BIG-FG: A Bi-directional Interaction Graph Framework with Filter Gate Mechanism for Chinese Spoken Language Understanding

Wentao Zhang[1], Bi Zeng[1], Pengfei Wei[1(✉)], and Huiting Hu[2]

[1] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China
wpf@gdut.edu.cn

[2] School of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510006, China

Abstract. Spoken language understanding (SLU) primarily entailing slot filling and intent detection has been studied for many years with achieving significant results. However, in Chinese SLU tasks, Some models fail to take word-level information into account, and there is insufficient interaction between slot information and intent information. To address the aforementioned issues, we propose a novel bi-directional interaction graph framework with filter gate mechanism (**BIG-FG**) for Chinese spoken language understanding, which can make a fine-grained interaction directly with slot information and intent information, while also effectively fusing character-word semantic information. The model consists of two core modules: (1) bi-directional interaction graph (BIG), which is based on a multi-layer graph attention network with the bi-directional connections between intent information, slot information, and adjacent slot information, fully considering the correlation between slot filling and intent detection; (2) filter gate (FG), which enhances fusion performance by solving the problem of semantic ambiguity brought by direct fusion of character-word semantic information. Experiments on two datasets demonstrate that our model outperforms the best benchmark model by 0.39% and 2.65% in the Overall(Acc) evaluation metric, respectively, and accomplishes the state-of-the-arts performance.

Keywords: Chinese Spoken Language Understanding · Bi-directional Interaction Graph · Filter Gate · Graph Attention Network

1 Introduction

In intelligent dialogue systems, spoken language understanding (SLU) is critical, which typically includes two main subtasks: intent detection and slot filling [1].

L. Iliadis et al. (Eds.): ICANN 2023, LNCS 14257, pp. 304–315, 2023.
https://doi.org/10.1007/978-3-031-44216-2_25

Given an utterance for example, “play music on youtube”, the intent label is “PlayMusic”, and the slots are labeled in order {O, O, O, B-service}.

In the past, researchers focused mostly on English SLU and proposed different methods including gate-based methods [2–4], attention-based methods [5,6] and GATs-based methods [7,8]. In contrast to English SLU, Chinese SLU faces the difficulty of word segmentation. When there is an error in the word segmentation in Chinese SLU, it leads to error propagation, and, as a result, a slot filling error. To avoid this problem, [9] established a new collaborative memory network model based on the character to avoid the introduction of word segmentation. [10] introduced a two-stage modeling approach at the character level, exploiting the crossover effects between intent and slot information. However, it is commonly understood that Chinese word segmentation is critical for interpreting slots in an utterance. Given the utterance “我/想/听/稻香 (I want to listen to Rice Fragrance)” as an example, we use “/” to split the words in an utterance. If the model is based on character level, it is likely to wrongly predict the slot of “稻 (rice)” as “Rice_name”. However, by using the information of “稻香 (rice fragrance)” in the word segmentation, the model can easily predict “稻香 (rice fragrance)” slot as “Song”. To inject the word information into the Chinese SLU, [11] introduced a word adapter to combine information about characters and words. However, they did not consider the influence of redundant information in the word adapter. For example, in the example utterance mentioned above, the single word “稻香 (song)” and the single character “稻 (rice_name)” are semantically different, which causes semantic ambiguity when injecting word information, and the model lacks the guidance of slot information on intent detection.

To address these issues, we propose a unique bi-directional interaction graph framework to jointly model slot filling and intent detection, taking into account the correlation between slot information and intent information in Chinese SLU as well as alleviating the redundant information caused by the direct fusion of character-word semantic information: (1) bi-directional interaction graph, which uses slot information and intent information as feature nodes and creates bi-directionally connected edges between slot and intent information, and interacts via a multi-layer graph attention network; (2) filter gate. To fuse character-word semantic information efficiently, the redundant information caused by direct fusion is eliminated utilizing a filter gate fusion mechanism which can control the propagation of effective semantic information.

In summary, the following is the contributions of this work:

- We propose a bi-directional interaction graph for interacting with slot and intent information that takes into account the reciprocal facilitation of slot filling and intent detection.
- We propose a filter gate to limit the impact of redundant information owing to directly fusing character-word semantic information.
- Experiments on CAIS and SMP-ECDT datasets demonstrate that our model outperforms the best benchmark model and accomplishes the state-of-the-arts performance.

2 Related Works

Slot Filling and Intent Detection. The researchers proposed many implicit joint models considering the relationship between slot filling and intent detection tasks [6,12–14]. Essentially, they fail to make an explicit relationship between the two tasks. Later, some of researchers started to explore intent-augmented joint models and proposed many execellent approaches [2,3,11,15–17]. Nevertheless, these models do not account for the guiding role of slot information in intent detection. Recently, researchers have begun to explore models in which two tasks guide each other [4,10,18–21].

Graph Neural Networks. Currently, graph neural networks are performing very well in many fields. [22] applied graph attention networks to short text classification. [23] improved the performance of aspect-level sentiment classification by clarifying the dependencies between words through graph attention networks. Due to some limitations of GCN, the researcher proposed new approaches [24,25]. In SLU, [7,8] improved model performance by building effective interaction graphs. In our BIG-FG, a bi-directional interaction graph is built based on a multi-layer graph attention network to explicitly model the relationship between the two tasks, fully considering the mutual facilitation between intent detection and slot filling to enhance the performance of the model.

3 Approach

This work contributes to implementing slot filling and intent detection for Chinese SLU. In this section, the proposed approach will be introduced in detail. The overall framework of the model is shown in Fig. 1 (a). Firstly, the text encoding layer is introduced to realize the vectorized representation of characters and words in an utterance. Secondly, we propose the adaptive fusion module to obtain the slot and intent information. Next, the intent nodes and slot nodes are learned through an interrelated connection fusion using the bi-directional interaction graph. Finally, through a cooperative learning schema, slot filling and intent detection are optimized concurrently.

3.1 Text Encoding Layer

Following [11], we utilize a novel text encoding structure to obtain character-word information representation, which consists primarily of an embedding encoder, a self-attention, and a Bi-LSTM.

Character Encoding. Given a Chinese utterance $\boldsymbol{X} = \{x_1, x_2, x_3, \cdots, x_T\}$, T denotes the number of characters. Firstly, each character is transformed into a character vector $\boldsymbol{E}^c = \{\boldsymbol{e}_1^c, \boldsymbol{e}_2^c, \cdots, \boldsymbol{e}_T^c\}$, and secondly, high-level semantic information is obtained by self-attention and Bi-LSTM, respectively.

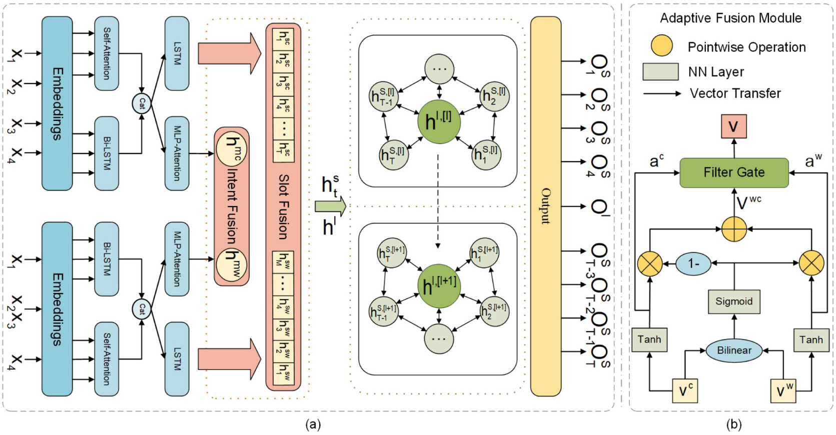

Fig. 1. The overall architecture of our proposed bi-directional interaction graph framework with filter gate fusion mechanism, where Cat denotes the concatenation operation. The internal structure of adaptive fusion module is shown in (b).

The self-attention [26] captures the features of the context of the characters in the utterance and the Bi-LSTM portrays the sequence from two directions, which can express more semantic information. We feed the characters vector $\boldsymbol{E}^c$ into self-attention and Bi-LSTM, respectively. The output vectors are $\boldsymbol{H}^A = \{\boldsymbol{h}_1^A, \boldsymbol{h}_2^A, \cdots, \boldsymbol{h}_T^A\}$ and $\boldsymbol{H}^L = \{\boldsymbol{h}_1^L, \boldsymbol{h}_2^L, \cdots, \boldsymbol{h}_T^L\}$, respectively.

Finally, the semantic information $\boldsymbol{h}_t^c = [\boldsymbol{h}_t^A, \boldsymbol{h}_t^L]$ is obtained by concatenating self-attention and Bi-LSTM. The final output sequence of the character-level semantic information is $\boldsymbol{H}^c = \{\boldsymbol{h}_1^c, \boldsymbol{h}_2^c, \cdots, \boldsymbol{h}_T^c\}$.

Word Encoding. We use an external CWS (Chinese Word Segmentation) system. Given the Chinese utterance $\boldsymbol{X}$, we obtain the word sequences $\boldsymbol{E}^w = \{\boldsymbol{e}_1^w, \boldsymbol{e}_2^w, \cdots, \boldsymbol{e}_M^w\}(M \leqslant T)$ by word segmentation and vectorization. The rest of the encoding part is the same as the character encoding, and the final word-level semantic encoding output is denoted as $\boldsymbol{H}^w = \{\boldsymbol{h}_1^w, \boldsymbol{h}_2^w, \cdots, \boldsymbol{h}_M^w\}$.

3.2 Adaptive Fusion Module

Fusion Module. Figure 1(b) represents the structure diagram of the adaptive fusion layer module. As one of the contribution points of this work, a filter gate mechanism is proposed based on bilinear fusion to reduce the impact of redundant information brought by character-word fusion. Following [11], with the word vector $\boldsymbol{v}^w \in \mathbb{R}^d$ and character vector $\boldsymbol{v}^c \in \mathbb{R}^d$ as inputs, we obtain the normalized word and character vectors $\boldsymbol{a}^w$, $\boldsymbol{a}^c$ by $Tanh$ activation function. After that, the weights are calculated using a bilinear function, and then they are weighted and summed to obtain $\boldsymbol{v}^{wc}$. The procedure for calculating is as follows:

$$\boldsymbol{a}^{w} = tanh(\boldsymbol{W}_{aw}\boldsymbol{v}^{w} + \boldsymbol{b}^{aw}) \tag{1}$$

$$\boldsymbol{a}^{c} = tanh(\boldsymbol{W}_{ac}\boldsymbol{v}^{c} + \boldsymbol{b}^{ac}) \tag{2}$$

$$\lambda = sigmoid(\boldsymbol{v}^{c}\boldsymbol{W}_{\lambda}\boldsymbol{v}^{w} + b^{\lambda}) \tag{3}$$

$$\boldsymbol{v}^{wc} = (1-\lambda)\boldsymbol{a}^{c} + \lambda\boldsymbol{a}^{w} \tag{4}$$

where $\boldsymbol{W}_{aw}$, $\boldsymbol{W}_{ac}$, $\boldsymbol{W}_{\lambda}$ are the trainable matrix weights and $\boldsymbol{b}^{aw}$, $\boldsymbol{b}^{ac}$, b^{λ} are the bias values of the linear transformation.

Considering the potential redundant information, we propose a filter gate to specifically utilize fusion feature. When the fusion feature $\boldsymbol{v}^{wc}$ is advantageous, the filter gate will combine both the fusion and original features, and obtain explicit slot boundary information. The calculation process is as follows:

$$\boldsymbol{f}_{\boldsymbol{c}} = \boldsymbol{W}_{fc}[\boldsymbol{a}^{c}, \boldsymbol{v}^{wc}] + \boldsymbol{b}^{fc} \tag{5}$$

$$\boldsymbol{f}_{w} = \boldsymbol{W}_{fw}[\boldsymbol{a}^{w}, \boldsymbol{v}^{wc}] + \boldsymbol{b}^{fw} \tag{6}$$

$$f_{g} = sigmoid(\boldsymbol{W}_{g}[\boldsymbol{f}_{c}, \boldsymbol{f}_{w}] + b^{wc}) \tag{7}$$

$$\boldsymbol{v} = f_{g} * \tanh(\boldsymbol{W}_{v}\boldsymbol{v}^{wc} + \boldsymbol{b}^{g}) \tag{8}$$

where $\boldsymbol{W}_{fc}$, $\boldsymbol{W}_{fw}$, $\boldsymbol{W}_{g}$, $\boldsymbol{W}_{v}$ are the trainable matrix weights; $\boldsymbol{b}^{fc}$, $\boldsymbol{b}^{fw}$, b^{wc}, and $\boldsymbol{b}^{g}$ are the bias values of the linear transformation, $[,]$ represents the concatenation operation, and $\boldsymbol{v}$ is the final fusion output. The above formula for the adaptive fusion layer can be abbreviated as $\boldsymbol{v} = AFM(\boldsymbol{v}^{c}, \boldsymbol{v}^{w})$.

Intent Fusion. We employ MLP attention to obtain an informative representation of the entire utterance $\boldsymbol{h}^{mc} \in \mathbb{R}^{d}$. Similarly, we can also obtain the word-level representation of the information of the whole utterance $\boldsymbol{h}^{mw} \in \mathbb{R}^{d}$ and get the fused intent information representation $\boldsymbol{h}^{I}$ through the adaptive fusion layer.

$$\boldsymbol{h}^{I} = AFM(\boldsymbol{h}^{mc}, \boldsymbol{h}^{mw}) \tag{9}$$

Slot Fusion. By unidirectional LSTM, we can obtain more appropriate slot information $\boldsymbol{h}_{t}^{sc}$, $\boldsymbol{h}_{t}^{sw}$. Then, through the adaptive fusion layer, the fused slot information is obtained, denoted as $\boldsymbol{h}_{t}^{S}$[1].

$$\boldsymbol{h}_{t}^{S} = AFM(\boldsymbol{h}_{t}^{sc}, \boldsymbol{h}_{f_{align}(t,\boldsymbol{w})}^{sw}) \tag{10}$$

3.3 Bi-directional Interaction Graph Module

Another contribution point of this work is to carry out the interaction of slot information and intent information. We propose a bi-directional interaction graph module, as shown in Fig. 1. By constructing different edges and feature nodes, a multi-layer graph attention network is utilized to fully interact with the information between the intent and slot.

[1] Given a word sequence $\boldsymbol{w}$={“打”,“开”,“相机”}, $f_{align}(t, \boldsymbol{w})$ provides the index of the word that goes with the t-th character in $\boldsymbol{w}$ (e.g.,$f_{align}(1, \boldsymbol{w})$=1,$f_{align}(3, \boldsymbol{w})$=3,$f_{align}(4, \boldsymbol{w})$=3).

Graph Attention Network. The GAT is a crucial network structure in the domain of deep learning which utilizes the attention mechanism to perform adaptive weighting of various neighboring edges, significantly enhancing the expressive capability. Given a series of feature nodes $\boldsymbol{Z} = \{\boldsymbol{z_1}, \boldsymbol{z_2}, \cdots, \boldsymbol{z_N}\}$, N is the total number of nodes. The graph attention network generates new node features, $\boldsymbol{Z}' = \{\boldsymbol{z}'_1, \boldsymbol{z}'_2, \cdots, \boldsymbol{z}'_N\}$ as the output.

$$\alpha_{ij} = \frac{exp(f(\boldsymbol{a}^T[\boldsymbol{W}_z\boldsymbol{z}_i, \boldsymbol{W}_z\boldsymbol{z}_j]))}{\sum_{k' \in N_i} exp(f(\boldsymbol{a}^T[\boldsymbol{W}_z\boldsymbol{z}_i, \boldsymbol{W}_z\boldsymbol{z}_{k'}]))} \tag{11}$$

$$\boldsymbol{z}'_i = \overset{K}{\underset{k=1}{\|}} \sigma(\sum_{j \in N_i} \alpha^k_{ij}\boldsymbol{W}^k_z\boldsymbol{z}_j) \tag{12}$$

where α^k_{ij} denotes the attention weight at k-th head, $\boldsymbol{W}^k_z$ denotes the k-th trainable matrix weight, and $\|$ denotes the concatenation operation. In this work, the multi-headed graph attention layer is directly adopted to the bi-directional interaction graph.

Bi-directional Interaction Graph. The correlation between intent and slot is quite essential in SLU. The slot is a reflection of character-level information and the intent is a reflection of sentence-level information.

For the feature nodes of the bi-directional interaction graph, we concatenate the intent hidden layer representation $\boldsymbol{h}^I$ obtained from the intent fusion layer and the slot hidden layer representation $\boldsymbol{h}^S_t$ obtained from the slot fusion layer to be the nodes of the bi-directional interaction graph $\boldsymbol{H}^{[l]}_g = \{\boldsymbol{h}^{I,[l]}, \boldsymbol{h}^{S,[l]}_1, \boldsymbol{h}^{S,[l]}_2, \cdots, \boldsymbol{h}^{S,[l]}_T\}$. $\boldsymbol{h}^{I,[l]}$ denotes the intent hidden layer feature of the l-th layer and $\boldsymbol{h}^{S,[l]}_t$ denotes the slot hidden layer feature of the l-th layer.

For the edges of the bi-directional interaction graph, we establish a bi-directio- nal connection between the each slot and intent node, and due to a correlation between contexts, bi-directional connections are also established between slot and slot at adjacent location.

In order to fully interact with the feature between the intent and slot, a multi-layer bi-directional interaction graph is constructed. For a bi-direction graph with $(l+1)$ layers of interaction, a hidden layer feature of the bi-direction interaction graph at $(l+1)$-th layer can be obtained, and this hidden layer feature is used as the final output.

$$\boldsymbol{H}^{[l+1]}_g = \textit{multi-head}\,GAT^{[l]}(\boldsymbol{H}^{[l]}_g) \tag{13}$$

$$\boldsymbol{h}^{fI}, \boldsymbol{h}^{fs}_t = \boldsymbol{h}^{I,[l+1]}, \boldsymbol{h}^{S,[l+1]}_t \tag{14}$$

where $\textit{multi-head}\,GAT^{[l]}$ represents the multi-head graph attention network at l-th layer, $\boldsymbol{h}^{I,[l+1]}$ and $\boldsymbol{h}^{S,[l+1]}_t$ are the intent feature and slot feature at $(l+1)$-th layer, respectively, $\boldsymbol{h}^{fI}$ is the output of the intent feature, and $\boldsymbol{h}^{fs}_t$ represents the output of the slot feature.

Through the linear layer, $\boldsymbol{h}^{fI}$ and $\boldsymbol{h}^{fs}_t$ are used for intent detection and slot filling, respectively. $\boldsymbol{y}^I = softmax(\boldsymbol{W}_{fI}\boldsymbol{h}^{fI})$ and $\boldsymbol{y}^S_t = softmax(\boldsymbol{W}_{fs}\boldsymbol{h}^{fs}_t)$, where $\boldsymbol{W}_{fI}$ and $\boldsymbol{W}_{fs}$ are trainable parameters. $O^I = argmax(\boldsymbol{y}^I)$ is the predicted intent tags and $O_t^S = argmax(\boldsymbol{y}_t^S)$ is the predicted slot labels in an utterance.

3.4 Loss Function

In this work, cross-entropy is employed as the loss function. The training objective for combining intent and slot loss is to reduce the value of the following loss function:

$$\mathcal{L}_\theta = -\mu \sum_{i=1}^{N_I} \hat{y}_i^I log(y_i^I) - (1-\mu) \sum_{t=1}^{T} \sum_{i=1}^{N_S} \hat{y}_t^{S,i} log(y_t^{S,i}) \tag{15}$$

where N_I indicates the number of intent labels, T indicates the number of characters in an utterance, N_S indicates the number of slot labels, μ is a hyperparameter, $\hat{y}^I$ and $\hat{y}^S$ indicate the true tags of the intent and the true tags of the slot, respectively.

4 Experiment

4.1 Datasets and Evaluation Metrics

To verify the feasibility of the approach, two openly accessible Chinese datasets CAIS [9] and SMP-ECDT[2] [11] are selected to conduct experiments. The CAIS dataset contains 7995 training sets, 994 validation sets, and 1024 test sets. There are 1655 training sets, 413 validation sets, and 508 test sets in the SMP-ECDT dataset. Following [2,16], we assess the effectiveness of Chinese SLU intent prediction using precision, slot filling using F1 score, and utterance-level semantic frame parsing using overall precision. In this work, the Chinese natural language processing system (Language Technology Platform, LTP) is adopted to acquire Chinese word segmentation[3].

4.2 Implementation Details

We conduct experiments with the GPU of Tesla A100 and PyTorch framework. All of the model weights begin with a uniform distribution as the initialization. The dropout rate is 0.5. The number of layers of graph attention network is 2, and the hidden layer dimension of each layer is 128, using 8 heads. 1.0 is chosen as the value for the maximum norm of gradient clipping. The L2 norm coefficient is 10^{-6}. The Adam uses a learning rate of 5×10^{-4} to update all parameters.

[2] https://conference.cipsc.org.cn/smp2019/evaluation.html.

[3] http://ltp.ai/.

Table 1. Main results on CAIS and SMP-ECDT.

Models	CAIS			SMP-ECDT		
	Slot (F1)	Intent (Acc)	Overall (Acc)	Slot (F1)	Intent (Acc)	Overall (Acc)
Slot-Gated Full Atten [2]	81.13	94.37	80.83	60.91	86.02	53.75
SF-ID Network [19]	84.85	94.27	82.41	63.90	88.85	55.67
CM-Net [9]	86.16	94.56	-	-	-	-
Stack-propagation [16]	87.64	94.37	84.68	71.32	91.06	63.75
MLWA [11]	88.61	95.16	86.17	75.26	94.22	67.58
GAIR [10]	88.92	95.45	86.86	76.08	94.56	68.58
BIG-FG	**89.83**	**95.65**	**87.25**	**76.56**	**95.72**	**71.23**

4.3 Baseline Models

In order to compare with other researchers' models, some meaningful baseline models are selected including Slot-Gated [2], CM-Net [9], SF-ID Network [19], MLWA [11], Stack-Propagation [16] and GAIR [10]. We take advantage of these models' published performance data from the literature [11] on the CAIS dataset. On the SMP-ECDT dataset, we execute the published code of the comparative models utilizing the split test set, with the exception of CM-Net [9] that fails to share codes.

4.4 Main Results

Table 1 shows the primary results and some comparative baselines of the proposed model on the CAIS and SMP-ECDT datasets. From the results, we can notice that the GAIR model without injecting word information performs somewhat better than the MLWA model with injecting word information on all metrics. This result occurs since the GAIR model is two-stage and takes into account the bi-directional correlation between intent and slot information, whereas MLWA just takes into account the influence of intent information on slots, despite the addition of word information. Our proposed BIG-FG model is compared with the GAIR model. On the CAIS dataset, we accomplish a 0.91% increase in Slot (F1), a 0.20% increase in Intent (Acc), and a 0.39% increase in Overall (Acc). On the SMP-ECDT dataset, we accomplish improvements of 0.48% on Slot (F1), 1.16% on Intent (Acc), and 2.65% on Overall (Acc). From the results, we observe that our model performs better than the top baseline model and achieves state-of-the-art performance. We attribute the improvement to the following reasons: (1) Our model in this work introduces word information while using character information, which solves the problem of ambiguous word boundary; (2) Considering the mutual facilitation between slot filling and intent detection, effective interaction is carried out through the BIG module to achieve mutual communication between the two tasks and enhance the model performance; (3) The filter gate based on bilinear fusion reduces redundant information, resulting in enhanced model performance. However, the MLWA model fails to consider the influence of redundant information and the guiding role of slot information on intent detection, and the GAIR model ignores the influence of Chinese word segmentation.

Table 2. Ablation study on CAIS and SMP-ECDT datasets.

Models	CAIS			SMP-ECDT		
	Slot (F1)	Intent (Acc)	Overall (Acc)	Slot (F1)	Intent (Acc)	Overall (Acc)
w/o intent fusion	87.13	94.37	85.83	73.91	94.42	67.75
w/o slot fusion	86.85	95.27	86.01	74.91	95.55	68.67
w/o filter gate	89.80	95.45	86.75	76.36	94.97	69.84
w/o BIG	88.60	94.46	85.37	75.39	95.23	69.09
BIG-FG	**89.83**	**95.65**	**87.25**	**76.56**	**95.72**	**71.23**

4.5 Analysis

To investigate the effect of the components in the proposed model, ablation experiments are carried out to verify the effectiveness. Table 2 shows the results of the ablation experiments. In addtion, We also explore the effect of the number of BIG layers and different word segmentors on the performance of the model.

Effect of Intent Fusion Layer. To explore whether the intent fusion layer plays a role in the model, the intent fusion layer is removed in this experiment, meaning that the intent information provided by the characters and words is utilized directly, which is named as *w/o intent fusion*. From the results in *w/o intent fusion* row in Table 2, we can observe 1.28% and 1.30% drops on Intent (Acc) on the CAIS and SMP-ECDT datasets, respectively, and other evaluation metrics also decrease, which demonstrates that the intent fusion layer can efficiently fuse the intent information provided by the characters and words, and achieve improved semantic features for intent detection.

Effect of Slot Fusion Layer. Similarly, we remove the slot fusion layer to investigate whether the slot fusion layer enriches semantic knowledge of slot, which is named as *w/o slot fusion*. The experimental results are shown in *w/o slot fusion* row in Table 2. Slot (F1) decreases by 2.98% on the CAIS dataset and by 1.65% on the SMP-ECDT dataset. This indicates that directly fusing the slot information provided by characters and words is not effective, and the slot fusion layer in our model can improve the information representation of slot.

Effect of Filter Gate. The filter gate reduces the redundant information that would be produced by the direct fusion of character-word semantic information. To investigate the specific effect of the filter gate in our model, we remove the filter gate and directly fuse character-word semantic information with bilinearity, which is named as *w/o filter gate*. The *w/o filter gate* row in Table 2 shows that the Slot (F1), Intent (Acc), and Overall (Acc) decrease on both CAIS and SMP-ECDT datasets, indicating that bilinear fusion is not effective when there is no filter gate, and the filter gate plays an important role in the BIG-FG model.

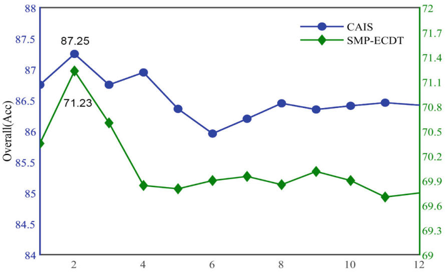

Fig. 2. Effect of the number of BIG layers and the horizontal axis indicates the number of layers of the BIG.

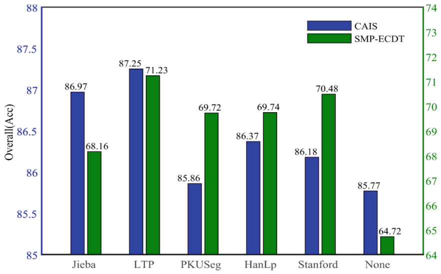

Fig. 3. Effect of word segmentors and the horizontal axis indicates the different word segmentors.

Effect of Bi-directional Interaction Graph. To explore the effect of the proposed bi-directional interaction graph module, we remove the bi-directional interaction graph module and directly use LSTM as the decoder. The intent detection and slot filling are modeled independently. It is named as *w/o BIG* in this experiment. The *w/o BIG* row in Table 2 shows that the evaluation metrics on two datasets drop more, which indicates that the independent modeling of intent and slot information is worse than the explicit joint modeling of intent and slot information and the BIG fully interacts with the features of intent and slot, thus enhancing the overall performance.

Effect of the Number of BIG Layers. To investigate the influence of the number of BIG layers, we plot the relationship between the Overall (Acc) and the number of layers of BIG, as shown in Fig. 2. It can be clearly seen from the figure that the model performance improves with the number of layers of the BIG, and the best performance on the CAIS and SMP-ECDT datasets is achieved when the number of layers of the BIG is 2, after which it decreases to gradually stabilize. The reason is that when there are too many layers, the model tends to exhibit excessive smoothing, resulting in the inclusion of redundant information. In general, choosing the optimal number of layers of BIG can improve the performance of the model.

Effect of Different Word Segmentors. We choose five different word segmentors for our experiments to investigate the impact of different word segmentors on the performance of our model, including Jieba[4], LTP, PKUSeg[5], HanLp[6], and Stanford[7]. Furthermore, we add an additional set of experiments without word segmentation information to validate the benefits of adding word segmentation information. Figure 3 shows the results of the experiments. We use Overall (Acc) as evaluation indice, and we can see that different word segmentation methods

[4] https://github.com/fxsjy/jieba.
[5] https://github.com/lancopku/PKUSeg-python.
[6] https://github.com/hankcs/HanLP.
[7] https://stanfordnlp.github.io/CoreNLP/.

perform differently. However, in our model, the LTP method has the best word segmentation effect. It is noteworthy that the model with the addition of word information has better effect than model without the addition of word information, which indicates the effectiveness of word segmentation.

5 Conclusion and Future Work

In this work, we proposed a novel bi-directional interaction graph framework with a filter gate mechanism for Chinese SLU. While reducing redundant information, we effectively fused the semantic information provided at the character level as well as word level. Furthermore, we took advantage of the correlation between intent and slot feature to enrich the semantic representations of intent and slot, thus improving the model performance. Experiments on SMP-ECDT and CAIS datasets showed that our model achieved the best performance. In the future, we will consider adding some prior knowledge and exploring a new fusion mechanism to fully fuse word information to enhance the performance of Chinese spoken language understanding tasks.

Acknowledgement. This work was supported in part by the National Science Foundation of China under Grant 62172111 and in part by the Natural Science Foundation of Guangdong Province under Grant 2019A1515011056.

References

1. Tur, G., De Mori, R.: Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. Wiley, Hoboken (2011)
2. Goo, C.W., et al.: Slot-gated modeling for joint slot filling and intent prediction. In: Proc. NAACL, pp. 753–757 (2018)
3. Li, C., Li, L., Qi, J.: A self-attentive model with gate mechanism for spoken language understanding. In: Proceedings of EMNLP, pp. 3824–3833 (2018)
4. Sun, C., Lv, L., Liu, T., Li, T.: A joint model based on interactive gate mechanism for spoken language understanding. Appl. Intell. **52**, 6057–6064 (2022)
5. Chen, M., Zeng, J., Lou, J.: A self-attention joint model for spoken language understanding in situational dialog applications. arXiv preprint arXiv:1905.11393 (2019)
6. Liu, B., Lane, I.: Attention-based recurrent neural network models for joint intent detection and slot filling. In: Proceedings of Interspeech, pp. 685–689 (2016)
7. Ding, Z., Yang, Z., Lin, H., Wang, J.: Focus on interaction: a novel dynamic graph model for joint multiple intent detection and slot filling. In: Proceedings of IJCAI, pp. 3801–3807 (2021)
8. Qin, L., Wei, F., Xie, T., Xu, X., Che, W., Liu, T.: GL-GIN: fast and accurate non-autoregressive model for joint multiple intent detection and slot filling. In: Proceedings of ACL-IJCNLP, pp. 178–188 (2021)
9. Liu, Y., Meng, F., Zhang, J., Zhou, J., Chen, Y., Xu, J.: CM-Net: a novel collaborative memory network for spoken language understanding. In: Proceedings of EMNLP-IJCNLP, pp. 1051–1060 (2019)

10. Zhu, Z., Huang, P., Huang, H., Liu, S., Lao, L.: A graph attention interactive refine framework with contextual regularization for jointing intent detection and slot filling. In: Proceedings of ICASSP, pp. 7617–7621 (2022)
11. Teng, D., Qin, L., Che, W., Zhao, S., Liu, T.: Injecting word information with multi-level word adapter for Chinese spoken language understanding. In: Proceedings of ICASSP, pp. 8188–8192 (2021)
12. Guo, D., Tur, G., Yih, W.t., Zweig, G.: Joint semantic utterance classification and slot filling with recursive neural networks. In: Proceedings of SLT, pp. 554–559 (2014)
13. Hakkani-Tür, D., et al.: Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM. In: Proceedings of Interspeech, pp. 715–719 (2016)
14. Xu, P., Sarikaya, R.: Convolutional neural network based triangular CRF for joint intent detection and slot filling. In: Proceedings of ASRU. pp. 78–83 (2013)
15. Ma, Z., Sun, B., Li, S.: A two-stage selective fusion framework for joint intent detection and slot filling. IEEE Trans. Neural Netw. Learn. Syst. 1–12 (2022)
16. Qin, L., Che, W., Li, Y., Wen, H., Liu, T.: A stack-propagation framework with token-level intent detection for spoken language understanding. In: Proceedings of EMNLP-IJCNLP, pp. 2078–2087 (2019)
17. Zhou, B., Zhang, Y., Sui, X., Song, K., Yuan, X.: Multi-grained label refinement network with dependency structures for joint intent detection and slot filling. arXiv preprint arXiv:2209.04156 (2022)
18. Chen, D., Huang, Z., Wu, X., Ge, S., Zou, Y.: Towards joint intent detection and slot filling via higher-order attention. In: Proceedings of IJCAI, pp. 4072–4078 (2022)
19. Haihong, E., Niu, P., Chen, Z., Song, M.: A novel bi-directional interrelated model for joint intent detection and slot filling. In: Proceedings of ACL, pp. 5467–5471 (2019)
20. Wang, J., Wei, K., Radfar, M., Zhang, W., Chung, C.: Encoding syntactic knowledge in transformer encoder for intent detection and slot filling. In: Proceedings of AAAI, vol. 35, pp. 13943–13951 (2021)
21. Zhang, C., Li, Y., Du, N., Fan, W., Yu, P.: Joint slot filling and intent detection via capsule neural networks. In: Proceedings of ACL, pp. 5259–5267 (2019)
22. Linmei, H., Yang, T., Shi, C., Ji, H., Li, X.: Heterogeneous graph attention networks for semi-supervised short text classification. In: Proceedings of EMNLP-IJCNLP, pp. 4821–4830 (2019)
23. Huang, B., Carley, K.: Syntax-aware aspect level sentiment classification with graph attention networks. In: Proceedings of EMNLP-IJCNLP, pp. 5469–5477 (2019)
24. Wang, S., Wu, Z., Chen, Y., Chen, Y.: Beyond graph convolutional network: an interpretable regularizer-centered optimization framework. arXiv preprint arXiv:2301.04318 (2023)
25. Chen, M., Wei, Z., Huang, Z., Ding, B., Li, Y.: Simple and deep graph convolutional networks. In: Proceedings of ICML, pp. 1725–1735 (2020)
26. Vaswani, A., et al.: Attention is all you need. In: Proceedings of NeurIPS, pp. 6000–6010 (2017)