The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/2514-9288.htm

ID-SF-Fusion: a cooperative model of intent detection and slot filling for natural language understanding

Meng Zhu and Xiaolong Xu
Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing, China

Received 22 March 2023
Revised 24 December 2023
Accepted 24 December 2023

Abstract

Purpose – Intent detection (ID) and slot filling (SF) are two important tasks in natural language understanding. ID is to identify the main intent of a paragraph of text. The goal of SF is to extract the information that is important to the intent from the input sentence. However, most of the existing methods use sentence-level intention recognition, which has the risk of error propagation, and the relationship between intention recognition and SF is not explicitly modeled. Aiming at this problem, this paper proposes a collaborative model of ID and SF for intelligent spoken language understanding called ID-SF-Fusion.

Design/methodology/approach – ID-SF-Fusion uses Bidirectional Encoder Representation from Transformers (BERT) and Bidirectional Long Short-Term Memory (BiLSTM) to extract effective word embedding and context vectors containing the whole sentence information respectively. Fusion layer is used to provide intent–slot fusion information for SF task. In this way, the relationship between ID and SF task is fully explicitly modeled. This layer takes the result of ID and slot context vectors as input to obtain the fusion information which contains both ID result and slot information. Meanwhile, to further reduce error propagation, we use word-level ID for the ID-SF-Fusion model. Finally, two tasks of ID and SF are realized by joint optimization training.

Findings – We conducted experiments on two public datasets, Airline Travel Information Systems (ATIS) and Snips. The results show that the Intent ACC score and Slot F1 score of ID-SF-Fusion on ATIS and Snips are 98.0 per cent and 95.8 per cent, respectively, and the two indicators on Snips dataset are 98.6 per cent and 96.7 per cent, respectively. These models are superior to slot-gated, SF-ID NetWork, stack-Prop and other models. In addition, ablation experiments were performed to further analyze and discuss the proposed model.

Originality/value – This paper uses word-level intent recognition and introduces intent information into the SF process, which is a significant improvement on both data sets.

Keywords Intent detection, Slot filling, BERT, Natural language understanding, BiLSTM, Word-level

Paper type Research paper

1. Introduction

It has always been an important research goal in the field of artificial intelligence to build an intelligent dialogue system (Valizadeh and Parde, 2022) that can understand human language, analyze the intents of sentences and give smooth and correct responses. A typical dialogue system is usually composed of the following three modules (Deriu *et al.*, 2021): natural language understanding (NLU), dialogue management and natural language generation. Among them, NLU (Samant *et al.*, 2022) is a core part of dialogue system. NLU aims at extracting the intent information and the corresponding semantic slot information from the input sentence to form a semantic framework. NLU consists of two tasks: intent detection (ID) and semantic slot filling (SF). ID is to identify the main intent of a paragraph of text (Samant *et al.*, 2021), which is akin to finding the central idea, while the goal of SF is to extract the statement information important to intent of the input sentence, which is the important slot information (Liu *et al.*, 2022). By completing the two tasks of ID and semantic SF in the NLU module, the system can identify the types of tasks that users

Data Technologies and Applications

2514-9288
DOI 10.1108/DTA-03-2023-0088

want the system to complete and the important semantic parameters to complete these tasks, and then form a structured semantic framework representation. For example, for the sentence "show flight from Beijing to New York today," the semantic framework representation obtained through the NLU module is shown in Table I.

Generally, ID and semantic SF are regarded as two independent tasks and modeled separately. ID is regarded as a text classification task (Zhang, 2021), while SF is usually regarded as a sequence labeling problem (Ye *et al.*, 2022). Some scholars (Goo *et al.*, 2018) pointed out that the two tasks influence and restrict each other. In order to make full use of the correlation of these two tasks, joint recognition models (Hakkani-Tür *et al.*, 2016; Liu and Lane, 2016; Chen *et al.*, 2019) for ID and semantic SF are proposed. These models improve the performance of spoken language understanding through the interaction between the two tasks. In order to capture long-distance dependencies in sentences, the attention mechanism (Vaswani *et al.*, 2017) was introduced into the model (Liu and Lane, 2016). However, these models (Hakkani-Tür *et al.*, 2016; Liu and Lane, 2016; Chen *et al.*, 2019; Wang *et al.*, 2018) only implicitly model the dependency relationship between the two tasks by sharing the underlying parameters and joint loss function. In recent years, some scholars began to study and model the explicit relationship between ID and semantic SF. Goo *et al.*, 2018 and Li *et al.*, 2018 respectively proposed a gating mechanism to introduce intent information into SF and explicitly modeled the relationship between the two tasks. However, the gating mechanism cannot fully summarize and remember the intent information, and because the two tasks of intent identification and SF use implicit vector interaction information, the interpretability of the model is reduced. Qin *et al.* (2019) proposed a joint model based on stack propagation and word-level ID, using the stack propagation mechanism to directly introduce intent information into the SF task to assist the completion of the SF task. However, the two unidirectional Long short-term memory (LSTM) structure used by the stack model makes the latter classification result depend on the classification result of the previous unit, limiting the parallel processing ability of the model. Fan *et al.* (2022) pointed out that slot words are of great help to the results of semantic SF, and proposed a joint recognition model with slot-related intents. Before the SF task, a slot word recognition layer was added, which was used to judge whether it is a slot or not, and the result of slot word recognition was introduced into the task of SF. However, the superstructure of this method still uses the gating mechanism, which cannot fully summarize and remember the intention information. Chen *et al.* (2019) proposed joint recognition model of ID and SF based on Bidirectional Encoder Representation from Transformers (BERT) (Devlin *et al.*, 2019), which made use of the powerful coding ability of BERT pretrained model, but did not model the relationship between the two tasks, and did not make full use of the association between the two tasks.

In this paper, we proposed a new model named ID-SF-Fusion to address the above problems. We conducted experiments on two public datasets, Snips and Airline Travel Information Systems (ATIS), and compared our model with the models in Goo *et al.*, 2018, Hakkani-Tür *et al.* (2016), Liu and Lane (2016), Chen *et al.* (2019), Li *et al.*, 2018, Qin *et al.* (2019), Haihong *et al.*, 2019, and Zhang *et al.* (2019). The results of experiments show that the

Table I.
Example of semantic framework

Sentence	show	flight	from	Beijing	to	New	York	today
Slot Label	O	O	O	B-dept	O	B-arr	I-arr	B-date
Intent				find flight				

Source: Table by authors

proposed model achieves good results in ID and SF. In addition, we performed ablation experiments to explore our model. In general, the contributions of this paper are as follows:

(1) Inspired by Qin *et al.* (2019), based on BERT, this paper proposes to use a word-level ID module. On the one hand, word-level ID can improve the effect of ID by voting the final intent. On the other hand, word-level ID can retain more useful intent information for SF. If the intent of some words in a sentence is incorrectly predicted, the other correct words will still contribute to the corresponding slot prediction.

(2) We proposed intent–slot information fusion layer for our model. In this paper, we focus on the interaction between intent information and slot information. For the problem of how to apply intent information to SF, we proposed a fusion layer to model the relationship between the two tasks, which outputs intent–slot fusion information by using similarity to perform the SF module.

2. Related Work

2.1 Independent modeling

The earliest approaches treat the two tasks as two independent modules and complete them in pipelined method. Rule-based methods (Chen *et al.*, 2017) were initially applied to these two tasks. The advantages of rule-based methods are highly stable and fast, and they can well extract the required intent and slot information. However, these methods rely on manual extraction by experts, which not only consume time but also human resources. In addition, due to the differences in scenarios, the rules formulated are not universal, resulting in the inability of the system to migrate from one scenario to another.

Later, machine learning and deep learning methods were applied to the tasks. Common methods for ID include machine learning methods such as support vector machine (Joachims, 1998), Naive Bayes (McCallum *et al.*, 1998), and deep learning methods such as LSTM (Fang, 2016). SF is often treated as a sequence labeling problem and common sequence labeling formats include BIO and BIOS. Commonly used machine learning methods for SF include hidden Markov model, conditional random field (CRF) and so on. With the rise of deep learning, Peng and Yao (2015) and Mesnil *et al.* (2015) used recurrent neural network (RNN) to complete SF task. Later, Yao *et al.* (2014) used LSTM network to accomplish SF.

2.2 Joint modeling

Some researchers have pointed out that the joint modeling of the ID and SF can achieve better results than the independent modeling. At present, the joint recognition method of ID and SF has become the mainstream method. Joint modeling is not only simple at structure but also can get better F1 score. The joint modeling method based on neural network structure maintains the technical advantages of deep learning, does not require manual definition of features, and only need word vectors and character vectors to reach the standard level. There are two main forms of joint recognition models. The first is to let the two tasks share the underlying feature data, and then output corresponding results for the two tasks respectively. The weighted sum of the loss functions of the two tasks is the loss function of the model. The second is to combine the two tasks into a single task and solve for the joint probability of ID and SF.

Zhang and Wang (2016) proposed to use RNN to extract the underlying shared features, and then to identify the intent and semantic slots respectively, which achieved good results. However, RNN adopts a linear sequence structure to continuously input information from front to back, which is not good at capturing the long-term dependence relationship in the text. And to some extent, there is the problem of gradient disappearance or gradient explosion. Attention mechanism has been widely used in the field of natural language processing since it was proposed in 2014, because it can learn deep representation features and capture long-distance dependencies in sequences. Liu and Lane (2016) proposed to add alignment information and attention mechanism into bidirectional RNN to further improve the performance of NLU.

Further, Zhang *et al.* (2019) argued that the method based on RNN does not consider the hierarchical relationship between slot and intent, and it is risky to condense all information into a vector. What's more, SF based on word level can provide clues for sentence-level ID, and the result of ID can also help the idea of SF. In 2018, they proposed a joint recognition model of ID and SF based on capsule neural network, and completed ID and SF based on three-layer capsule network model and dynamic routing algorithm.

In 2018, Goo *et al.*, 2018 proposed that ID and SF are interdependent. The results of ID have an important impact on SF task, and the results of SF can also affect ID. They employ gating mechanism to monitor SF with the result of ID and improve SF performance. In the same year, Li *et al.*, 2018 also proposed another gating mechanism, proposing a new gating mechanism model with self-attention to make full use of the semantic association between slots and intents. The model obtains a neural network intent-augmented embedding based on a self-attention mechanism. Through end-to-end joint learning, the objectives of both tasks are optimized simultaneously.

Qin *et al.* (2019) pointed out two shortcomings on model using gating mechanism in 2019, and proposed a stack frame that combines token-level ID mechanism. The stack propagation framework is used to directly take the result of ID as the input of the SF task, which directly uses the intent information to predict the slot result. In addition, Qin *et al.* performed ID for each individual word, while the intent of the whole sentence was voted on by the intent result of each word. If a word-level intent prediction is wrong, other predicted words can still give the correct guide to the slot of the corresponding word.

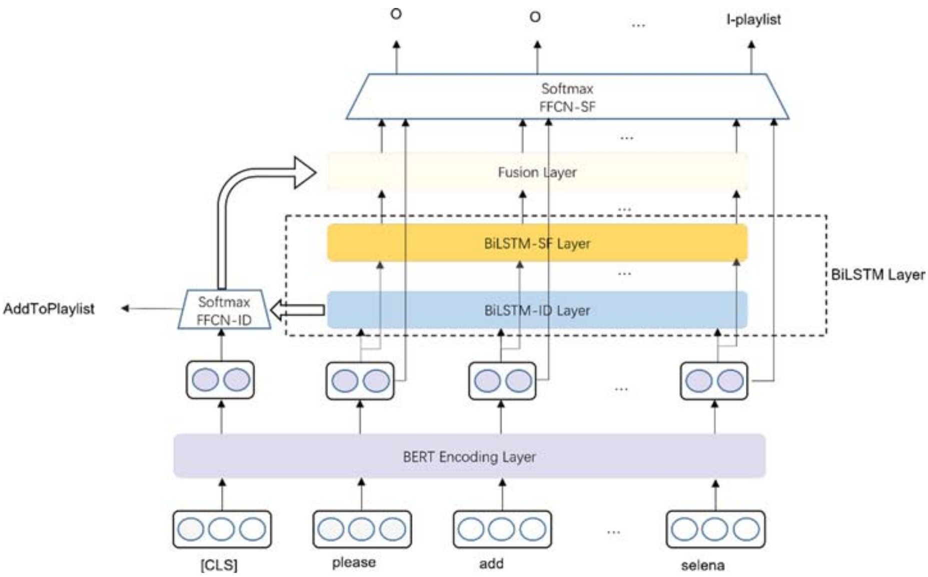
Chen *et al.* (2019) of Alibaba proposed joint recognition model for ID and SF based on BERT model. BERT can dynamically generate semantic representations of words according to different context, which can represent sentence features better than traditional word embedding, and achieve better results in the joint recognition task of intent recognition and SF. However, this method directly sent output of BERT to the intent classifier and the slot classifier. The two tasks simply share the coding layer, resulting no relationship between the two tasks are learned.

Fan *et al.* (2022) improved the model proposed by Goo *et al.* by adding a slot recognition layer that distinguishes slot words from general words. But the superstructure still uses Goo's gating mechanism, which does not adequately summarize and remember intent information. Han *et al.* (2021) proposed a bidirectional joint model for ID and SF, which achieved mutual performance improvement between ID and SF through BERT and bidirectional joint NLU mechanism Intent2slot and slot2Intent. However, in essence, this method is still a shared coding layer, and the two tasks are processed separately based on the coding layer, which does not make full use of the effect of the result of ID on the SF task.

In this paper, we focus on the idea of ID based on word level and propose an intent–slot information fusion layer to generate intent–slot information, which model the relationship between the two tasks, and finally complete the tasks of ID and SF.

3. Model

In this section, we describe the proposed joint model of ID and SF. The structure of the proposed model is shown in Figure 1. In the coding layer, the BERT model is adopted to encode, because the BERT model can be used to obtain word feature vectors containing rich information for the downstream modules. We believe that only relying on the location information in the BERT model is not enough to characterize the dependence relationship



Source: Figure by authors

Figure 1. The framework of ID-SF-Fusion

between word sequences, so we set up a Bidirectional Long Short-Term Memory (BiLSTM) layer, which uses two independent BiLSTM structures, namely BILSTM-ID and BILSTM-SF, to extract the intent context vector and slot context vector respectively. Then, we concatenate the feature vectors of [CLS] placeholders with the context vectors extracted by BiLSTM-ID, respectively, and transmit the obtained results to softmax function for word-level ID. The sentence-level intent result is determined by voting on the ID result at the word level, and then the intent result of the whole sentence is obtained. In addition, we send the word-level intent results output by the softmax layer to the intent–slot information fusion layer to obtain the intent–slot fusion information, and concatenate the results of this layer with the word feature vectors of the coding layer. Then we take the results as the input of the fully connection. Finally, a softmax layer is added to classify slot labels to obtain the results of slot labels coding layer.

The input data for the ID and SF are user statements in the form of text sentences. These sentences are typically labeled as a sequence of words that is the same length as the number of words in the input sentence. The output of SF is a slot label sequence of the same length as the input sequence, and the result of ID is a single label. Different slot labels and intent labels are mapped to integers for numerical representation. Let n be the length of the longest input sentence in the dataset and extend all sentences to the length of n by filling with the filler "[PAD]." And finally the word sequence $\omega_1, \omega_2, \ldots \omega_n$, is formed. Then, we insert a special placeholder "[CLS]" at the beginning of the sequence to obtain a sequence of length $n + 1$ $\omega_0, \omega_1, \ldots \omega_n$, where ω_0 stands for "[CLS]." We use BERT pretrained model as the underlying encoding layer to obtain the word feature vectors, and the feature vectors of the ith word ω_i is $c_i \in R^{d_e}$:

$$c_i = PretrainedLM(\omega_{0:n}, i) \tag{1}$$

where d_e represents the size of the word embedding vectors.

3.1 BiLSTM layer

Based on the feature that BiLSTM can extract temporal features, this paper adopts two BiLSTM structures to extract intent context vector information and slot context vector information respectively.

BiLSTM-ID sublayer: The output of the encoding layer $c = (c_1, c_2, \ldots, c_n)$ is used as the input of this layer, and the BiLSTM reads the sequence forward and backward. The forward LSTM reads the word sequence in the original order and generates a hidden state vector fh_i^{ID} at each time step. Similarly, the reversed LSTM generates a sequence of hidden state vectors $(bh_n^{\mathrm{ID}}, bh_{n-1}^{\mathrm{ID}}, \ldots, bh_1^{\mathrm{ID}})$. The intent context vector h_i^{ID} of each final step is concatenated from the forward and backward state vectors:

$$h_i^{\mathrm{ID}} = [fh_i^{\mathrm{ID}}, bh_{n-i}^{\mathrm{ID}}] \tag{2}$$

In this way, the intent context vector at each time step contains the information of the whole sentence.

BiLSTM-SF sublayer: The structure is akin to the BiLSTM-ID sublayer above. Their difference is that the BiLSTM-ID sublayer captures intent information, while the BiLSTM-SF sublayer captures slot information. This layer also takes the encoding layer output $c = (c_1, c_2, \ldots c_n)$ as input, using bidirectional LSTM to process the sequence forward and backward. The forward LSTM processes the word sequence in the original order and generates a hidden state vector fh_i^{SF} at each time step. The reversed LSTM generates hidden state vectors sequence $(fh_n^{\mathrm{SF}}, fh_{n-1}^{\mathrm{SF}}, \ldots fh_1^{\mathrm{SF}})$. The slot context vector h_i^{SF} at each final step is concatenated by the forward and backward state vectors:

$$h_i^{\mathrm{SF}} = [fh_i^{\mathrm{SF}}, fh_{n-i}^{\mathrm{SF}}] \tag{3}$$

Similarly, the slot context vector at each step contains information about the entire sentence.

3.2 Word-level ID layer

In the architecture of our proposed model, we use the idea of a word-level ID, which provides word-level intent features for the input sequence of words. Based on this idea, ID at the word level can be temporarily viewed as a sequence labeling problem, with the input word sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ and the output intent sequence $O = (o_1^{\mathrm{ID}}, o_2^{\mathrm{ID}}, \ldots, o_n^{\mathrm{ID}})$. Finally, according to the principle of "the minority obeys to the majority," the final intent label of the sentence is obtained by all word-level intent label.

The first special start character "[CLS]" encoded by BERT previously contains the information of the whole sentence. What's more, a BiLSTM-ID layer is adopted to extract the temporal features between sentences to obtain more intent information. In order to make full use of the sentence information contained in the feature vector c_0 corresponding to the beginning character "[CLS]" and the intent context vector h_i^{ID} corresponding to the words at each time step, we get the intent information vector I_i^{ID} at word level by concatenating the two:

$$I_i^{\mathrm{ID}} = [c_0, h_i^{\mathrm{ID}}] \tag{4}$$

After that we adopted a common classification task strategy. We input the obtained word-level intention information vector into a feedforward neural network. Then we use a softmax layer for intention prediction, and use the softmax layer for word-level intention classification.

Finally, the word-level intention classification probability y_i^{ID} is obtained. Further, we can get the word intention classification label o_i^{ID}:

$$y_i^{\mathrm{ID}} = softmax(W^{\mathrm{ID}} \cdot I_i^{\mathrm{ID}} + b^{\mathrm{ID}}) \tag{5}$$

$$o_i^{\mathrm{ID}} = argmax(y_i^{\mathrm{ID}}) \tag{6}$$

where y_i^{ID} is the intent result of the ith word; W^{ID} is the weight matrix of the feedforward neural network, and b^{ID} is its bias; o_i^{ID} is the result of the ID of the ith word.

Finally, the sentence-level ID result O^{ID} is voted on by all word-level ID results:

$$O^{\mathrm{ID}} = argmax \sum_{i=1}^{n} \sum_{j=1}^{n_I} \alpha_j 1[o_i^{\mathrm{ID}} = j] \tag{7}$$

where n is the length of the sentence, n_I is the number of intent labels; 1 is the indicator function; α_j is a 0–1 vector $\alpha \in R^{n_I}$; the value of ith element is 1, and all the others are 0; *argmax* indicates the operation that returns the index with the highest value in α.

There are two main advantages to using word-level ID:

(1) In our proposed framework, ID at the word level can provide more features for SF, which can simplify error propagation and provide more useful information for SF tasks. Compared with word-level ID, if the intent prediction result of the whole sentence is wrong, the wrong intent may have a negative impact on all slots. However, in word-level ID, if the intent prediction of some steps in the sentence is misaligned, other correct word-level intent information can still act on the SF task.

(2) Since the feature vectors corresponding to bidirectional LSTM and "[CLS]" are concatenated, each time step can grasp the context information of the whole sentence, and we can regard each time step prediction in a sentence as an individual prediction of utterance intent. Therefore, this method can reduce the prediction variance and improve the performance of ID. The experiment verifies the effectiveness of word-level ID.

3.3 Intent–slot information fusion layer

We propose a fusion mechanism to fuse the intent information and the slot context information to obtain intent–slot fusion vector. Specifically, this layer receives the slot context vector from the BiLSTM-SF layer and the intent vector from the word-level ID layer, and unifies the dimension size between them by multiplying their respective weight matrices. Then, a consistency weight is calculated by the cosine similarity of the obtained results:

$$\beta_i = cosine\_similarity(w \cdot y_i^{\mathrm{ID}}, v \cdot h_i^{\mathrm{SF}}) \tag{8}$$

Among them, *cosine_similarity* is a function to calculate the degree of familiarity between vectors; w and v are the corresponding weight matrices of y_i^{ID} and h_i^{SF}, respectively; β_i is the cosine similarity of intent information and semantic slot information at the ith time step. β_i can be regarded as the degree of consistency between y_i^{ID} and h_i^{SF}. The larger β_i is, the more intent information and semantic slot information focus on the same part of the input sequence, which means the correlation between intent and slot are stronger, and the context vector is more reliable to share the prediction results.

Finally, we multiply the slot context vector h_i^{SF} and the intent context vector y_i^{ID} by the calculated weights, respectively, and concatenate the results to obtain the intent–slot fusion vector:

$$e_i = [\beta_i \cdot y_i^{ID}, \beta_i \cdot h_i^{SF}] \tag{9}$$

where e_i is the intent–slot fusion vector at the ith time step, which will be used as the input of the intent-recognition layer.

3.4 SF layer

The SF layer defines the SF as a BIO-based sequence labeling task. First, a sequence of vectors $I = (I_1^{SF}, I_2^{SF}, \ldots, I_n^{SF})$ is created, where each I_i^{SF} is composed of the concatenation of the intent–slot fusion vector e_i and the feature vector c_i output by BERT coding layer:

$$I_i^{SF} = [e_i, c_i] \tag{10}$$

Then, each vector I_i^{SF} is passed to the feedforward neural network, and a softmax layer is used to obtain the distribution of the predicted intent labels. Finally, the index number of the slot label with the largest probability value is calculated by *argmax* function:

$$y_i^{SF} = softmax(W^{SF} \cdot I_i^{SF} + b^{SF}) \tag{11}$$

$$o_i^{SF} = argmax(y_i^{SF}) \tag{12}$$

where y_i^{SF} is the slot output of the ith word; W^{SF} is the weight matrix of the feedforward neural network, and b^{SF} is its bias; o_i^{SF} is the result of the ID of the ith word.

Thus, the prediction result of the SF task of the whole sentence is $o^{SF} = (o_1^{SF}, o_2^{SF}, \ldots, o_n^{SF})$.

3.5 Joint training

Unlike most existing joint recognition models, our framework adopts word-level intent recognition, where we transform sentence-level classification tasks into word-level predictions and use words in the intent–slot fusion layer. level of intent information, so the loss function for intent recognition can be expressed as

$$L_1 \triangleq -\sum_{j=1}^{m}\sum_{i=1}^{n} \hat{y}_{j,i}^{ID} \log(y_{j,i}^{ID}) \tag{13}$$

Similarly, the loss function for SF can be expressed as

$$L_2 \triangleq -\sum_{j=1}^{m}\sum_{i=1}^{n} \hat{y}_{j,i}^{SF} \log(y_{j,i}^{SF}) \tag{14}$$

where $\hat{y}_{j,i}^{ID}$ and $\hat{y}_{j,i}^{SF}$ are the correct intent label and the correct slot label; n is the number of words in a sentence, and m is the number of all sentences.

Finally, the joint loss L of the whole model is the sum of the loss L_1 of the intent identification task and the loss L_2 of the SF task:

$$L = L_1 + L_2 \tag{15}$$

4. Datasets and Experimental Setup

4.1 Datasets

To evaluate our model, we conduct experiments on two publicly benchmark datasets ATIS [1] and Snips[2].

ATIS: This dataset is organized and constructed by US Defense Advanced Research Projects Agency based on Airline booking Information data. The data in ATIS are all related to Airline booking. As shown in Table II, ATIS is composed of 4,478 sentences of training text, 500 sentences of verification text and 893 sentences of test text, with an average of almost 15 words per sentence. ATIS has 120 kinds of slot labels and 21 kinds of intent labels, and the size of the glossary is composed of 722 words.

Snips: Snips is a spoken language comprehension dataset obtained from Snips personal voice assistant. Snips is derived from the information collected by voice assistant. So compared with ATIS, Snips has a wider range. As shown in Table II, Snips consists of 13,084 training data texts, 700 verification texts and 700 test texts. The dataset has 72 kinds of slot labels and 7 kinds of intent labels. Compared with ATIS dataset, the vocabulary size is larger, consisting of 11,241 words.

In this paper, the two datasets, ATIS and Snips, are used to assess the performance of the proposed model. Since the two datasets are used by many scientific research experiments, there are reliable, which provides great help to evaluate the performance of the experiments.

4.2 Evaluation metrics

The evaluation metrics we used are Intent ACC, Slot F1 and Sentence ACC.

1. Intent ACC

For the task of ID, we always treat it as a classification task. In the classification task, there are only correct and incorrect classification results for a sentence. So, we pay more attention to its accuracy. Intent ACC can be defined as the percentage of the samples with correct ID in the total samples. The calculation method is shown in equation (16)

$$Intent \quad ACC = \frac{correct\_num}{total\_num} * 100\% \tag{16}$$

where *correct_num* is the number of samples with correct prediction results, and *total_num* is the number of all samples.

	ATIS	Snips
Vocabulary	722	11,241
Train	4,478	13,084
Dev	500	700
Test	893	700
Number of slot labels	120	72
Number of intent labels	21	7

Source: Table by authors

Table II. Information of ATIS and Snips datasets

2. Slot F1

For the SF task, we used F1 value as its evaluation index, and the calculation formula is as follows:

$$p = \frac{TP}{TP + FP} * 100\% \tag{17}$$

$$R = \frac{TP}{TP + FN} * 100\% \tag{18}$$

$$F1 = \frac{2PR}{P + R} * 100\% \tag{19}$$

where *TP* is the number of correct answers predicted, *FP* is the number of incorrect predictions of other classes as this class and *FN* is the number of entities of this class predicted as other classes.

3. Sentence ACC

Sentence ACC can be defined as the percentage of the total number of samples whose ID and SF are all correct in the sample. The calculation formula is as follows:

$$Sentence \quad ACC = \frac{both\_correct}{total\_num} * 100\% \tag{20}$$

where *both_correct* is the number of samples that are all correct for ID and SF.

4.3 Baseline

We compare our model with existing baseline joint models for ID and SF, which include

Joint Seq (Hakkani-Tür *et al.*, 2016): A multitask modeling method proposed by Hakkani-Tur *et al.* in 2016, using BiLSTM to construct a sequence-based Joint model, and RNN-LSTM architecture is proposed. By establishing a Joint multi-domain model, joint modeling of SF and ID is realized;

Attention-Based (Liu and Lane, 2016): An attention-based model proposed by Liu and Lane in 2016, which uses the attention mechanism to let the network learn the relationship between intent and slot labels. Combined with the encoder-decoder model, a joint model of intent prediction and SF is proposed;

Slot-Gated (Goo *et al.*, 2018): A model based on gating mechanism proposed by Goo *et al.* in 2018, which aims to better explore the correlation between SF and ID. They simulate the dependence of slot on intent by introducing slot gating mechanism, and use intent context vector to model the slot-intent relationship;

Self-Attentive (Li *et al.*, 2018): A novel self-attention model based on intent enhancement gate mechanism proposed by Li *et al.* in 2018. This model first obtains neural network intent enhancement embedding based on self-attention mechanism, and then uses intent semantic representation as the door of the marker slot label, and uses the semantic association between the slot and the intent to achieve the completion of the intent recognition and SF tasks;

SF-ID NetWork (Haihong *et al.*, 2019): A bidirectional correlation ID and SF joint recognition model proposed by Haihong *et al.* in 2019 introduces SF-ID networks to establish a connection between ID and SF by directly connecting the SF subnet and the

ID subnet, and designs an iterative mechanism to enhance the two-way connection, where the SF-ID netWork (SF-first) model executes the SF subnet first and the SF-ID netWork (ID-first) model executes the ID subnet first;

CAPSULE-NLU (Zhang *et al.*, 2019): A joint recognition model for ID and SF based on capsule neural network proposed by Zhang *et al.* in 2018 completes ID and SF by the three-layer capsule network model and dynamic routing algorithm. This model realizes the synergistic effect of word-level SF and sentence-level intent recognition;

Joint BERT (Chen *et al.*, 2019): A Joint recognition model of intent and slot based on BERT proposed by Chen *et al.* This model uses the special label [CLS] added at the beginning for ID. SF not only uses the results of Transformer output directly for sequence annotation, but also adds CRF layer for global optimization of slot labels.

Stack-Prop (Qin *et al.*, 2019): A stack propagation framework proposed by Qin *et al.* in 2019 combines word-level intent recognition mechanisms. This model utilizes the stack propagation framework to directly exploit the results of intent recognition as input to the SF task, which directly uses intent information to predict slot results.

4.4 Experimental settings

The BERT pretraining model we use is the "bert-base-uncased" which is trained by the Wikipedia and BookcCorpus. Wikipedia owns words of magnitude 2500M, the model cased the following tasks. The word magnitude of BookCorpus is 800M, and the rich dataset ensures the reliability of the model output. The "bert-base-uncased" pretraining model has 12 hidden layers, outputs of 768-dimensional tensor, 12 self-attention heads, and a total of 110 M parameters. All hyperparameters are fine-tuned during training. We set the maximum sentence length to 50 and the batch-size size to 32. We used Adam optimizer to optimize the model, and conducted experiments with Adam initial learning rate of {1e-5, 2e5, 3e-5, 4e-5, 5e-5} and Dropout to 0.1. We trained 40 epochs on each dataset and calculated the ID accuracy, the F1 value of SF, and the sentence semantic recognition accuracy after each epoch on the validation set. We chose the model corresponding to the checkpoint that worked best on the validation to apply to the test.

4.5 Experimental results and analysis

Table III shows the specific performance of our proposed model versus the baseline. As shown in Table III, our model shows the best performance on both datasets compared to the baseline model.

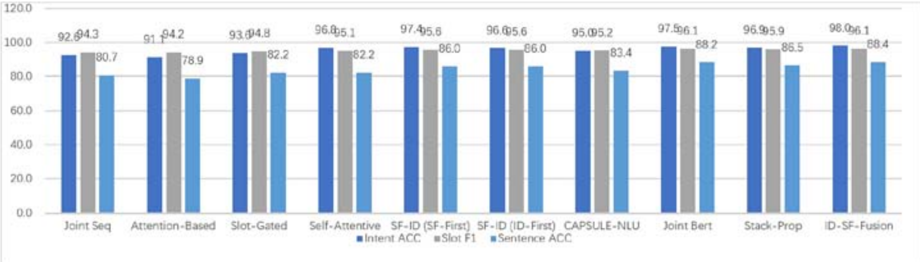
As shown in Figure 2(a), our model outperforms all baseline models on the dataset ATIS. Compared with Stack-Prop, the state-of-the art model that does not use BERT pretrained model as the underlying coding layer, the Intent ACC, Slot F1 and Sentence ACC of our model on ATIS dataset increased by 1.1 per cent, 0.2 per cent and 1.9 per cent, respectively. We also find that our proposed model outperforms the Joint BERT model. As shown in Figure 2(a), compared with the Joint BERT model, the Intent ACC and Sentence ACC of our model on ATIS increased by 0.5 per cent and 0.2 per cent, respectively. On another dataset, Snips, our model also shows advanced performance. As shown in Figure 2(b), similarly to ATIS dataset, the performance of our model on Snips dataset is improved by 0.6 per cent, 2.5 per cent and 6 per cent respectively compared with Stack-Prop model in three indicators, and our model sentence ACC is improved by 0.1 per cent compared with the model using BERT.

For ATIS datasets, the intent label distribution is unbalanced, from 3,309 samples at the highest frequency to 1 sample at the lowest frequency. This can negatively affect the overall performance of ID. In addition, the number of slots in ATIS is significantly more than the number of that in SNIPS, and these slots cover only one domain, resulting in a sea of slots being shared by multiple intents. For example, flight intents in ATIS contain 66 slot types,

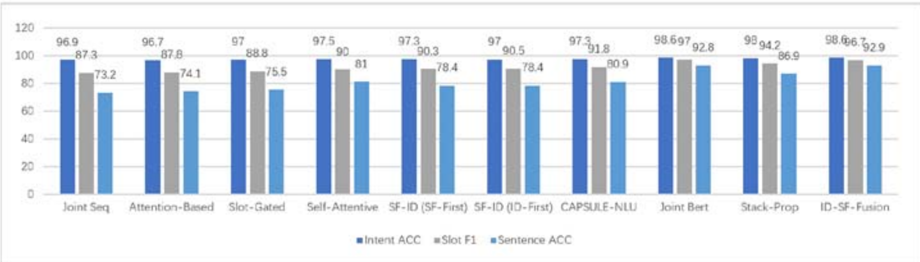
Table III. Experimental results and comparison of models (in %)

Models	ATIS Intent ACC	ATIS Slot F1	ATIS Sentence ACC	Snips Intent ACC	Snips Slot F1	Snips Sentence ACC
Joint Seq (Hakkani-Tür *et al.*, 2016)	92.6	94.3	80.7	96.9	87.3	73.2
Attention-Based (Liu and Lane, 2016)	91.1	94.2	78.9	96.7	87.8	74.1
Slot-Gated (Goo *et al.*, 2018)	93.6	94.8	82.2	97.0	88.8	75.5
Self-Attentive (Li *et al.*, 2018)	96.8	95.1	82.2	97.5	90.0	81.0
SF-ID (SF-First) (Haihong *et al.*, 2019)	97.4	95.6	86.0	97.3	90.3	78.4
SF-ID (ID-First) (Haihong *et al.*, 2019)	96.6	95.6	86.0	97.0	90.5	78.4
CAPSULE-NLU (Zhang *et al.*, 2019)	95.0	95.2	83.4	97.3	91.8	80.9
Joint BERT (Chen *et al.*, 2019)	97.5	96.1	88.2	98.6	97.0	92.8
Stack-Prop (Qin *et al.*, 2019)	96.9	95.9	86.5	98.0	94.2	86.9
Our model	98.0	96.1	88.4	98.6	96.7	92.9

Source: Table by authors



(a) Results on ATIS



(b) Results on Snips

Figure 2. Experimental results

Source: Figure by authors

of which 52 slots are repeated in other intents. In our proposed model, the validity of contextual cues provided by intents may be undermined by overlapping slots between different intents. However, the results show that the performance of SF can still be improved by exploiting the intents corresponding to these specific slots.

Compared to ATIS, the cross-domain topics of Snips are more complex. For example, the intents "BookRestaurant," "GetWeather" and "PlayMusic" are from different areas. Therefore, the vocabulary of Snips dictionary is much larger than ATIS's. The lexical diversity in Snips provides useful information for ID, and its intent distribution in the training dataset is well balanced. These factors explain why the ID accuracy on Snips dataset is relatively higher than that on ATIS dataset. In addition, Snips has relatively few overlapping slots (only 11 slots are shared with other intents, compared to ATIS's 79 slots). With fewer overlapping slots, our model flow further provides intent–slot fusion information about slots, resulting in an intent–slot classification accuracy of 98.6 per cent. On the other hand, Snips dataset has 7 intent tags and 72 slot tags, among which the intent "BookRestaurant" has the largest number of slot tags, with 15, and the intent "SearchCreativeWork" has the least number of slot tags with only 2. Through intent–slot fusion information, our model provides more clues to SF. However, compared with the experimental results of ATIS dataset, the score of slot F1, the performance of SF task on Snips dataset, is better, which is largely due to the larger number of samples contained in the training set.

In general, the proposed model performs well on both two datasets.

4.6 Ablation study

We performed the ablation experiments to further analyze and discuss the proposed model. Table IV shows the results of these ablation experiments.

4.6.1 Effect of CRF. Chen *et al.* (2019) argued that slot labels depend on the prediction of surrounding words. CRF is a very classical sequence labeling method. Deep learning adding with CRF has been widely used in the field of sequence labeling. In 2015, Zhou *et al.*, 2015 improved semantic role labeling by adding CRF layer to BiLSTM encoder. In order to explore the influence of CRF on the model results, we replaced the last softmax with CRF in the SF layer and carried out experiments. For comparison purposes, we named the model that uses CRF for SF as "Model-CRF."

The results of the experiment are shown in Figure 3, in which we can find that the performance of Model-CRF on ATIS dataset is comparable to ID-SF-Fusion. We believe that this is because of BERT's powerful extraction ability and the effect of intent–slot fusion layer. As a powerful encoding structure, BERT's output vectors fully model features and contain rich useful information. At the same time, the intent–slot fusion layer proposed in this paper has already transmitted the fusion vector rich in intent information and slot information to the SF task, so the addition of CRF has little impact on the performance of the model. However, in the Model-CRF, the three indicators all show a significant downward trend on the dataset Snips. The sample size of the training set of Snips dataset is almost three times that of ATIS. The larger training set and fewer intent labels and slot labels make the model generate vectors rich in slot information before CRF. We think that this result is due to the characteristics of the Snips dataset itself.

4.6.2 Effect of intent–slot information fusion layer. In order to explore the effect of our proposed intent–slot fusion layer, we further conduct ablation experiments. We remove the intent–slot information fusion layer, and replace it with the simple concatenation of intent information and slot information, and send the concatenation result to the SF layer. We named it "no-fusion."

Models	ATIS Intent ACC	ATIS Slot F1	ATIS Sentence ACC	Snips Intent ACC	Snips Slot F1	Snips Sentence ACC
Model-CRF	97.6	95.8	88.4	98.3	96.5	92.0
No-fusion	97.4	96.0	88.1	98.3	96.7	91.9
Sentence-level	97.8	95.8	87.7	98.1	96.3	91.6
Our model	98.0	96.1	88.4	98.6	96.7	92.9

Source: Table by authors

Table IV.
Ablation results (in %)

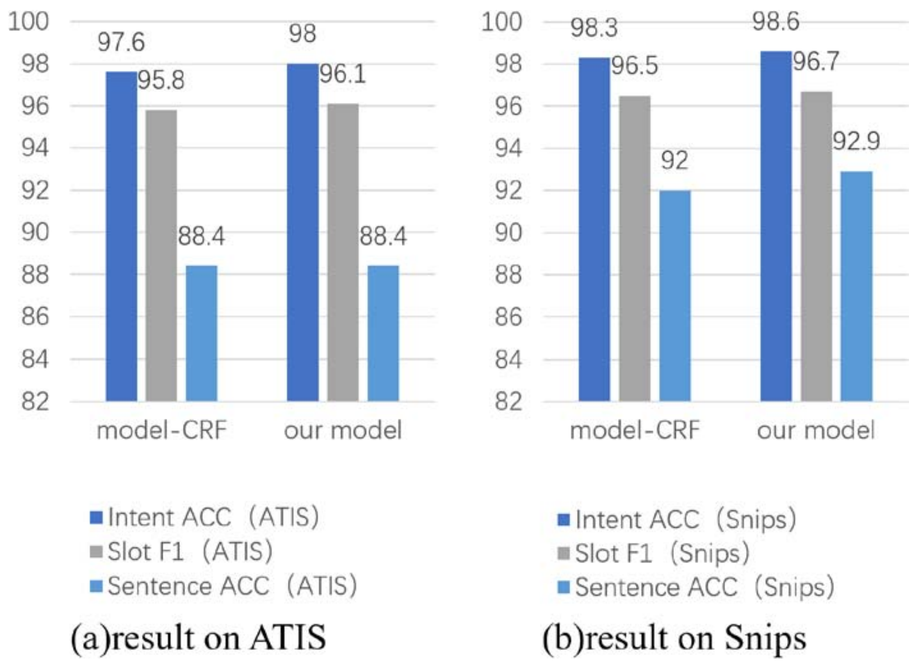


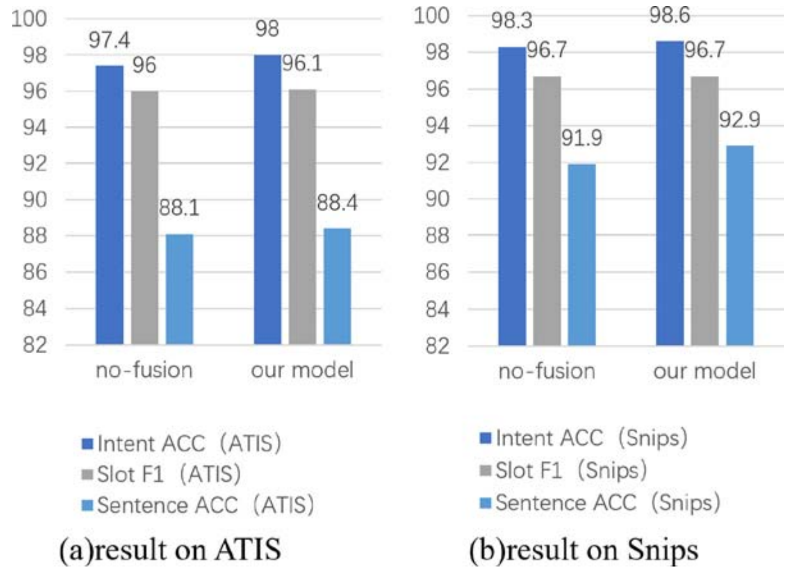
Figure 3.
Experimental results of ID-SF-Fusion and Model-CRF

Source: Figure by authors

As shown in Figure 4, the model with the fusion layer has a better performance. We believe is because the fusion layer captures the more useful part of intent information and slot information. Without the fusion layer, a large amount of useless information will interfere with the SF task, thus making the model less effective.

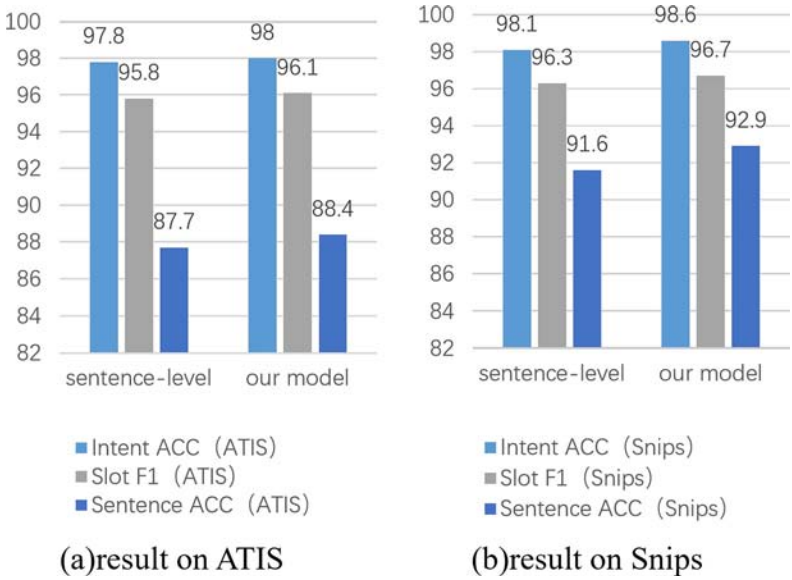
4.6.3 Effect of ID at the word level. In this section, in order to explore the influence of word-level ID on the model, we modify the part of ID into sentence-level ID for experiments. We concatenate the last hidden vector of the BiLSTM-ID layer and the word feature vector corresponding to "[CLS]" for ID, and concatenate the output of softmax layer for ID with the output of BiLTSM-SF respectively. The rest of the model remains unchanged and we name it "sentence-level."

Figure 5 shows the experimental results of sentence-level on the two datasets. It can be found that word-level ID shows better performance than sentence-level ID, because intent prediction at each time step has similar advantages to integrated neural networks, which can reduce the prediction variance and improve the performance of ID. Therefore, our model provides more useful intent information for SF by introducing word-level intent detection. If we only use sentence-level intent information to the fusion layer, we get worse results, which shows that using word-level intent information is effective. The main reason for this is probably that the intent information at the word level can preserve useful features for each time step.



Source: Figure by authors

Figure 4.
Experimental results of ID-SF-Fusion and no-fusion



Source: Figure by authors

Figure 5.
Experimental results of ID-SF-Fusion and sentence-level

5. Conclusion

In this paper, we propose a new joint recognition model for intention recognition and SF. The model uses intention recognition based on word level to further improve performance and reduce error propagation. In addition, we propose an intent–slot fusion layer to extract the information needed by SF for the completion of SF task. The proposed model is applied on two publicly available datasets ATIS and Snips, and our proposed model achieves significant improvements in intent classification accuracy, SF F1, and sentential semantic frame accuracy. The BERT model has a sea number of parameters, so in the future, we will use the improved BERT structure to reduce the model parameters, optimize the time and space complexity as much as possible and further explore the application in Chinese language scenarios.

ORCID iDs

Meng Zhu https://orcid.org/0000-0001-5859-0478
Xiaolong Xu https://orcid.org/0000-0001-6254-5864

Notes

1. https://github.com/howl-anderson/ATIS_dataset
2. https://github.com/snipsco/nlu-benchmark/

References

Chen, H., Liu, X., Yin, D. and Tang, J. (2017), "A survey on dialogue systems: recent advances and new frontiers", *ACM SIGKDD Explorations Newsletter*, Vol. 2, pp. 25-35. doi: 10.1145/3166054.3166058.

Chen, Q., Zhuo, Z. and Wang, W. (2019), "BERT for joint intent detection and slot filling", *arXiv preprint arXiv:1902.10909*, available at: https://arxiv.org/pdf/1902.10909.pdf (accessed 21 April 2022).

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E. and Cieliebak, M. (2021), "Survey on evaluation methods for dialogue systems", *Artificial Intelligence Review*, Vol. 1, pp. 755-810. doi: 10.1007/s10462-020-09866-x.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019), "BERT: pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv.1810.04805*. available at: https://doi.org/10.48550/arXiv.1810.04805 (accessed 30 July 2022).

Fan, J.F., Wang, M.L., Li, C.L., Zhu, Z.Q. and Mao, L. (2022), "Intent-slot correlation modeling for joint intent prediction and slot filling", *Journal Of Computer Science And Technology*, Vol. 2, pp. 309-319. doi: 10.1007/s11390-020-0326-4.

Fang, I. (2016), "Deep learning for query sementic domains classification", available at: www.semanticscholar.org/paper/Deep-Learning-for-Query-Semantic-Domains-Fang/b0d8498ad8a88b07005194fda26ecfb92613d3c1 (accessed 06 July 2022).

Goo, C.W., Gao, G., Hsu, Y.K., Huo, C.L., Chen, T.C., Hsu, K.W. and Chen, Y.N. (2018), "Slot-Gated modeling for joint slot filling and intent prediction", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Association for Computational Linguistics, Stroudsburg, USA, pp. 753-757. doi: 10.18653/v1/N18-2118.

Haihong, E., Niu, P., Chen, Z. and Song, M. (2 019), "A novel bi-directional interrelated model for joint intent detection and slot filling", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, USA, pp. 5467-5471. doi: 10.18653/v1/P19-1544.

Hakkani-Tür, D., Tur, G., Celikyilmaz, A., Chen, Y.N., Gao, J., Deng, L. and Wang, Y.Y. (2016), "Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM", *Interspeech 2016*, Vol. 27, pp. 715-719, doi: 10.21437/Interspeech.2016-402.

Han, S.C., Long, S., Li, H., Weld, H. and Poon, J. (2021), "Bi-directional joint neural networks for Intent detection and slot filling", *Interspeech 2021*, Vol. 32, pp. 4743-4747, doi: 10.21437/Interspeech.2021-2044.

Joachims, T. (1998), "Text categorization with support vector machines: learning with many relevant features", *Proceedings of the Machine Learning: ECML*-98, Springer Berlin Heidelberg, Berlin, Heidelberg, Germany, pp. 137-142. doi: 10.1007/BFb0026683.

Li, C., Li, L. and Qi, J. (2018), "A self-attentive model with gate mechanism for spoken language understanding", *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, USA, pp. 3824-3833. doi: 10.18653/v1/D18-1417.

Liu, B. and Lane, I. (2016), "Attention-based recurrent neural network models for joint intent detection and slot filling", *Proc. Interspeech 2016*, Association for Computational Linguistics, Stroudsburg, USA, pp. 685-689. doi: 10.21437/Interspeech.2016-1352.

Liu, J., Yu, M., Chen, Y. and Xu, J. (2022), "Cross-domain slot filling as machine reading comprehension: a new perspective", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, pp. 673-685. doi: 10.1109/TASLP.2022.3140559.

McCallum, A. and Nigam, K. (1998), "A comparison of event models for naive Bayes text classification", *Proceedings of the 10th European Conference on Machine Learning*, Springer-Verlag, Berlin, Heidelberg, Germany, pp. 137-142. available at: www.kamalnigam.com/papers/multinomial-aaaiws98.pdf

Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-Tur, D., He, X., Heck, L., Tur, G., Yu, D. and Zweig, G. (2015), "Using recurrent neural networks for slot filling in spoken language understanding", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 3, pp. 530-539. doi: 10.1109/TASLP.2014.2383614.

Peng, B. and Yao, K. (2015), "Recurrent neural networks with external memory for language understanding", *Natural Language Processing and Chinese Computing*, Springer, Berlin, Heidelberg, Germany, pp. 25-35. doi: 10.1007/978-3-319-25207-0_3.

Qin, L., Che, W., Li, Y., Wen, H. and Liu, T. (2019), "A stack-propagation framework with token-level intent detection for spoken language understanding", *in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP -IJCNLP)*, Association for Computational Linguistics, Stroudsburg, USA, pp. 2078-2087. doi: 10.18653/v1/D19-1214.

Samant, A.P., Warhade, K. and Gunale, K. (2021), "Pedestrian intent detection using skeleton-based prediction for road safety", *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, IEEE, Piscataway, USA, pp. 238-24. doi: 10.1109/ACCESS51619.2021.9563293.

Samant, R.M., Bachute, M.R., Gite, S. and Kotecha, K. (2022), "Framework for deep learning-based language models using multi-task learning in natural language understanding: a systematic literature review and future directions", *IEEE Access*, Vol. 10, pp. 17078-17097. doi: 10.1109/ACCESS.2022.3149798.

Valizadeh, M. and Parde, N. (2022), "The AI doctor is in: a survey of task-oriented dialogue systems for healthcare applications", *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Stroudsburg, USA, pp. 6638-6660. doi: 10.18653/v1/2022.acl-long.458.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017), "Attention is all you need", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Association for Computational Linguistics, Stroudsburg, USA, pp. 6000-6010. available at: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, Y., Shen, Y. and Jin, H. (2018), "A bi-model based RNN semantic frame parsing model for intent detection and slot filling", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *Volume 2 (Short Papers)*, Association for Computational Linguistics, Stroudsburg, USA, pp. 309-314. doi: 10.18653/v1/N18-2050.

Yao, K., Peng, B., Zhang, Y., Yu, D., Zweig, G. and Shi, Y. (2014), "Spoken language understanding using long short-term memory neural networks", *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, Piscataway, USA, pp. 189-194. doi: 10.1109/SLT.2014.7078572.

Ye, J., Zhou, X., Zheng, X., Gui, T. and Zhang, Q. (2022), "Uncertainty-aware sequence labeling", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, IEEE, Piscataway, USA, Vol. 30, pp. 1775-1788. doi: 10.1109/TASLP.2021.3138680.

Zhang, C., Li, Y ., Du, N., Fan, W. and Yu, P.S. (2019), "Joint slot filling and intent detection via capsule neural networks", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, USA, pp. 5259-5267. doi: 10.18653/v1/P19-1519.

Zhang, X.D. and Wang, H.F. (2016), "A joint model of intent determination and slot filling for spoken language understanding", *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2993-2999. available at: https://dl.acm.org/doi/10.5555/3060832.3061040.

Zhang, Y. (2021) "Research on text classification method based on LSTM neural network model", *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, IEEE, Piscataway, USA, pp. 1019-1102. doi: 10.1109/IPEC51340.2021.9421225.

Zhou, J. and Xu, W. (2015) "End-to-end learning of semantic role labeling using recurrent neural networks", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Stroudsburg, USA, pp. 1127-1137. doi: 10.3115/v1/P15-1109.

Corresponding author
Xiaolong Xu can be contacted at: xuxl@njupt.edu.cn