H. Ototake H. Uchida Y. Takamaru K. Sakamoto K. Ishioroshi M. Mitamura T. Kando N. Mori T. Yuasa H. Sekine S. Inui K. Kimura, Y. : “overview of the ntcir-14 qalab-poliinfo task.”. In *Proceedings of the 14th NTCIR Conference*, 2019.
3. T. KUDO. : “mecab : Yet another part-of-speech and morphological analyzer”.
4. Hiroya Takamura, Takashi Inui, and Manabu. Okumura. : “extracting semantic orientations of words using spin model”. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pages 133–140, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.
5. Taiichi Hashimoto Toshinori Sato and Manabu Okumura. : “implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval (in japanese)”. In *Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing*, pages NLP2017–B6–1. The Association for Natural Language Processing, 2017.