LETTER

Speech-Act Classification Using a Convolutional Neural Network Based on POS tag and Dependency-Relation Bigram Embedding

Donghyun Yoo[†a)] **, Youngjoong Ko**[††b)] **and Jungyun Seo**[†c)]

SUMMARY In this paper, we propose a deep learning based model for classifying speech-acts using a convolutional neural network (CNN). The model uses some bigram features including parts-of-speech (POS) tags and dependency-relation bigrams, which represent syntactic structural information in utterances. Previous classification approaches using CNN have commonly exploited word embeddings using morpheme unigrams. However, the proposed model first extracts two different bigram features that well reflect the syntactic structure of utterances and then represents them as a vector representation using a word embedding technique. As a result, the proposed model using bigram embeddings achieves an accuracy of 89.05%. Furthermore, the accuracy of this model is relatively 2.8% higher than that of competitive models in previous studies.

key words: *Speech-Act Classification, Dependency-Relation, Convolutional Neural Network, Word Embedding, Bigram Embedding.*

1. Introduction

A speech-act is a linguistic action intended by a speaker, and speech-act classification refers to the analysis of the intention of the speaker in an utterance [1]. In particular, a speech-act classification must be highly accurate because it is the first step in processing an utterance in a dialogue system. However, it is difficult to perfectly interpret a speaker's intention because a speaker indicates his/her intention using various expressions [2]. In order to overcome this problem, most studies adopt supervised machine learning methods using various speech features including lexical, contextual, and other semantic features [3-5]. Recently, many deep learning techniques have been applied to natural language processing because they afford a significant advantage with regard to analyzing the semantics of lexical features based on a distributed representation. Thus, in this study, we attempt to improve the accuracy of speech-act classification through the use of better distributed representations for various features.

In this paper, we propose a new speech-act classification model using a convolutional neural network (CNN) based on bigram embedding. Most deep learning techniques use a distributed representation of features, called word embedding. Word embeddings are distributed representations of words that group words that share semantic or syntactic properties. CNNs have traditionally used only morpheme unigram embeddings, whereas the proposed model additionally uses bigram embeddings created by various bigram features such as morpheme, parts-of-speech (POS) tags, and dependency-relation bigrams. In particular, a model with POS tags and dependency-relation bigrams achieves the highest accuracy because it can represent structural information of utterances.

The remainder of this paper is organized as follows. The next section summarizes related work. In Section 3, we explain our proposed method in detail. In Section 4, experimental results and comparisons between the proposed method and previous ones are presented. The paper is finally concluded in Section 5.

2. Related Work

Most previous studies for speech-act classification have used various supervised machine-learning methods to automatically identify a speaker's intention. Lee [3] designed a speech-act classification method using the hidden Markov model (HMM) to estimate speech-act probabilities and this method was improved upon by using smoothed class probabilities from a decision tree. Choi [4] proposed a maximum entropy model (MEM) to determine the speech-act of current utterances using previous utterances as contextual information. Song [5] recommended a support vector machine (SVM) model to preferentially analyze classes with lower distribution when training among a set of classes.

Since deep learning techniques have recently shown better performances in many fields such as image classification [6], many natural language researchers have applied deep learning techniques using word embeddings to many natural language processing problems [7,8]. Collobert [7] constructed a model for various natural language processing techniques. Kim [8] presented a convolutional neural network model for sentence classification. These studies used word embeddings with a morpheme unit for the convolutional neural network, but we utilize bigram embeddings of various bigram features such as POS tag bigram and dependency-relation bigram, which contain syntactic structural information within utterances.

† The author is with #908 R-hall, Sogang University, Sinsu-dong, Mapo-gu, 121-742, Republic of Korea.
†† Dong-A University, 840, Hadan 2-dong, Saha-gu, Busan, 604-714, Republic of Korea.
a) E-mail: babuluve@gmail.com
b) E-mail: youngjoong.ko@gmail.com (Corresponding author)
c) E-mail: seoyj@sogang.ac.kr

3. Proposed Model

Fig. 1 outlines the proposed model, which consists of three steps. The first step is the extraction of five types of features: previous utterance speech-act, morpheme unigram, morpheme bigram, POS tag bigram, and dependency-relation bigram. In the next step, the extracted features are converted into a distributed representation using a word embedding technique during pre-training. Finally, CNN is trained with an effective combination of the extracted features.

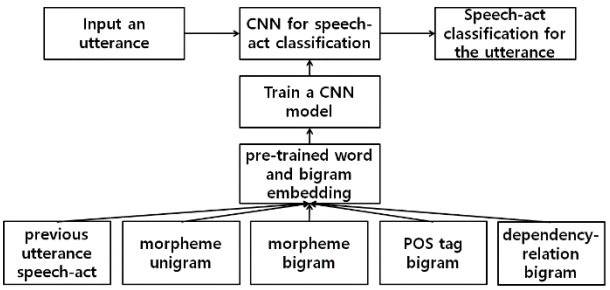

Fig. 1. Overview of speech-act classification using unigram and bigram embedding

3.1 *Word Embedding Technique*

Word embeddings express the latent values of features as distributed representations. They are trained using the neural network language model [9]. It is also possible to express words that are represented by large data in a real-valued vector form. Mikolov [10] proposed a neural network language model to remove the hidden layer for word embeddings: the continuous bag-of-words (CBOW) and Skip-gram model. CBOW predicts the current word based on the context, and Skip-gram specifies the word based on the context. In particular, the removal of the hidden layer results in faster training speed compared to other models.

3.2 *Convolutional Neural Network (CNN)*

A convolutional neural network (CNN) is a deep learning technique. In particular, many studies have proved the effectiveness of CNN for natural language processing. Kim proposed the CNN architecture for sentence classification in [8]. Here, we use Kim's CNN architecture for speech-act classification. However, unlike Kim's model, we applied bigram embeddings of various features into the CNN architecture to improve the performance of speech-act classification.

The CNN architecture has various layers. In the first layer, extracted features are transformed into a distributed presentation using pre-trained word embeddings. A convolution layer produces new features using a filter that applies to each window of words on a current word to produce a feature map. For the max pooling layer, the maximum value is selected from each feature map to capture the most important feature. The proposed model uses multiple filters to obtain multiple features. Finally, the max pooling layer is fully connected to the softmax layer whose output is the probability distribution over classes.

3.3 *Feature Extraction for Speech-act Classification*

In previous studies, several kinds of features have been used for speech-act classification, such as lexical features (morpheme unigrams and POS tag bigrams) and a contextual feature (speech-act of previous utterance) [1,11]. However, we believe that there exist more effective features, which can be used to further improve speech-act classification using feature distributed representations in deep learning. Thus, different features are considered in this research to reflect the syntactic information of utterances, such as morpheme and dependency-relation bigrams. Consequently, five types of features are used in speech-act classification: the speech-act of previous utterance, morpheme unigram, morpheme bigram, POS tag bigram, and dependency-relation bigram.

In previous studies, the speech-act of previous utterances is considered important as the speech-act of the current utterance highly depends on that of the previous utterance as contextual information. In addition to the above-mentioned lexical and contextual features, the proposed model attempts to utilize the syntactic information of utterances based on POS tags and dependency-relation bigrams for speech-act classification. In actual, a POS tag bigram provides a pattern of consecutive POS tags in the syntactic structure of an utterance and a dependency-relation bigram affords the same for syntactic information between constituents, such as subjects, objects, and verbs, in whole utterances; for example, consider the input utterance is "네, 이름을 알려주시고 얼마 동안 머무르실 건지 가르쳐 주십시오. (Yes, please tell me your name and how long you intend to stay here)", the analysis result from the dependency parser is shown in Fig. 2. They represent relationships between constituents and these relationships are called dependency-relations. Choi [12]'s dependency parser is used in the paper. An arc represents relationships between two constituents analyzed by a dependency parser. Boxes in the first layer contain morphemes analyzed by a POS tagger and boxes in the second layer contain relationship names from the dependency parser.

However, we do not use all the head-dependent relationships in the result from the dependency parser. Relation ① is excluded because it includes unnecessary syntactic information like interjections. Further, we use other relations to learn word embeddings. These head-dependent relationships are extracted as dependency-relation bigrams as shown in Fig. 2 and 3.

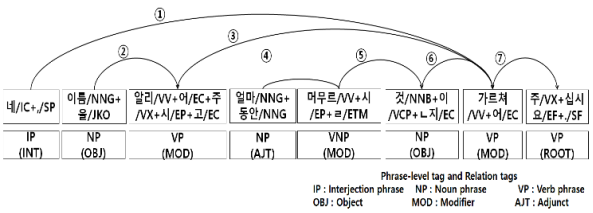

Fig. 2. Analysis result of a dependency parser

relation ② : 이름을/NNG+을/JKO__알리/VV+어/EC+주/VX+시/EP+고/EC
your name__tell me
relation ③ : 알리/VV+어/EC+주/VX+시/EP+고/EC__가르쳐/VV+어/EC
tell me __ teach
relation ④ : 얼마/NNG+동안/NNG__머무르/VV+시/EP+ㄹ/ETM+것/NNB+이/VCP+ㄴ지/EC how long __stay
relation ⑤ : 머무르/VV+시/EP+ㄹ/ETM__것/NNB+이/JKS
stay__here
relation ⑥ : 것/NNB+이/VCP+ㄴ지/EC__가르쳐/VV+어/EC
here__teach
relation ⑦ : 가르쳐/VV+어/EC__주/VX+십시요/EF+./SF
teach__please

Fig. 3. Example of extracted dependency-relation bigrams

All the above-mentioned four types of features must be converted into vector representations to be used in CNN. These four features including morpheme unigram, morpheme bigram, POS tag bigram, and dependency-relation bigram are represented as vectors by a word embedding technique, whereas the speech-act of previous utterances is initialized randomly as the same size of the vector as that of the other three types of features.

Table 1. Example of extracted bigram features

Example	네, 이름을 알려주시고 얼마 동안 머무르실 건지 가르쳐 주십시오. (Yes, please tell me your name and how long stay here)
Morpheme unigram	네/IC, 이름/NNG, 을/JKO, …
Morpheme bigram	네/IC_이름/NNG, 이름/NNG__을/JKO, 을/JKO__알리/VV, …
POS tag bigram	IC__NNG, NNG__JKO, JKO__VV, …
Dependency -relation bigram	이름을/NNG+을/JKO___알리/VV+어/EC+주/VX+시/EP+고/EC, 알리/VV+어/EC+주/VX+시/EP+고/EC___가르쳐/VV+어/EC, …

Table 1 shows the four types of unigram and bigram features used in this research. Features are extracted through the following steps. First, morpheme unigram features are created by a lexical/POS_tag after a POS tagger analyzes the input utterance; morpheme bigrams are extracted using the sliding window technique with a window size of 2. Then, the input utterance annotated by the POS tagger is syntactically analyzed by a dependency parser and the dependency-relation bigrams are extracted by the processes explained in Fig. 2 and 3. Finally, the extracted unigram or bigram features are converted into distributed representations as vectors using the embedding technique of CBOW and they are used as inputs of the input layer of CNNs. The corpus for feature embedding is composed of articles from KBS News[a)] and its size is approximately 3G bytes.

4. Experiments

In this section, we compare the proposed model with those proposed in previous studies over several experiments to demonstrate the effectiveness of the proposed model.

4.1 *Experimental Settings*

To evaluate the proposed model, we used the reservation tasks corpus that was used in the previous studies [3, 4, 5, 11] which was transcribed from real conversions occurring when making hotel, airline, and tour reservation. That corpus consists of 528 dialogues, 10,285 utterances, and 17 types of speech-acts. The corpus was divided into training (428 dialogues and 8,349 utterances) and testing data (100 dialogues and 1,936 utterances). Since the corpus has no standard development set, 20% of the training data was randomly selected as a development set. The development set is used for tuning the parameters of CNNs.

In order to evaluate speech-act classification, we used the accuracy measure in Eq. (1).

$$\text{Accucary} = \frac{Number\ of\ correct\ sentences}{Number\ of\ test\ sentences} \qquad (1)$$

We exploited the architecture of Kim [8]'s model using rectified linear units (ReLU) as an activation function. The window sizes of filters are 3, 4, and 5 with 100 feature maps. Dropout rate is 0.5, and mini-batch size is 50 with the Adadelta update rule [13]. The dimension of the embeddings is set to 64.

4.2 *Experimental results*

Table 2 shows the final experimental results for combinations of various features with SVM and CNN.

Table 2. Experiment results of combination of various unigram and bigram features

	SVM	CNN
previous utterance speech-act + morpheme unigram	85.05	88.23

[a)] http://news.kbs.co.kr/

previous utterance speech-act + morpheme unigram + morpheme bigram	84.65	87.29
previous utterance speech-act + morpheme unigram + POS tag bigram	**85.95***	88.35
previous utterance speech-act + morpheme unigram + dependency-relation bigram	85.89	88.11
previous utterance speech-act + morpheme unigram + POS tag bigram +dependency-relation bigram	85.79	**89.05***

* A significant test is conducted by one-tailed McNemar's test and the improvement of the proposed CNN is statistically significant, p<0.01 (p=0.000102).

Overall, the performance of CNN is much better than that of SVM in all the combinations. For SVM, the best performance is the combination of morpheme unigram and POS tag bigram, but other SVMs with dependency-relation bigrams also achieve very similar performances. However, the best among CNNs is from the combination of morpheme unigram, POS tag bigram, and dependency-relation bigram. The relative difference of the best performance of SVM and CNN is 3.6%.

The combination of morpheme unigram and the speech act of previous utterances is considered as a baseline model; its performance was fairly high. As three different bigrams are added to the baseline model, their effectiveness is evaluated through experiments. The highest accuracy was achieved when the baseline model was combined with POS tag and dependency-relation bigrams. The accuracy is 0.93% relatively higher than that of the baseline model. While dependency-relation bigrams contain syntactic structure information within the utterance, they do not contain information from the sequence of words. Thus, further improvements in the combination model can be realized by adding POS tag bigrams into dependency-relation bigrams.

4.3 *Comparison of the proposed model and previous studies*

Table 3. Comparisons with previous studies

	Accuracy (%)
Lee's HMM [3]	81.50
Choi' MEM [4]	83.57
Kim's SVM [12]	86.62
Song's SVM [5]	86.54
Proposed model	**89.05**

As shown in Table. 3, Kim's [11] model achieved the highest accuracy in previous models. However, our proposed model achieved 2.8% relatively higher performance than that of Kim's model.

5. Conclusions

In this study, we presented a CNN model for speech-act classification using POS tag and dependency-relation bigrams. The proposed CNN model using POS tag and dependency-relation bigrams obtained the highest accuracy of 89.05%. We think that these features well reflect the syntactic information of utterances to speech-act classification.

In the future, we will continue to study how to reflect other information of utterances to realize higher performances in speech-act classification.

References

[1] S. Kim, Y. Lee, and J. Lee. 2008. Korean Speech Act Tagging using Previous Sentence Features and Following Candidate Speech Acts. *Journal of KISS : Software and Applications*, Vol. 35, No. 6, pp. 374-385.

[2] C. Doran, J. Aberdeen, L. Damianos and L. Hirschman, 2003, "Comparing Several Aspects of Human-Computer and Human-Human Dialogues," *Current and New Directions in Discourse and Dialogue*, pp. 133-159.

[3] S. Lee and J. Seo, 2005, "A Korean Speech Act Analysis System Using Hidden Markov Model with Decision Trees," *International Journal of Computer Processing of Oriental Languages*, vol.15, pp. 231-243.

[4] W. Choi, H. Kim and J. Seo, 2005, "An Integrated Dialogue Analysis Model for Determining Speech Acts and Discourse Structures," *IEICE Transactions on Information and Systems*, vol.E88-D, No.1, pp.150-157.

[5] N. Song, K. Bae and Y. Ko, 2016, "Effective Korean Speech-act Classification Using the Classification Priority Apllication and a Post correction Rules," *Journal of KIISE*, Vol. 43, No. 1, pp. 80-86.

[6] A. Krizhevsky, S. Ilya and G. E. Hinton, 2012, "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. pp. 1097-1105.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, 2011, "Natural language processing (almost) from scratch," *The Journal of Machine Learning Research*, Vol. 12, pp. 2493-2537.

[8] Y. Kim, 2014, "Convolutional neural networks for sentence classification," In *Proceedings of EMNLP*. pp. 1746-1751.

[9] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, 2003, "A neural probabilistic language model," *Journal of Machine Learning Research*, pp. 1137-1155.

[10] T. Mikolov, W. T. Yih, and G. Zweig, 2013, "Linguistic Regularities in Continuous Space Word Representations," In *HLT-NAACL*, vol.13, pp. 746-751.

[11] K. Kim and J. Seo, 2003, "Decision of the Korean Speech Act using Feature Selection Method," *Journal of KISS : Software and Applications*, Vol. 30, No. 3·4 , pp. 278-284.

[12] D. Choi, J. Park, and K. Choi, 2012, "Korean treebank transformation for parser training," *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, vol.29, no.6, pp. 78-88.

[13] M. D. Zeiler, 2012, "ADADELTA: an adaptive learning rate method," *arXiv preprint* arXiv:1212.5701.