Extraction of Paraphrases using Time Series Deep Learning Method

Ryuichi Omi, Yoko Nishihara, *Member, IAENG,* Ryosuke Yamanishi, *Member, IAENG*

***Abstract*—We propose a new method to extract paraphrases of inappropriate expressions using long short-term memory (LSTM) as one of the time series deep learning methods. Inappropriate expressions are often described indirectly. To extract inappropriate expressions described indirectly, the meanings of expressions must be identified. The meanings of expressions may vary depending on the domain and context where the expressions are used. The proposed method uses LSTM to learn the series of responses on thread on a bulletin board system. LSTM obtains a model for detecting responses containing inappropriate expressions. When the model evaluates a response as inappropriate, the method extracts words from the response. If a word appears frequently in responses evaluated as inappropriate, the method evaluates the word as a paraphrase for an inappropriate expression. We conducted preliminary experiments. It was confirmed that the method could extract paraphrases for inappropriate expressions.**

***Index Terms*—Paraphrases for inappropriate expressions, Time series deep learning, Bulletin board system, Word vector with distributed representation**

I. Introduction

THERE are many contents on the Web that are regarded as inappropriate for young people. Social media like bulletin board systems, blogs, social networking services are used widely. Young people also use the social media for their communication. Some of the threads on the BBSs have responses that are inappropriate topics for the young people. The inappropriate contents are uploaded everywhere on the Web.

In order to prevent the young people from inappropriate information, filtering methods have been proposed and invented. Most of the filtering methods check whether inappropriate expressions are included in the information or not. If an inappropriate expression is detected, the filtering method filters the information out.

Kikuchi et al. proposed a method to detect inappropriate Web sites with high accuracy. Their method calculates the probability of inappropriateness by using the co-occurrence probability of two words. The method applies a Bayesian filter for detecting inappropriate Web sites. The Bayesian filter is used frequently for filtering spam mails [2]. Methods using Bayesian obtain conditions when information is evaluated and labeled with inappropriate. The methods learns many appropriate and inappropriate information.

Most of the filtering methods use dictionaries for inappropriate words for labeling information as inappropriate. The dictionaries contain words and expressions with inappropriate meanings. However, the meanings of words vary depending on context and domain where the words are used. There are paraphrases of inappropriate words that are not contained in the dictionaries. In order to avoid being filtered, some of the contents on the Web use indirect expressions for inappropriate words and expressions. A word “leaves” is one of the examples. Generally, the meaning of “leaves” is a flattened structure of a higher plant. However, the word “leaves” is often used as a paraphrase of cannabis. The information relating to the not-allowed drugs (at least in Japan) is not appropriate for young people. Such information should be filtered out. However, the meaning of word and expression depends on the domain and context in which it is used. In order to filter out the inappropriate expressions described indirectly, the change of the word meaning should be captured.

In this research, we focus on the meaning of a word varies depending on domain and context. We propose a new method to extract paraphrases of inappropriate words using Long short-term memory (LSTM) [5], which is one of the method of time series deep learning. For extracting inappropriate information, the recent studies often use the neural network methods [4], [8], [1].

Manuscript received December 26, 2018; revised January 8, 2019.

R.Omi, Y. Nishihara and R. Yamanishi are with Ristumeikan University. e-mail: nisihara@fc.ritsumei.ac.jp

II. Proposed method

The method takes thread data of bulletin board system as the input. Thread data contains some inappropriate expressions. Firstly, each response contained in the thread data is given its index with chronological order. Each response is parsed by a morphological analyzer and divided into words. When an obvious inappropriate word is included in a response, the response is given a label as inappropriate. Each response is transformed to a vector of words using fastText [3]. The vectors with labels are learned by LSTM and a model is obtained. If the model evaluates a response as inappropriate and the response does not have an obvious inappropriate word, the method extracts words from the response. If the extracted word appears frequently in the responses evaluated as inappropriate, the method outputs the word as a paraphrase of inappropriate expression.

A. Input: Set of thread data of bulletin board system

In this research, we use an age-restricted thread data on electronic bulletin board system. Each response on a thread is parsed by a morphological analyzer into words. Each response is checked whether it contains an obvious inappropriate word or not. The method uses a list of NG words used in a Web site of video hosting service as a set of obvious inappropriate word. If a response contains an obvious inappropriate word, the method labels the response with 1 as inappropriate. If not, the method labels the response with 0 as appropriate.

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

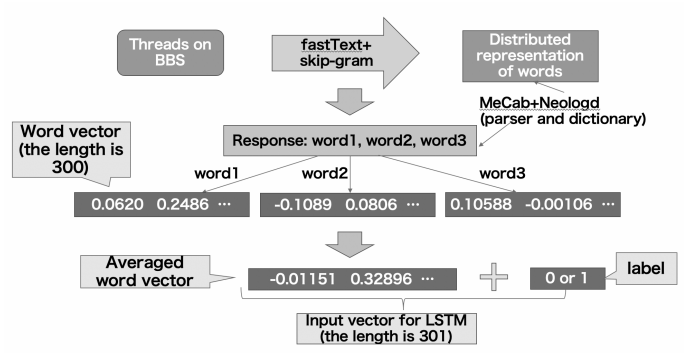

Fig. 1. Flow of transformation of a response to an averaged word vector.

B. Transformation of response to Averaged Word Vector

The method transforms each response into an averaged word vector. The method uses LSTM to learn the context of threads. LSTM accepts the same length input to learn and make a model. The number of words contained in a response is different from each other.

Fig.1 shows the flow of transformation of a response to an averaged word vector. The method obtains distributed representations of words. Each word in a response is represented as a word vector by the distributed representations. Then, the all vectors of word contained in a response are averaged and a label is added the last of a vector. The averaged word vectors are used for input to LSTM.

The method uses fastTex with skip-gram to make the distributed representations of words. In this research, we used about 9,000 threads on a bulletin board system in Japan. The threads were in age-restricted categories. In the threads, there were approximately 1,000,000 responses with obvious inappropriate words and approximately 5,600,000 responses without obvious inappropriate words.

C. Making of Time Series Model using LSTM

The system uses LSTM as a time series deep learning method to make a model. The averaged word vectors are inputted to LSTM.

D. Extraction of Paraphrases of Inappropriate Expressions

The learned model can evaluates whether each response is inappropriate or not. If the model evaluates a response as inappropriate, the model gives a high probability to the response. If a response with a high probability does not contain an obvious inappropriate word, the method collects all of the words in a response as the candidates of paraphrases of inappropriate words. If the collected word appears frequently in the responses evaluated as inappropriate, the word can be assumed as a paraphrase of an inappropriate word.

III. Preliminary Experiment

We conducted a preliminary experiment to investigate whether paraphrases of inappropriate words can be extracted or not. The experimental procedures were as follows.

Firstly, we obtained a word vector model by learning morphologically analyzed corpus. We used MeCab [7] as the morphological analyzer for parsing responses. MeCab's dictionary was NEologd [9] updated on 18th, August, 2018. The used corpus had 10,300 threads with age-restriction. The model was obtained by learning the corpus. The threads had about 1 million responses with obvious inappropriate words, about 5.6 million responses without obvious inappropriate words. LSTM made a model that can evaluates whether the sixth response was inappropriate or not when five consecutive responses were given.

In this preliminary experiment, the above process was performed on one thread, and responses with probability more than 0.4 were checked by manually whether paraphrases of inappropriate words were contained or not. The one thread had 1,000 responses.

TABLE I
Experimental Results: the numbers of evaluated responses.

		Inappropriate expression in response	
		YES	NO
output	inappropriate	64	72
	not appropriate	123	741
total		187	813

A. Discussion on Experimental Results

Table I shows the numbers of responses. The results showed that LSTM evaluated correctly for 64 inappropriate responses and 741 appropriate responses. The accuracy was 80.5%. The results indicated that LSTM could learn a model for the thread.

Note, 72 responses without obvious inappropriate words were evaluated as inappropriate. We checked the 72 responses manually and found that 22 responses contained the paraphrases for inappropriate words. Through the preliminary experiments, it was confirmed that the paraphrases for inappropriate words could be extracted by the proposed method.

IV. Conclusion and Future Works

In this paper, we proposed an extraction method for paraphrases for inappropriate words using time series deep learning method. Preliminary experiments were carried out in this paper. We prepared 10,300 threads on a bulletin board system. The threads were in categories with age-restricted. The time series of responses was learned by LSTM, and a model was constructed.

In the preliminary experiments, we investigated whether the proposed method extracts paraphrases for inappropriate words or not. We tested 1,000 responses, and obtained 72 responses that might contain paraphrases for inappropriate words. We found that 22 responses contained paraphrases for inappropriate words.

We will conduct an evaluation experiment on all threads and evaluate the usefulness of the proposed method in the future.

References

[1] Betty van Aken, Julian Risch, Ralf Krestel, Alexander Loser: Challenges for Toxic Comment Classification: An In-Depth Error Analysis, 2nd Workshop on Abusive Language Online to be held at EMNLP 2018, (2018).

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[2] Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Georgios Sakkis, Constantine D. Spyropoulos, Panagiotis Stamatopoulos: Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach, Proceedings of the workshop "Machine Learning and Textual Information Access," 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 1-13, (2000),
[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Trans. of the Association for Computational Linguistics, Vol.5, pp.135-146, (2017).
[4] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, Vassilis P. Plagianakos: Convolutional Neural Networks for Toxic Comment Classification, SETN '18 Proceedings of the 10th Hellenic Conference on Artificial Intelligence, (2018).
[5] Sepp Hochreiter, and Jurgen Schmidhuber. Long short-term memory, Neural Computation, Vol.9, No.8, pp.1735-1780, (1997).
[6] Takuya Kikuchi, Akira Utsumi: Harmful site filtering method based on cooccurrence information of words, Information Processing Society of Japan, Vol.9, No.6, pp.1-6, (2013).
[7] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In The Proc. of EMNLP-2004, pp. 230-237, 2004.
[8] Julian Risch and Ralf Krestel: Aggression Identification Using Deep Learning and Data Augmentation, Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pp.150-158, (2018).
[9] https://github.com/neologd/mecab-unidic-neologd/ (accessed on 25th, December, 2018)

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)