CUTKB at *NTCIR-14 QALab-PoliInfo* Task

Toshiki Tomihira and Yohei Seki

University of Tsukuba Ibaraki, Japan 305-8550
tomihira@klis.tsukuba.ac.jp
yohei@slis.tsukuba.ac.jp

Abstract. The news has spread quickly due to the development of social media. It will be a problem when the *fake news* has gone viral. Since the fake news was related to the politics in many cases, we focused on automatic estimation of fact-checkability using Japanese Regional Assembly Minutes Corpus. To verify the fact-checkability in the sentences correctly, it is essential to focus on the sentences which contain the evidence of the facts. In this paper, we explain our deep learning model to estimate "Relevance", "Fact-checking", and "Stance" for the sentences of the Minutes. Furthermore, we investigate whether the model combining *CNN* and *LSTM* is effective to check the facts.

Keywords: Fake News · Fact check · LSTM · Deep learning

Team Name CUTKB

Subtasks Classification subtask

1 Introduction

News has spread quickly with the development of social media. In many cases, traditional media such as television and newspapers delivered the articles after the third party organizations deliberated and edited them. In the social media, the users sometimes posted and delivered the article without deliberation, and they had no gatekeeper to secure information. In recent years, it has become an issue to spread the *fake news* through social media. A lot of fake news on politics was circulated already. It was known that the fake news had a big influence in the US presidential election. When the information including lies spread more quickly than the correct information[11], it will be a serious problem.

Organizations such as "Snopes", "FactCheck" and "Politifact" have announced the news with investigating the authenticity of them manually as measures to exclude the fake news. The manual efforts, however, will take some time, while the information on lies is transmitted quickly. In order to estimate the authenticity about the political news automatically, it will be necessary to check the facts in the Minutes for the politician who were referred in the news articles. In this paper, we summarized our research that challenged the classification subtask in *NTCIR-14 QALab-PoliInfo*, by focusing on checking facts in the Minutes relating to the politics[3].

To check the facts for the Minutes relating to the politics in the classification subtask, organizers provided three submodules. The first module is to check whether the target sentence can be confirmed and verified factually. The second module is to extract relevant sentences from the Minutes. The third module is to check whether relevant target texts will be supportive or disprovable. We focus on the first step, since it is essential to check whether sentences are verifiable.

In the previous researches, the models incorporating multi-task learning[4] and CNN[7] were proposed. We propose a model combining CNN and LSTM for fact-checking with convolution and sequence operations to take into the consideration of the contexts in the Minutes appropriately. In checking the facts, we supposed that it was essential to focus on the sentences which contained the evidence of facts. Our system attained relatively high scores among the task participants for fact-checking module. In this paper, we explain our model to predict “Relevance”, “Fact-checking”, and “Stance” in order to estimate the facts, especially focusing on the model combining CNN and LSTM for fact-checking.

2 Related work

Recently, many researchers pay attention to fake news checking since it had an impact in the election of the United States. Therefore, tasks such as Fact-Checking[1], Fake News Challenge[2] and Rumor-Eval[3] for the purpose of countermeasures against Fake News are held. Especially Fact-Checking task is similar in purpose to this task and it provides the president’s discussion dataset. In addition, researchers develop datasets using real news as a practical research and study to distinguish them based on information sources. Although we can think of several factors for the judgment of fake news, we particularly describe previous research on Fact-checking and stance detection mentioned in this paper.

Fact-checking It is regarded as a problem to answer lie information at the QA community site. Therefore, research on Fact-checking at the QA community site has been conducted for a long time[6]. Ma and colleagues are challenging the fact-checking by using RNN model for actual microblog data[5]. In addition, Kochkina and colleagues attempt on Rumor-Eval by using multi-task learning because it requires multidirectional of learning to confirm facts[4].

Stance As previous studies, LSTM[15] and Multi-layer perceptron[10] have been used to determine the position. In particular, Mohtarami et al. use End-to-End Memory Networks to judge the status of “agree”, “disagree”, “discuss” and “irrelevant”. Further they record state-of-the-art in the Fake News Challenge[7].

[1] http://www.fakenewschallenge.org/

[2] http://alt.qcri.org/clef2018-factcheck/

[3] http://alt.qcri.org/semeval2017/task8/

They state that not only LSTM model which focus on the sequence but also needs CNN to check the facts.

Some of these previous studies use English news and Minutes as datasets, but few studies use Japanese datasets. In this paper, we propose a model using CNN and LSTM, which fit to check the facts of the politics in the Minutes written in Japanese, based on the previous researches.

3 Dataset

In this study, we challenge the classification subtask in *QALab-PoliInfo* and used the dataset provided by the organizers [3]. We used 10,000 pieces of data that were manually labeled as "Relevance", "Fact-checkability", and "Position" correspond to the theme and the content text of Minutes. "Relevant" is a binary value of presence / absence, "Fact-checkability" is a binary value of possible / impossible, and "Position" is a label which takes three values: support / against / other. "Relevance" is the relationship between the theme of the Minutes and each sentence. Fact-checkability determines whether a sentence can contain specific expressions such as facility names, dates, and amounts. "Position" also determines the standing position of each Minutes document for the subject. The dataset has 14 types of task sentences and 10,000 sentences in total. We used 31,807 training data and 3,412 test data. A detailed data set description is provided in the task description paper.

4 Approach

This section describes our approach at the classification subtask in *QALab-PoliInfo*. We created a prediction model suitable for classifying the three types of labels, "Relevance", "Fact-checkability" and "Stance" described in the data set section.

4.1 Relevance

We regard the relevance task as binary classification whether each two inputs pair ("Topic" and "Utterance" columns) is relevant or not. We used a modified version of *MaLSTM* model, as proposed by Mueller et al[8]. The configuration for performing two-input binary classification with the LSTM model is shown in Figure 1. Due to the two inputs, we defined optimizer as *Manhattan distance* between two LSTMs obtained from the topic and from the utterance.

$$optimizer = exp(-||h^{(left)} - h^{(right)}||_1) \tag{1}$$

The parameters of the two LSTMs in Figure 1 are same. As parameters for Embedding, the vocabulary number is consisted of 1,000 words and the maximum length is 200. The hidden units of LSTM is 50, the gradient clipping norm is 1.25, the batch size is 64, and the number of epochs is 25 at the time of learning.

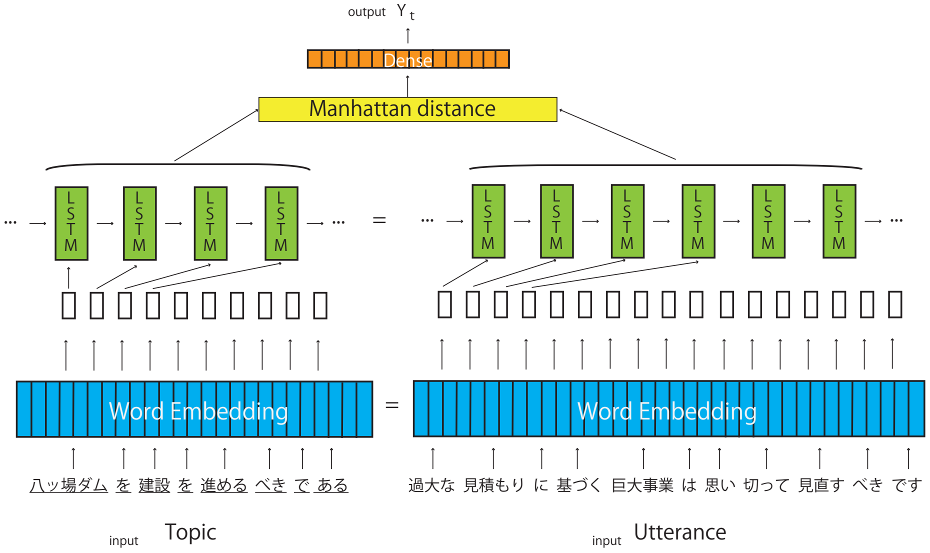
output Y t
Dense
Manhattan distance
LSTM
=
Word Embedding
=
Word Embedding
八ッ場ダム を 建設 を 進める べき で ある
過大な 見積もり に 基づく 巨大事業 は 思い 切って 見直す べき です
input Topic
input Utterance

Fig. 1. LSTM model for “Relevance” classification

4.2 Fact-checkability

In order to check the fact, it is useful to consider whether specific/concrete expressions such as the facility name, date, and money amount are included in the sentence or not. Of those expressions, the phrases relating to the fact-checking is difficult to find because their contents are different in each sentence. On the other hands, factual events tend to be referred in many documents of Minutes. Figure 2 shows an example of fact checkable sentences defined in the task: “the local government should be able to handle vacant houses. ” The green underlined parts are fact checkable parts. They refer to the similar facts relevant to the same topic. As indicated by the red underlines, it can be seen that these three sentences describe similar facts (use cases of vacant houses). We found that the similar facts were referred in the fact checkable sentences in the several cases.

空き家の活用につきましては、これまで福島県空き家・古民家相談センターを設立し、 県内への定住希望者に対する情報提供や改修相談等を実施してまいりました。
空き家の活用事例としましては、綾町と諸塚村において、空き家を改修した再生利用が 行われております。
厚生環境委員会で奄美大島を視察させていただいたとき、空き家対策として、 人口減対策として市が所有者から空き家を借り上げ、空き家を改修し、Ｕターン、 Ｉターン、奄美への移住者に低家賃で貸し出して人口減対策に取り組んでいるという話を 聞いてまいりました。

Fig. 2. Example sentences

Therefore, when we perform the fact-checkability classification task at the word level, it is more essential to focus on the words and phrases relevant to other documents as clues. We supposed that CNN-based model with convolution operation is appropriate to estimate the words/phrases relevance to the words/phrases in the other documents. Several previous researches proved that CNN model was suitable for document classification as in [2, 14].

On the other hand, task-relating verbs such as "実施してまいりました (have been conducted)", "行われております (have done) " and "取り組んでいる (are challenging)" are important features to estimate factuality, as shown by the blue underlines in Table 2. To take account of such context, LSTM using sequential information is suitable. We supposed that CNN and LSTM were necessary module to take these important features into consideration to estimate the factuality.

Mohtarami et al. proposed fake news detection system using CNN and LSTM [7]. The inputs were claims and news body sentences, which were the candidates of the evidence. In our dataset, the data set was Minutes, which did not contain the evidence information. We implement our fact-checkability estimation system by performing convolution and sequence prediction to take the relationship between the Minutes into consideration as a substitute for evidence (news body sentences) in Mohtarami's paper. Also, as prior research on a model combining CNN and LSTM, the CLDNN model was proposed by Sainath et al.[9]. They clarified the effectiveness of their model to estimate the sentence sentiment[12] and to identify claims[1].

Figure 3 shows the combined prediction model of LSTM and CNN. As parameters for Embedding, the vocabulary number is consisted of 1,000 words, and the maximum length is 150. As the parameters of CNN shown in Figure 3, the dropout was 0.2, the MaxPooling size was 4, and the activation function for convolution was *ReLU*. Also, as the parameters of LSTM shown in Figure 3, the number of hidden units was 100, activation was *Tanh*, and recurrent activation was hard *Sigmoid*. The total connection layer of the output part used *Sigmoid* as activation function and binary cross-entropy as loss function.

4.3 Stance

We estimate the standing position of "support", "disapproval" and "no matter" of a Minutes for the theme. For example, a support position is likely to contain positive words, such as "期待できる (promising)" or "考えられる (conceivable)". Also, if we are in a disappointing position, we think that the end of the sentence will be negative, such as "ふさわしくない (improper)" or "よくない (unfavorable)". This seems to be close to the positive / negative classification. There are previous studies using LSTM for positive / negative classification[13]. In this experiment, a simple LSTM prediction model shown in Figure 4 was used.

As parameters for Embedding, the vocabulary number is consisted of 1,000 words and the maximum length is 150. At the time of learning, the activation function is *ReLU*, dropout is 0.5, and the activation function of the last output layer is a sigmoid function. Also, the loss function is sparse categorical cross-entropy.

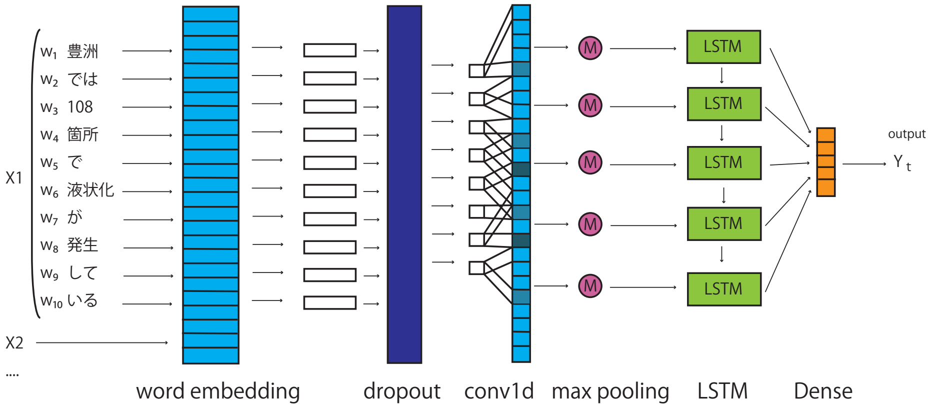

Fig. 3. LSTM-CNN model for “fact-checkbility” classification

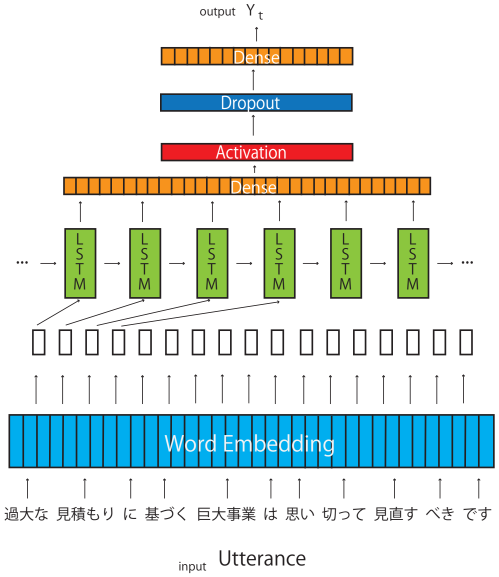

Fig. 4. LSTM model for “Stance” classification

5 Results

The results submitted in the formal run at the classification subtask in *QALab-PoliInfo* are shown in Table 1, Table 2, and Table 4.

5.1 Relevance

From the Relevance classification results in Table 1, we found that all responses of test data have been answered "existence" in the formal run. In this experiment, when creating embedding, it corresponds only to training data, so when a new word in test data comes out, it cannot cope. There are only 14 types of task statements in this data set, so we create a embedding with this data set, it will be difficult to apply the new test data is given. It is scarce to judge the relevance of new words or sentences, and we thought that the output is biased to one of the labels. We thought that the model has also been over-trained as data imbalance since we did not perform under sampling in training data.

Table 1. Result of relevance in formal run

	existence		absence	
Accuracy	Recall	Precision	Recall	Precision
0.865	1.000	0.865	0.000	NaN

5.2 Fact-checkability

The Fact-checkability results in Table 2 were the second highest score among all participants. As stated in the approach, it was confirmed that the model using LSTM and CNN is effective. Comparing the existence and absence scores, the absence was higher. In this model, we found that it was easier to predict "absence". The reason for that is considered that the number of data was uneven. As we mentioned in the approach, sentences that do not appear in other sentences are likely to be judged as in fact not checkable since we use convolution. The sentences that judged as absence is more diversity so the existence score seems to be lower.

Table 2. Result of fact-checkability in formal run

	existence		absence	
Accuracy	Recall	Precision	Recall	Precision
0.730	0.523	0.647	0.843	0.764

Table 3 shows Accuracy, Recall, and Precision with different evaluation metrics. N1: one or more; N2: two or more assessors; and N3: the label given by three or more assessors is regarded as the correct answer. SC uses the number of labels given by the creator as the weight of the correct score. The score was higher for all three people who gave a correct answer than for SC, so it is considered that the result regardless of people is better identified.

Table 3. All result of fact-checkability in formal run

Gold		existence		absence	
Standard	Accuracy	Recall	Precision	Recall	Precision
N1	0.966	0.782	0.978	0.406	0.938
N2	0.810	0.863	0.865	0.660	0.673
N3	0.918	0.944	0.945	0.841	0.839
SC	0.730	0.843	0.763	0.523	0.646

5.3 Stance

The results of Stance classification is shown in Table 2. Similar to Relevance, there was an imbalance in the training data since we did not under sample. The learning data was biased with about 25,000 “other”, about 4,000 “agree” and about 2,500 “disagree”. In addition, the score is low due to the data shaping problem of the submission data. Table 5 shows the results of retesting the formal run data without changing the model.

Table 4. Result of stance agreeing in formal run (problem)

	agree		disagree		other	
Accuracy	Recall	Precision	Recall	Precision	Recall	Precision
0.033	0.015	0.677	0.017	0.625	0.038	0.778

Table 5. Result of stance agreeing in formal run (fixed)

	agree		disagree		other	
Accuracy	Recall	Precision	Recall	Precision	Recall	Precision
0.807	0.269	0.627	0.212	0.610	0.963	0.824

6 Conclusion

In this experiment, the results of Relevance and Stance could not be expected to improve the score due to overlearning for data inequality, but with Fact-checkability, it is possible to expect the score improvement since the effect of the targeted approach was obtained. In particular, it was clarified that both convolution and sequence operations were necessary to estimate the fact-checkability. From the data set, it was confirmed that the sentences including the fact checkable information shared similar facts with the target sentence provided in the task. Therefore, it was shown that the convolution considering the relevance to other sentences was effective.

From the results of the formal run, we need to adjust the models of Relevance and Stance in future. We also need to adjust model parameters for Fact-checkability to improve the recall. The subject of the issue was limited in the data set, and the current model was specialized for the subject of the task in learning.

It seems that multi-task learning will be effective in order to attain more generic estimation even if provided the limited data set. In the future, while improving models and parameters, we aim to implement more general-purpose models in consideration of practical applications.

Acknowledgments This work was partially supported by JSPS Grants-in-Aid for Scientific Research (B) (#19H04420)

References

1. GUGGILLA, Chinnappa; MILLER, Tristan; GUREVYCH, Iryna. CNN-and LSTM-based Claim Classification in Online User Comments. In: Proceedings of the 26th International Conference on Computational Linguistics (COLING '16): Technical Papers. 2016. p. 2740-2751.
2. KIM, Yoon. Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP '14). 2014. 1746-1751.
3. 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.
4. KOCHKINA, Elena; LIAKATA, Maria; ZUBIAGA, Arkaitz. All-in-one: Multi-task Learning for Rumour Verification. In: Proceedings of the 27th International Conference on Computational Linguistics (COLING '18). 2018. p. 3402–3413.
5. MA, Jing; GAO, Wei; MITRA, Prasenjit; KWON, Sejeong; JANSEN, J. Bernard; WONG, Kam-Fai; CHA, Meeyoung. Detecting Rumors from Microblogs with Recurrent Neural Networks. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI '16). 2016. p. 3818-3824.
6. MIHAYLOVA, Tsvetomila; NAKOV, Preslav; MARQUEZ, Lluis; BARRON-CEDENO, Alberto; MOHTARAMI, Mitra; KARADZHOV, Georgi; GLASS, Jamese. Fact Checking in Community Forums. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI '18). 2018. p. 5309-5316.
7. MOHTARAMI, Mitra; BALY, Ramy; GLASS, Jamese; NAKOV, Preslav; MARQUEZ, Lluis; MOSCHITTI, Alessandro. Automatic Stance Detection Using End-to-End Memory Networks. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL '18) (Volume 1: Long Papers). 2018. p. 767-776.

8. MUELLER, Jonas; THYAGARAJAN, Aditya. Siamese Recurrent Architectures for Learning Sentence Similarity. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI '16). 2016, p. 2786-2792.
9. SAINATH, Tara N., et al. Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '15). IEEE, 2015. p. 4580-4584.
10. THORNE, James, et al. Fake News Detection using Stacked Ensemble of Classifiers. In: Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism. 2017. p. 80-83.
11. VOSOUGHI, Soroush; ROY, Deb; ARAL, Sinan. The Spread of True and False News Online. Science, 2018, 359.6380: 1146-1151.
12. WANG, Jin, et al. Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL '16) (Volume 2: Short Papers). 2016. p. 225-230.
13. WANG, Yequan, et al. Attention-based LSTM for Aspect-level Sentiment Classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP '16). 2016. p. 606-615.
14. ZHANG, Xiang; ZHAO, Junbo; LECUN, Yann. Character-level Convolutional Networks for Text Classification. In: Advances in Neural Information Processing Systems (NIPS '15). 2015. p. 649-657.
15. ZUBIAGA, Arkaitz, et al. Discourse-Aware Rumour Stance Classification in Social Media Using Sequential Classifiers. Information Processing & Management, 2018, 54.2: 273-290.