Using Convolutional Neural Network with BERT for Intent Determination

Changai He,Sibao Chen*
Key Lab of IC&SP of MOE, School of Computer Science and Technology, Anhui University, Hefei 230601, Anhui, China
PKU Shenzhen Institute, Shenzhen, China. changai.he@imsl.org.cn

Shilei Huang
Shenzhen Raisound Technologies, Co., Ltd
Shenzhen, China
shilei.huang@imsl.org.cn

Jian Zhang
PKU Shenzhen Institute
Shenzhen, China
jian.zhang@imsl.org.cn

Xiao Song*
PKU Shenzhen Institute
Shenzhen, China
xiao.song@imsl.org.cn

***Abstract*—We propose an Intent Determination (ID) method by combining the single-layer Convolutional Neural Network (CNN) with the Bidirectional Encoder Representations from Transformers (BERT). The ID task is usually treated as a classification issue and the user's query statement is usually of short text type. It has been proven that CNN is suitable for conducting short text classification tasks. We utilize BERT as a sentence encoder, which can accurately get the context representation of a sentence. Our method improves the performance of ID with the powerful ability to capture semantic and long-distance dependencies in sentences. Our experimental results demonstrate that our model outperforms the state-of-the-art approach and improves the accuracy of 0.67% on the ATIS dataset. On the ground truth of the Chinese dataset, as the intent granularity increases, our method improves the accuracy by 15.99%, 4.75%, 4.69%, 6.29%, and 4.12% compared to the baseline.**

***Keywords*-CNN; BERT; ID; context representation.**

I. Introduction

Spoken dialogue system, voice assistant, automatic customer service, etc. are the hot spots of current natural language processing research [1]. The success of these applications depends not only on speech recognition but also on text understanding. In human-machine, Spoken Language Understanding (SLU) aims to automatically determinate the intent of the user as expressed in natural language. As an important part of the SLU system, ID is an accurate understanding of the user's intents.

Generally speaking, the SLU system first needs to transcribe user's voices into text via automatic speech recognition (ASR), or the input of users is text typed [2], then determine the user's intents and combine the corresponding constraints, and finally these information can be delivered to the dialogue or task management system to satisfy the special needs of the user.

An example utterance sentence is shown in Table I, which uses the In/Out/Begin (IOB) and intent label representation. The intent label is used for the ID task, the slot label is used for the slot filling (SF) task. ID and SF are two major tasks in SLU. The sample is taken from the airline travel information system (ATIS) corpus [3], which is widely used in the SLU domain.

Table I
ATIS utterance example with intent and slot annotation

Sentence	Slot label	Intent label
show	O	
flights	O	
from	O	
boston	B-dept	flight
to	O	
new	B-arr	
york	I-arr	
today	B-date	

The ID task is usually treated as a classification issue and the SF task as a sequence labeling issue [2]. Applying the information from one task to another can promote each other and achieve joint prediction. We will take more expense and effort to mark intent and slot than to mark intent only when we prepare training data for training a new SLU system. In practice, SF does not bring significant improvement in the ID task. We focus on only utilize the intent labels to improve the ID ability in this work.

In this work, we divide the ID tasks into two main stages, one is the text representation, another one is text classification.

1) text representation: In recent years, the unsupervised pre-training models have achieved excellent performance in many natural language tasks, such as Word2Vec, ELMo, BERT [4]. These models can be used to conduct the work of the text representation stage, among them the semantic information learned by BERT is more accurate and complete. BERT consists of multi-layer transformers, each of the transformer is composed of a self-attention sub-layer with multiple attention heads.

2) text classification: In the text classification stage, the text classification methods based on deep learning mainly include model based on CNN and RNN or their improved versions. Compared with the classification method based on traditional machine learning, the text classification method based on deep learning does not need to extract the key features of corpus text manually. The user's query statement is usually a short text type. In short

978-1-7281-5014-7/19/$31.00

text classification tasks, CNN is better than RNN [5]. In the ID task, the accuracy of the LSTM model is 1.48% higher than that of the RNN model [6], and GRU is an improvement of the LSTM model.

So we try to combine CNN, LSTM, and GRU with BERT respectively to deal with the ID tasks, where only the intent labels are used. We observe the performance of these models in conducting ID task on the ATIS corpus (English dataset) and the Chinese dataset. The experimental results demonstrate that the method of the CNN combined with BERT is superior to other methods. To our best knowledge, this work is the first time to utilize the CNN combined with BERT to conduct the issue of ID under the condition only the intent labels are utilized, and our model outperforms the state-of-the-art approaches on English dataset and achieves excellent performance compared to the baseline on the Chinese dataset.

II. Related Work

Many researchers have performed research on ID task, the main directions of current research are divided into methods based on rule templates [7] and methods based on classification models [8], [9], [6], [10].

ID based on rule templates generally requires artificial construction of rule templates and category information to classify user intent. Different expressions lead to an increase in the number of rule templates, which requires a lot of manpower and resources. Although the method based on rule templates without a large amount of the training data, which can achieve very high accuracy, it cannot solve the high-cost issue of reconstructing the template when the intent text is changed.

The purpose of ID is to determinate the intent of a sentence, which can be regarded as a standard classification task [11]. The determination of the user's intent based on the classification model is mainly divided into the traditional machine learning methods and deep learning methods. The classification methods based on the traditional machine learning [10] [12] [13] [14], it is necessary to manually extract the key features of the corpus text, such as word features, n-gram, etc., and then implement the intent classification by training the intent classifier. This method is not only costly but also does not accurately understand the deep semantic information of the user.

In recent years, with the continuous development of deep learning, scholars have been explored CNN[15][16], recurrent neural networks (RNN)[6][17], long short-term memory (LSTM)[18][19], etc. to conduct the ID task. Sarikaya et al. used deep belief networks (DBNs) to handle routing classification problems [20]. Tur et al. proposed a sentence simplification method to deal with the issue of increased error rates caused by more complex, longer and more natural sentences [1]. As for joint work on ID and SF, Xu et al. proposed a neural network (NN) version of the triangular CRF (TriCRF) model, which utilized CNN to extract features and shared by two tasks [9]. Guo et al. found that recursive neural networks (RecNNs) have well performance in joint work of ID and SF [8]. Zhang et al. utilized GRU to learn the representation of each time step to predict the label for each slot and used the max-pooling layer to capture the global features of sentences to perform the ID task [2]. Liu et al. proposed using the RNNs framework for performing joint work of ID, SF, and language modeling (LM) task on ATIS corpus [11]. But these models can not accurately capture the context information of sentences.

In this paper, we propose using CNN built on top of BERT for the ID task. BERT has been proven effective for learning contextualized word representations, and CNN is suitable for conducting short text classification tasks. We verify the effectiveness of our proposed method on the ATIS dataset compared with other methods, and the experimental results on the Chinese dataset show our method can achieve excellence performance.

III. Methodology

A. Task Description

In the previous work [8], [9], [11], most of them regard the ID task as a classification task, and we follow this idea in this work. Before training the ID model, we need to divide the intent of the sentence corpus, for example, I want to query the weather of Shenzhen tomorrow, we can mark this sentence with a query_weather label. After the marking is completed, the training, development, and test sets are divided according to a certain ratio. Finally, the data is fed to the network framework shown in Figure 1 to train the ID model. In detail, the ID task can be defined as a given marker dataset $S = (X, Y)$, where $x_i \in X$ is the utterance and $y_i \in Y$ is its intent label, and the ID task is an attempt to correctly associate the utterance x_i with the label y_i.

B. Framework Description

The overall framework of the proposed method in Figure 1. $x_1, x_2, ..., x_n$ represents a sentence of length n. The BERT part consists of 12 layers of the transformer. The transformer is composed of 12 different attention heads, each of which can focus on different types of component combinations. Input representation can unambiguously represent a text sentence in one token, then get the context word embedding through the transformer encoder. We utilize d to denote the dimension of the word embedding. If the sentence length is n, then the dimension of the sentence matrix is $n \times d$, we use the same zero-padding strategy as in [21]. The CNN part consists of three filter region sizes: 2, 3 and 4, with a total of 128 filters. The CNN part first performs convolutions on the sentence matrix generate feature maps, then 1-max pooling layer records the largest number from each feature map, a single variable feature vector is generated from all 128 maps, and the 128 features are connected to form a feature vector of the penultimate layer. The last softmax layer receives this feature vector as input and uses it to determinate the intent [22]. Here we assume $d = 5$, $n = 7$.

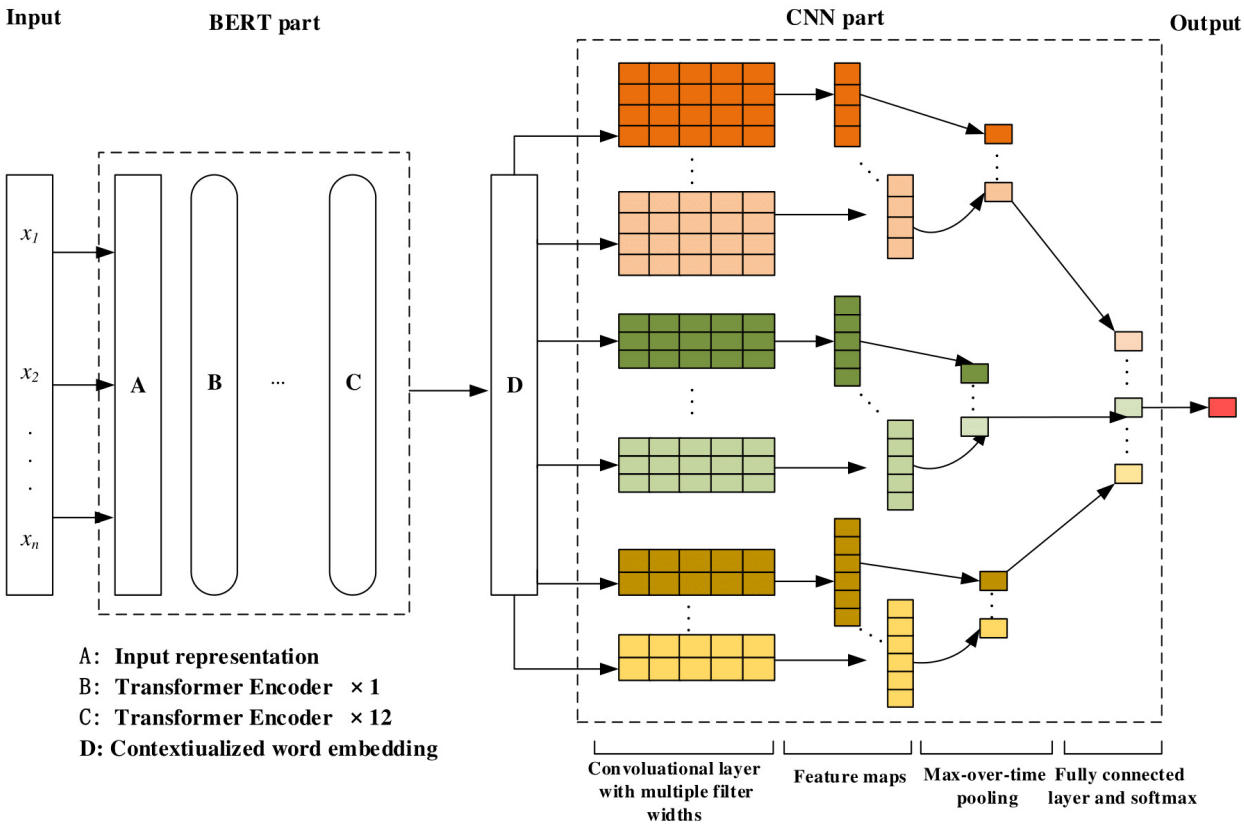

Figure 1. The framework of our method.

C. Our Method

In this section, we will describe our approach in detail to predict the user's intent. In order to conduct this issue, Liu et al. utilized the RNNs framework conduct joint work of ID, SF, and LM, which highlight is real-time parsing. This framework can provide an optimal intent, slot resolution and a prediction of a word for the moment T input to the present.

In fact, although slot filling brings some benefits to the ID task, it does not significantly improve the effect of ID and requires more expense and effort to mark the slot label. In view of the powerful semantic sentence understanding of BERT, which has been proven effective for learning contextualized word representations, and the user's query statement is usually of short text type. It has been proven that CNN is suitable for conducting short text classification tasks. We propose CNN built on top of BERT to conduct the issue of ID. As we all know that CNN congenital convolution operations are not suitable for sequence-level text, but the sentence-level word vector obtained by BERT has just made up for this shortcoming. BERT can get the context representation of a sentence, and the CNN can get the feature maps from the contextualized word embedding, and the prediction labels are given through the softmax layer.

1) The BERT part: Recently, researchers from the Google AI language team opened the source code of the BERT project[1] , a library for pre-training language representation. BERT has the ability to capture semantics and long-distance dependencies of sentences, and BERT enhances the generalization ability of the word vector model, so which can accurately capture the context information of sentences compared to previous pre-training models. The BERT part consists of 12-layer transformers and input representation, each of the transformer is composed of a self-attention sub-layer with multiple attention heads. Compared to the deep learning strategy previously ID, we utilize BERT to obtain the context representation of the sentence, which catch the context representation of the sentence easier and more accurately.

2) The CNN part: The CNN model has made remarkable achievements in computer vision and speech recognition, and it can also play a very important role in natural language processing. In a way, word is to text what pixel is to image, different from CNN in image recognition, the width of convolution kernel here is fixed. Parameters of each convolution kernel (filter) are Shared, which means that a filter can only identify the same type of features. Convolution has the function of local feature extraction, so CNN can be used to extract the key information similar to n-gram in sentences.

The feature map generated after convolution is no longer a matrix, but a column vector with a width of 1. Then

[1]https://github.com/google-research/bert

Table II
ID ON THE YTBD DATASET OF DIFFERENT GRANULARITIES

Model	ID (acc)				
	10-intents	20-intents	30-intents	40-intents	50-intents
CNN	82.89%	85.21%	83.94%	81.04%	80.81%
BERT-LSTM	93.88%	89.37%	88.31%	86.51%	83.57%
BERT-GRU	93.16%	89.96%	87.68%	86.10%	83.57%
BERT-CNN(ours)	**93.88%**	**89.96%**	**88.63%**	**87.33%**	**84.93%**

1-max pooling is adopted in the pooling layer, that is, each filter gets the maximum value in the column vector. After the pooling layer is a fully connected layer, the operation of the hidden layer connecting the output layer, the softmax function is used to output the distribution probabilities of different intents. At the same time, a dropout operation is added to prevent overfitting. The key of short text classification is to accurately extract the main idea of sentences, and the way to extract the main idea of sentences is to extract the keywords of sentences as features. The convolution and pooling process of CNN is a feature extract process, so CNN is more suitable for conducting short text classification task.

IV. EXPERIMENTS

In our experiments, we used the independent training RNN intent model as a baseline on the ATIS corpus. Liu et al. joint online spoken language understanding and language modeling with RNN achieved the state-of-the-art performance in ID task on the ATIS corpus, at the same time, which provided the result of the independent training RNN intent model [11].

The user's query statement is usually a short text type. In short text classification tasks, CNN is better than RNN [5]. In the ID task, the accuracy of the LSTM model is 1.48% higher than that of the RNN model [6], and GRU is an improvement of the LSTM model. So we try to combine CNN, LSTM, and GRU with BERT respectively to deal with the ID tasks, where only the intent labels are used. At the same time, we do the experiment CNN without the BERT model on ATIS corpus, then we also utilize the above model to do experiments on the Chinese dataset, which consists of questions from Yuetongbao customer service platform[2] , which is a public platform with 70 million users for Electronic Toll Collection (ETC) in China. we refer to the Chinese dataset as Yuetongbao Dataset (YTBD).

A. Dataset

1) ATIS: The ATIS corpus is widely used in SLU research. In this work, we followed the same ATIS corpus setup used in [3]. The ATIS contains 18 different intent labels and 5,781 sentences, there are 4,978 sentences for training, 893 sentences for testing, and then take 893 out of the training set as the development set. The average length of each piece of data is approximately 15 words. We report the results of 10-fold cross-validation.

[2]https://www.96533.com/

2) YTBD: There is no public dataset on the Chinese intent determination task, so we collected the questions on the Yuetongbao customer service platform and manual marked them. Most of the styles of these questions are colloquial. After screening, we manual marked 84,107 training data and 968 test data for intent. The average length of each piece of data is approximately 15 words. There are 50 kinds of intent labels in this dataset. We divide them into five data sets. The criteria for dividing are: first, take all the data of the first ten classes in the training set, and then take all the data of the first twenty classes in the training set, and so on. Do the same on the test set. Finally, we will get 5 training sets and 5 test sets with 10 intents, 20 intents, ..., 50 intents, and We take 15 percent of each training sets as development sets. The purpose of this is to observe if CNN with BERT framework also has an advantage in a fine-grained ID task.

Table III
ID ON THE ATIS DATASET

Model	ID (acc)
CNN	74.79%
online+RNN-LU (independent ID)	97.87%
online+RNN-LU (intent+slot)	98.43%
BERT-LSTM	96.31%
BERT-GRU	96.86%
BERT-CNN (ours)	**98.54%**

B. Evaluation Metrics

Our experiments were evaluated using accuracy, the metric accuracy for ID can be computed as:

$$acc = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

TP is the number of positive classes predicted as positive. TN is the number of negative classes predicted as negative. FN is the number of positive classes predicted as negative. FP is the number of negative classes predicted as positive. Some sentences of the ATIS dataset contain multiple intents, and our process method is that so long as the identified intents belong to one of the multiple intents, they are counted as correct identification.

C. Results Analysis

In all the experiments shown in Table II and Table III, we set the max length of the input sentence to 32, and input 32 data per batch, we automatically stop

training, when the accuracy of more than 1000 steps is not improved.

Table II shows the accuracy scores of our method and the CNN model without BERT, the BERT-based LSTM model, the BERT-based GRU model on the YTBD dataset of different granularities for conducting the ID task. Our method achieves excellent performance. Comparing with the CNN model without BERT, our method improves the accuracy of ID at each granularity. Comparing with the BERT-LSTM model and the BERT-GRU model, our method achieves the best performance at each granularity, but the ability to determinate intent not improves much. Table III shows the accuracy scores of our method and the CNN model without BERT, online+RNN-LU (independent ID) model: independent training RNN intent model, online+RNN-LU (intent+slot) model: joint model with recurrent intent + slot label context, the BERT-based LSTM model, the BERT-based GRU model on the ATIS dataset for conducting the ID task. Our method is superior to the latest technology in ID. Comparing with the CNN model, the accuracy is improved by more than 23%. Comparing with the online+RNN-LU (independent ID) model, it illustrates the effectiveness of our method in dealing with the ID task, our method only improved by 0.67%, mainly because the baseline score is relatively high. Comparing with the online+RNN-LU (intent+slot) model, it illustrates that our method achieves a comparable result. Comparing with the BERT-LSTM model and the BERT-GRU model, it verifies that the CNN model is more suitable for dealing with short text classification issues.

V. Conclusion

In this work, we proposed a method by combining BERT with CNN to conduct the ID task. The result on the ATIS corpus shown that our model outperforms the state-of-the-art approaches. The results on the Chinese dataset shown that our model improves the accuracy by 15.99%, 4.75%, 4.69%, 6.29%, and 4.12% compare to the baseline as the intent granularity increases, and it also shows that even if only the intent label is used, the ID model with superior performance can be trained, which will help us save expense and effort in the process of conduct ID task.

We will apply the ID model trained on the YTBD dataset to an actual project, and plan to improve the robustness of the ID model in the future.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China under Grant 61976004, and in part by Basic Research in Shenzhen (Discipline Layout) under Grant JCYJ20170817155939233.

References

[1] G. Tur, D. Hakkani-Tur, L. Heck, and S. Parthasarathy, "Sentence simplification for spoken language understanding," in *IEEE International Conference on Acoustics*, 2011.

[2] X. Zhang and H. Wang, "A joint model of intent determination and slot filling for spoken language understanding." in *IJCAI*, 2016, pp. 2993–2999.

[3] G. Tur, D. Hakkani-Tur, and L. Heck, "What is left to be understood in atis?" in *Spoken Language Technology Workshop*, 2011.

[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[5] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of cnn and rnn for natural language processing," *arXiv preprint arXiv:1702.01923*, 2017.

[6] S. V. Ravuri and A. Stolcke, "Recurrent neural network and lstm models for lexical utterance classification." 2015.

[7] J. Prager, D. Radev, E. Brown, A. Coden, and V. Samn, "The use of predictive annotation for question answering in trec8," 1999.

[8] D. Z. Guo, G. Tur, W. T. Yih, and G. Zweig, "Joint semantic utterance classification and slot filling with recursive neural networks," in *Spoken Language Technology Workshop*, 2015.

[9] P. Xu and R. Sarikaya, "Convolutional neural network based triangular crf for joint intent detection and slot filling," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 78–83.

[10] P. Haffner, G. Tur, and J. H. Wright, "Optimizing svms for complex call classification," in *IEEE International Conference on Acoustics*, 2003.

[11] L. Bing and I. Lane, "Joint online spoken language understanding and language modeling with recurrent neural networks," 2016.

[12] A. SCHAPIRE and Y. Singer, "A boosting-based system for text classification," *Machine Learning*, vol. 39, no. 1, p. 2, 2000.

[13] K. M. Schneider, "A comparison of event models for naive bayes anti-spam e-mail filtering," in *Tenth Conference on European Chapter of the Association for Computational Linguistics*, 2003.

[14] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.

[15] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[16] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. R. Mohamed, G. Dahl, and B. Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39–48, 2015.

[17] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," *Statistics*, pp. 285–290, 2015.

[18] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and F. F. Li, "Every moment counts: Dense detailed labeling of actions in complex videos," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 375–389, 2018.

[19] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5142–5154, 2016.

[20] R. Sarikaya, G. E. Hinton, and B. Ramabhadran, "Deep belief nets for natural language call-routing," in *IEEE International Conference on Acoustics*, 2011.

[21] Y. Kim, "Convolutional neural networks for sentence classification," *Eprint Arxiv*, 2014.

[22] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," *arXiv preprint arXiv:1510.03820*, 2015.