A Deep Joint Model of Multi-scale Intent-Slots Interaction with Second-Order Gate for SLU

Qingpeng Wen[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. Slot filling and intent detection are crucial tasks of Spoken Language Understanding (SLU). However, most existing joint models establish shallow connections between intent and slot by sharing parameters, which cannot fully utilize their rich interaction information. Meanwhile, the character and word fusion methods used in the Chinese SLU simply combines the initial information without appropriate guidance, making it easy to introduce a large amount of noisy information. In this paper, we propose a deep joint model of Multi-Scale intent-slots Interaction with Second-Order Gate for Chinese SLU **(MSIM-SOG)**. The model consists of two main modules: (1) the Multi-Scale intent-slots Interaction Module (MSIM), which enables cyclic updating the multi-scale information to achieve deep bi-directional interaction of intent and slots; (2) the Second-Order Gate Module (SOG), which controls the propagation of valuable information through the gate with second-order weights, reduces the noise information of fusion, accelerates model convergence, and alleviates model overfitting. Experiments on two public datasets demonstrate that our model outperforms the baseline and achieves state-of-the-art performance compared to previous models.

Keywords: Intent Detection · Slot Filling · Multi-Scale intent-slots Interaction Module (MSIM) · Second-Order Gate (SOG)

1 Introduction

For task-oriented dialogue systems, Spoken Language Understanding (SLU) is a critical component [17], it includes two subtasks Intent Detection (ID) and Slot Filling (SF) [4]. SF is a sequence labeling task to obtain the slot information of the utterance; ID is a classification task to identify the intent of the utterance. An example of a simple Chinese SLU is shown in Fig. 1.

B. Luo et al. (Eds.): ICONIP 2023, CCIS 1966, pp. 42–54, 2024.
https://doi.org/10.1007/978-981-99-8148-9_4

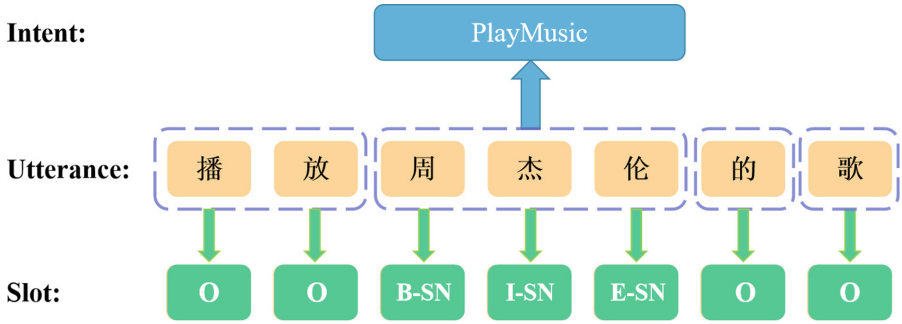


Fig. 1. An example of Chinese SLU, where `B-SN` denotes `B-singer_name`, `I-SN` denotes `I-singer_name` and `E-SN` denotes `E-singer_name`, the blue dashed box denotes word segmentation and the yellow box denotes character segmentation.

The main challenge in English SLU research is correlating ID and SF effectively. In response, Xu et al. [18] proposed a variable-length attention encoder-decoder model in 2020, in which SF is guided by intent information and achieves intent-enhanced association, but it lacks a bi-directional correlation between ID and SF. Recent research [11,14,16] has demonstrated that ID and SF tasks can mutually reinforce each other. Accordingly, Li et al. [7] proposed a bi-directional correlation BiLSTM-CRF model in 2022, updating ID and SF in both directions, but the deep interaction remains unestablished.

Compared to English SLU, Chinese SLU also faces challenges in segmenting Chinese utterances and effectively integrating character information. Unlike English, Chinese lacks natural word separators, rendering character segmentation techniques unreliable. As shown in Fig. 1, the segmentation of '周杰伦(Jay Chou)' into '周(week)-杰(Jay)-伦(Aron)' by characters incorrectly predicts '周(week)' as 'Datetime_date'. However, we expect the model to correctly segment it into '周杰伦(Jay Chou)' and predict it as the slot label 'singer_name' by using a suitable Chinese Word Segmentation (CWS) system. To address this, Teng et al. [15] improved the CWS system using Multi-level Word Adapter to fuse character and word information, but it lacks bi-directional interaction between ID and SF and introduces noise and overfitting problems in the fusion mechanism. This paper proposes a deep joint model of Multi-Scale intent-slots Interaction with Second-Order Gate (MSIM-SOG) to better fuse character and word information and establish a deep bi-directional interaction between two tasks. Experimental results on two publicly datasets called CAIS and SMP-ECDT show that our model outperforms all other models and achieves SOTA performance.

To summarize, the following are the contributions of this paper:

- In this paper, we propose a deep joint model of Multi-Scale intent-slots Interaction with Second-Order Gate for Chinese SLU (MSIM-SOG), which optimizes the performance of Chinese SLU tasks and improves current joint model.

- A Multi-Scale intent-slots Interaction Module (MSIM) is proposed in this paper, which enables deep bi-directional interaction between ID and SF by cyclically updating the multi-scale information on intent and slots.
- A Second-Order Gate module (SOG) is proposed to fuse character and word information, control effective information propagation through the gate with second-order weights, reduce noise and accelerate model convergence.
- On the public CAIS and SMP-ECDT datasets, our model improves the semantic accuracy by **0.49%** and **2.61%** over the existing models respectively, and achieves the competitive performance.

For this paper, the code is public at https://github.com/QingpengWen/MSIM-SOG.

2 Related Work

English SLU Task: The Spoken Language Understanding (SLU) task consists of two main tasks: Intent Detection (ID) and Slot Filling (SF). Early research in ID often utilized common classification methods like SVM [2] and RNN [6]. While SF extracts semantic information through sequence labeling such as CRF [19] and LSTM [20]. However, these approaches commonly cause error propagation as they lack the interaction of ID and SF. Ma et al. [10] proposed a two-stage selective fusion framework that explored intent-enhanced models. However, it simply guided slot by intent, and the bi-directional relationship was still not established. Sun et al. [13] designed a bi-directional interaction model based on a gate mechanism to achieve bi-directional association between ID and SF.

Chinese SLU Task: Although these approaches have made great progress in English SLU, there are still some challenges in Chinese SLU include dealing with ambiguous words, the effective fusion of word and character information, and lacks of deep bi-directional interaction models for ID and SF. As a result, existing English SLU models cannot be directly applied to the Chinese SLU task. To address this, Zhu et al. [21] proposed a two-stage Graph Attention Interaction Refine framework to mitigate ambiguity in Chinese SLU, but it may incorrectly identify slot boundaries due to the absence of CWS system. Teng et al. [15] proposed a Multi-level Word Adapter to fuse character and word information, but it only used the intent guidance slot, while the fusion mechanism they used introduces noisy information and risks losing critical information.

3 Approach

In this section, we will introduce the MSIM-SOG model proposed in this paper in detail. The general model framework is illustrated in Fig. 2.

3.1 Char-Word Channel Layer

Based on MLWA [15], we construct a character-level and word-level channel layer (Char-Word Channel Layer), which obtains the complete character sequence information and utterance representation information for SF and ID tasks. The Char-Word Channel Layer consists of the Self-Attentive Encoder module, the LSTM module, and the MLP Attention module. Among them, the Self-Attentive Encoder module extracts the character and word encoding representation. Then the LSTM module is utilized to extract the contextual and sequence information for the SF task, while the MLP Attention module extracts the complete representation of the utterance for the ID task.

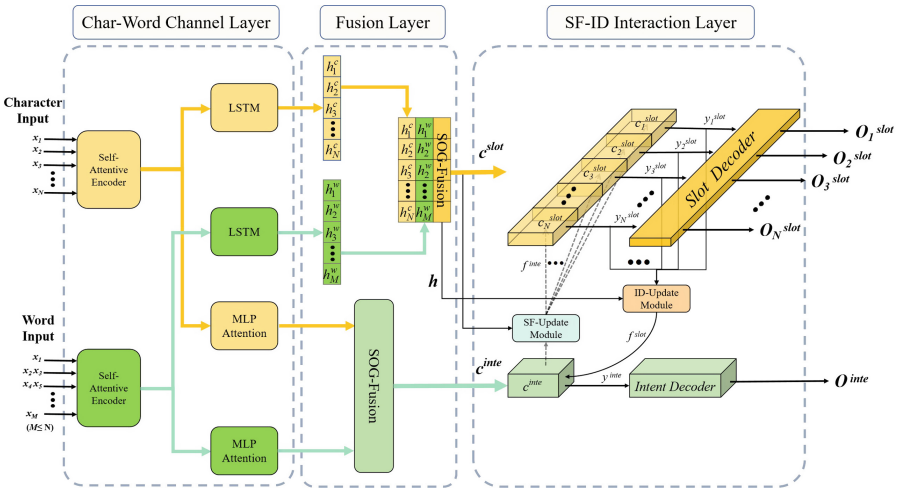


Fig. 2. The MSIM-SOG model proposed in this paper. The model includes our Char-Word Channel Layer, Fusion Layer and SF-ID interaction layer. The internal structure of the SOG Fusion module is shown in Fig. 3.

Self-attentive Encoder: The Self-Attentive Encoder mainly consists of an Embedding encoder, a Self-Attention encoder, and a BiLSTM encoder [3]. For a given Chinese utterance $c = \{c_1, c_2, ..., c_N\}$ containing N characters. The Embedding encoder converts it into the character vector $\mathbf{E}^c_{emb} \in \mathbb{R}^{N\times d} = \{\mathbf{e}^{c,e}_1, \mathbf{e}^{c,e}_2, \cdots, \mathbf{e}^{c,e}_N\}$. The BiLSTM encoder loops the input utterance forward and backward to obtain context-aware sequence feature information $H^c \in \mathbb{R}^{N\times d} = \{h^c_1, h^c_2, ..., h^c_N\}$, where $h^c_j \in \mathbb{R}^d = BiLSTM(e^{c,e}_j, h^c_{j-1}, h^c_{j+1})$ and the Self-Attention encoder captures the contextual information of each character in a valid sequence as $A^x \in \mathbb{R}^{N\times d} = softmax(\frac{Q\cdot K^T}{\sqrt{d^k}}) \cdot V$, where Q, K and V are matrices acquired by the application of different linear projections to the input vectors and d^k denotes the vector dimension. Subsequently, we concatenate these outputs to obtain the final character-level encoding representation as $\mathbf{E}^c \in \mathbb{R}^{N\times 2d} = \{\mathbf{e}^c_1, \mathbf{e}^c_2, \cdots, \mathbf{e}^c_N\}$.

For word-level encoding, we adopt the CWS system to capture the word segmentation sequence $w = \{w_1, w_2, ..., w_M\}$ (M ≤ N) by segmenting the utterance. And the final word-level encoding is denoted as $\mathbf{E}^w \in \mathbb{R}^{M\times 2d} = \{\mathbf{e}_1^w, \mathbf{e}_2^w, \cdots, \mathbf{e}_M^w\}$.

LSTM: In the LSTM module, we extract the contextual information of the character-level encoding $\mathbf{E}^c$ and capture the character sequence information to obtain the hidden state output $\mathbf{H}^c \in \mathbb{R}^{N\times 2d} = \{\mathbf{h}_1^c, \mathbf{h}_2^c, \mathbf{h}_3^c, \cdots, \mathbf{h}_N^c\}$, and use it for SF task, where $h_j^c = LSTM(e_j^c, h_{j-1}^c)$.

Equally, by extracting the word-level encoding information $\mathbf{E}^w$, we obtain the output of the hidden state is $\mathbf{H}^w \in \mathbb{R}^{M\times 2d} = \{\mathbf{h}_1^w, \mathbf{h}_2^w, \cdots, \mathbf{h}_M^w\}$.

MLP Attention: In the MLP Attention module, we extract the complete utterance representation information $\mathbf{S}^c \in \mathbb{R}^{2d}$ and the complete word-level representation information $\mathbf{S}^w \in \mathbb{R}^{2d}$ by computing the weighted sum of all hidden units $\mathbf{E}^c$ and $\mathbf{E}^c$ in the Self-Attentive Encoder.

3.2 Fusion Layer

Since the current fusion mechanism simply combines the initial information without the corresponding guidance, it is easy to introduce a large amount of noise and redundant information, thus missing useful information. To solve above problems, we propose a Second-Order Gate (SOG) module to fuse information, as shown in Fig. 3. The SOG module selects valid information from the first-order output of the gate mechanism (Eq. 2–Eq. 3) using the initial input vectors, and then performs second-order gating calculations through the gate neuron λ to enhance the efficient propagation of valuable information (Eqs. 4). This outputs the weight of the fused information as second-order, reducing noise and redundancy, improving information acquisition, and accelerating model convergence.

Given the input vectors $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^d$ and the output $\mathbf{h} \in \mathbb{R}^d$, then the SOG fusion is calculated as follows:

$$\lambda = f\left[W_x \cdot \tanh(\mathbf{x}) + W_y \cdot \tanh(\mathbf{y})\right] \tag{1}$$

$$\mathbf{h}_x = \lambda \cdot tanh(\mathbf{x}) + \mathbf{x} \tag{2}$$

$$\mathbf{h}_y = (1-\lambda) \cdot tanh(\mathbf{y}) + \mathbf{y} \tag{3}$$

$$\mathbf{h} = SOG(\mathbf{x}, \mathbf{y}) = \lambda \cdot \mathbf{h}_x + (1-\lambda) \cdot \mathbf{h}_y \tag{4}$$

where $\mathbf{W}_x$ and $\mathbf{W}_y$ are trainable parameters, $f(\cdot)$ denotes the activation function and λ is the gate neuron that controls the Fusion information weighting.

Subsequently, we use the fusion output $\mathbf{c}^{slot} \in \mathbb{R}^{N\times 2d} = \left\{\mathbf{c}_1^{slot}, \mathbf{c}_2^{slot}, .., \mathbf{c}_N^{slot}\right\}$ of the hidden information $\mathbf{H}^c$ and $\mathbf{H}^w$ as the input of the SF task, apply the output $\mathbf{h} \in \mathbb{R}^{N\times 2d} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N\}$ to update the intent information, and use the fusion output $\mathbf{c}^{inte} \in \mathbb{R}^{2d}$ of the representation information $\mathbf{S}^c$ and $\mathbf{S}^w$ as the input of the ID task. The calculation formula is as follows:

$$\mathbf{c}_j^{slot} = SOG(\mathbf{H}_j^c, \mathbf{H}_{f_{align}(j,\mathbf{w})}^w) \tag{5}$$

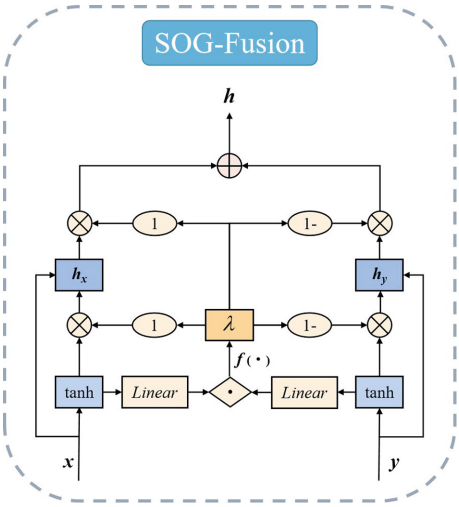


Fig. 3. SOG Fusion module, where $\mathbf{x}$ and $\mathbf{y}$ are fusion input vectors, $f(\cdot)$ denotes the activation function, λ is the gate neuron that controls the weight of the fused information, $\mathbf{h}_x$ and $\mathbf{h}_y$ are the selection information and $\mathbf{h}$ is the fusion output.

$$\mathbf{h}_j = SOG(\mathbf{H}^w_{f_{align}(j,\mathbf{w})}, \mathbf{H}^c_j) \tag{6}$$

$$\mathbf{c}^{inte} = SOG(\mathbf{S}^c, \mathbf{S}^w) \tag{7}$$

$$f_{align}(j, \mathbf{w}) = \begin{cases} 1 & j \leq len(\mathbf{w}_1) \\ \sum_{i=2}^{|\mathbf{w}|} i \cdot \mathbb{I}(\sum_{k=1}^{i-1} len(\mathbf{w}_k) < j \leq \sum_{k=1}^{i} len(\mathbf{w}_k)) & other \end{cases} \tag{8}$$

where $\mathbf{w}$ is the word sequence, $len(\cdot)$ counts the number of characters in a word, $\mathbb{I}(\cdot)$ is the indicator function. $j = \{1, 2, ..., N\}$ is each character's position index.

3.3 SF-ID Interaction Layer

To fully exploit the rich interaction information of intent and slots, we propose the Multi-Scale intent-slots Interaction Module (MSIM), which consists of SF-Update Module, ID-Update Module and Decoder Module. The MSIM module first uses the intent information to update the multi-scale information of slots obtained from the fusion layer and then uses them to guide the previous intent information. Finally, the deep bi-directional interaction between SF and ID is achieved through multiple interactions. A specific interaction is as follows.

SF-Update Module: For the updating of multi-scale slots information, we first obtains the update information $f^{inte} \in \mathbb{R}^{N \times 2d} = \left\{f_1^{inte}, f_2^{inte}, ..., f_N^{inte}\right\}$ by fusing $\mathbf{y}^{inte} \in \mathbb{R}^{2d}$ and $\mathbf{c}^{slot}$ using the SOG module, then the update information f^{inte} is calculated with $\mathbf{c}^{slot}$ to update the multi-scale slots information $\mathbf{y}^{slot} \in \mathbb{R}^{N \times 2d} = \left\{\mathbf{y}_1^{slot}, \mathbf{y}_2^{slot}, ..., \mathbf{y}_N^{slot}\right\}$, which is calculated as follows:

$$f_j^{inte} = SOG\left(\mathbf{c}_j^{slot}, \mathbf{y}^{inte}\right) \tag{9}$$

$$\mathbf{y}_j^{slot} = (w_j^I \cdot f_j^{inte}) \cdot \mathbf{c}_j^{slot} \tag{10}$$

where w_j^I is the trainable parameter and $j = \{1, 2, ..., N\}$ is the position index of each character. In the first cycle of interactions, we define $\mathbf{y}^{inte} = \mathbf{c}^{inte}$.

ID-Update Module: For the updating of intent information, similar to SF-Update module, we first obtains $f^{slot} \in \mathbb{R}^{2d}$ by fusing $\mathbf{y}^{slot}$ and $\mathbf{h}$ using the SOG module, then calculates f^{slot} with $\mathbf{c}^{inte}$ to update the intent information $\mathbf{y}^{inte}$. The calculation is as follows:

$$f^{slot} = \sum_{j=1}^{N} SOG(\mathbf{y}_j^{slot}, \mathbf{h}_j) \tag{11}$$

$$\mathbf{y}^{inte} = f^{slot} + \mathbf{c}^{inte} \tag{12}$$

Decoder Module: After the cyclic interaction, we decode the final information $\mathbf{y}^{slot}$ and $\mathbf{y}^{inte}$ by the Decoder module to obtain the final multi-scale slots output $\mathbf{O}^{slot} = \left\{\mathbf{O}_1^{slot}, \mathbf{O}_2^{slot}, ..., \mathbf{O}_N^{slot}\right\}$ and intent output $\mathbf{O}^{inte}$, which is calculated as follows.

$$P(\tilde{\mathbf{y}}^{slot} = j|\mathbf{c}^{slot}) = softmax[w^{S\text{-}O} \cdot (\mathbf{h}_N \oplus \mathbf{y}_j^{slot})] \tag{13}$$

$$P(\tilde{\mathbf{y}}^{inte}|\mathbf{c}^{inte}) = softmax[w^{I\text{-}O} \cdot \mathbf{y}^{inte}] \tag{14}$$

$$\mathbf{O}_j^{slot} = argmax\ [P\left(\tilde{\mathbf{y}}^{slot} = j \middle| \mathbf{c}^{slot}\right)] \tag{15}$$

$$\mathbf{O}^{inte} = argmax\ [P(\tilde{\mathbf{y}}^{inte}|\mathbf{c}^{inte})] \tag{16}$$

where $w^{S\text{-}O}$ and $w^{I\text{-}O}$ are trainable parameters, $\oplus$ denotes concatenation operation, $j = \{1, 2, ..., N\}$ is the position index of each character.

3.4 Joint Loss Function

According to Goo et al. [1], a joint training scheme with NLLOSS is used for optimization in this paper, and the joint loss function is calculated as follows:

$$\mathcal{L} = -logP\left(\hat{\mathbf{y}}^{inte} \middle| \mathbf{c}^{inte}\right) - \sum_{i=1}^{N} logP(\hat{\mathbf{y}}_i^{slot}|\mathbf{c}_i^{slot}) \tag{17}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct experiments on two public Chinese SLU datasets, CAIS [15] and SMP-ECDT [21], to evaluate the model validity. The CAIS dataset contains

7,995 training sets, 994 validation sets, and 1,024 test sets. The SMP-ECDT dataset contains 1,832 training sets, 352 validation sets, and 395 test sets.

In this paper, we use F1 score and accuracy to evaluate the accuracy of SF and ID, respectively. Moreover, we use sentence-level semantic accuracy to indicate that the output of this utterance is considered as a correct prediction when and only when the intent and all slots are perfectly matched.

4.2 Experimental Setting

In this paper, we set the dropout rate to 0.5, the initial learning rate is set to 0.001, the learning rate is adjusted dynamically by using the warmup strategy [8], and the Adam optimizer [5] is used to optimize the parameters of the model. For the CAIS dataset, we set the number_cycles to 3, while the SMP-ECDT dataset, we set it to 4. The model is trained on a Linux system using the PyTorch framework and Tesla A100, and multiple experiments are conducted with different random seeds to select the model parameter for evaluation on the test dataset that perform the best on the validation dataset.

4.3 Baseline Models

In this section, we select the following models for comparison, which are Slot-Gated Full Atten [1], a slot-oriented gating model to improve the semantic accuracy; CM-Net [9], a collaborative memory network to augment the local contextual representation; Stack-Propagation [12], a stack-propagation model to capture semantic knowledge of intent; MLWA [15], a multi-level word adapter that fuses word information with character information; GAIR [21], a two-stage Graph Attention Interaction Refine framework that leverages SF and ID information.

On the CAIS dataset, we uses the model performance from the paper GAIR [21]. On the SMP-ECDT dataset, we compare the published code of the model by running experiments separately, while the CM-Net [9] cannot be compared with this model due to the fact that the official codes are not provided.

4.4 Main Results

Table 1 presents the experimental results of the proposed MSIM-SOG model and the baseline models on the CAIS and SMP-ECDT datasets. From the analysis of the experimental results, we give the following experimental conclusions.

1. The MSIM-SOG model proposed in this paper outperforms the above model in all metrics and achieves state-of-the-art performance
2. Compared with the baseline model MLWA [15], our model achieves larger improvements. In detail, on the CAIS and SMP-ECDT datasets, our model improved Slot F1 Score by 1.66% and 2.95%, Intent Acc by 0.49% and 1.58%, and Semantic Acc by 1.18% and 3.78%, respectively. These results indicate that our model effectively fuses character and word information and enhances performance through a deep bi-directional interaction between ID and SF.

Table 1. The main results of the above models on the CAIS and SMP-ECDT datasets. The numbers with * indicate that the improvement of the model in this paper is statistically significant at all baselines, with $p < 0.05$.

Model	CAIS dataset			SMP-ECDT dataset		
	Slot F1 Score	Intent Acc	Semantic Acc	Slot F1 Score	Intent Acc	Semantic Acc
Slot-Gated Full Atten [1]	81.13	94.37	80.83	60.91	86.02	53.75
CM-Net [9]	86.16	94.56	–	–	–	–
Stack-Propagation [12]	87.64	94.37	84.68	71.32	91.06	63.75
MLWA [15]	88.61	95.16	86.17	75.76	94.65	68.58
GAIR [21]	88.92	95.45	86.86	77.68	95.45	69.75
MSIM-SOG	**90.27***	**95.65***	**87.35***	**78.71***	**96.23***	**72.36***

3. Compared with the current SOTA model GAIR [21], our model improved Slot F1 Score by 1.35% and 1.03%, Intent Acc by 0.20% and 0.78%, and Semantic Acc by 0.49% and 2.61% on the CAIS and SMP-ECDT datasets, respectively. These results show that our model, when utilizing a suitable CWS system and incorporating character information, outperforms the GAIR [21] model without CWS system.

The aforementioned outcomes demonstrate the advancement of the MSIM-SOG model proposed in this paper. We attribute these results to the following reasons: (1) The SOG module effectively fuses word and character information, enhancing model accuracy. (2) The deep interaction of ID and SF in MSIM improves performance by selecting effective multi-scale slots and intent information. (3) The use of a suitable CWS system and character information prevents incorrect slot identification and predictions.

4.5 Ablation Study

In this section, we conducted an ablation study to investigate the impact of the MSIM and SOG module on the performance enhancement of the MSIM-SOG model. We analyzed the effects by ablating four important modules and employing different approaches in the experiment (Table 2).

Table 2. Main results of ablation experiments on CAIS and SMP-ECDT datasets.

Model	CAIS dataset			SMP-ECDT dataset		
	Slot F1 Score	Intent Acc	Semantic Acc	Slot F1 Score	Intent Acc	Semantic Acc
MSIM w/o joint learning	87.61 (↓ 2.66)	94.56 (↓ 1.09)	85.27 (↓ 2.08)	76.83 (↓ 1.88)	95.02 (↓ 1.21)	70.75 (↓ 1.61)
MSIM w/o intent→ slot	88.75 (↓ 1.52)	95.35 (↓ 0.30)	86.61 (↓ 0.74)	77.69 (↓ 1.02)	95.67 (↓ 0.56)	71.79 (↓ 0.57)
MSIM w/o slot→ intent	89.69 (↓ 0.58)	95.15 (↓ 0.50)	86.75 (↓ 0.60)	78.35 (↓ 0.36)	95.47 (↓ 0.76)	71.68 (↓ 0.68)
MSIM-SOG w/o SOG	88.96 (↓ 1.31)	94.95 (↓ 0.70)	86.36 (↓ 0.99)	77.61 (↓ 1.10)	95.35 (↓ 0.88)	71.08 (↓ 1.28)
MSIM-SOG	**90.27**	**95.65**	**87.35**	**78.71**	**96.23**	**72.36**

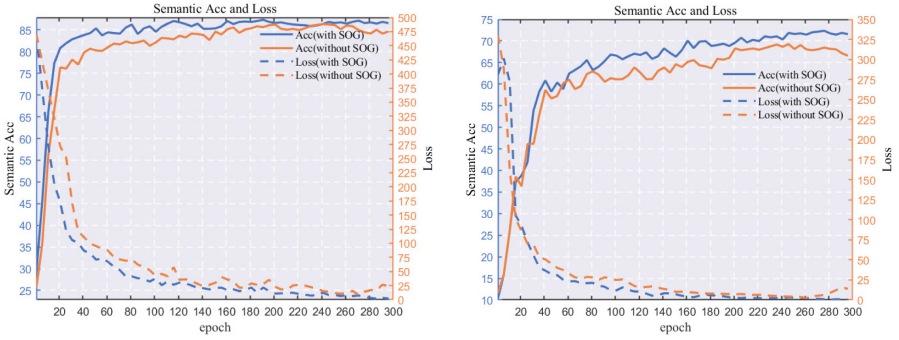
Effect on MSIM: To demonstrate the advancement of the MSIM module, we first ablated the joint learning strategy, directly feeding intent and slot information from the fusion layer into the decoder. The experimental results indicated a

significant drop in performance on both datasets compared to the original model, due to the lack of explicit interaction between intent and slot information. Subsequently, we conducted an ablation study on the unidirectional interaction of intent and slot, removing the SF-Update Module and ID-Update Module separately. The results indicated that the unidirectional interaction model had higher accuracy than the model without joint learning, but it performed significantly worse than the MSIM with deep bi-directional interaction. This confirms the mutual enhancement of multi-scale slots and intent information through deep interaction, aligning with previous studies.

Effect on SOG: To verify the advancement of the SOG module, we remove the SOG module and use MLWA [15] instead. The aforementioned experimental results demonstrate that the performance of SF and ID both decreased significantly. This indicates that the SOG module has a significant contribution in improving information acquisition, reducing the impact of model noise information and improving the learning ability of the model.

4.6 Convergence Analysis

To analyze the contribution of the SOG module in accelerating model convergence and reducing overfitting, we compared the semantic accuracy and loss curves of the model with and without the SOG module (replaced by MLWA [15]) after 300 epochs of training on the test set, as shown in Fig. 4a and Fig. 4b.



(a) On CAIS dataset (b) On SMP-ECDT dataset

Fig. 4. Semantic Acc and Loss Overall on CAIS and SMP-ECDT Dataset.

The results in Fig. 4a and Fig. 4b demonstrate that the model with the SOG module achieved convergence at around 117 and 160 epochs on the CAIS and SMP-ECDT datasets respectively, while the model without the SOG module reached convergence at 170 and 200 epochs. This indicates that the SOG module effectively accelerates model convergence and improves accuracy. On the loss curve, the model with the SOG module maintains relatively stable loss

after 200 epochs of training, whereas the model without the SOG module shows an increase in loss after 270 epochs on the CAIS dataset and 280 epochs on the SMP-ECDT dataset, suggesting that the SOG module effectively alleviates model overfitting. These results highlight the effectiveness of the SOG module in accelerating convergence and reducing overfitting.

4.7 Effect of Interation

To assess the impact of deep interactions between ID and SF in the MSIM model, we evaluated its performance with different depths of interaction levels on the CAIS and SMP-ECDT datasets using Semantic Acc.

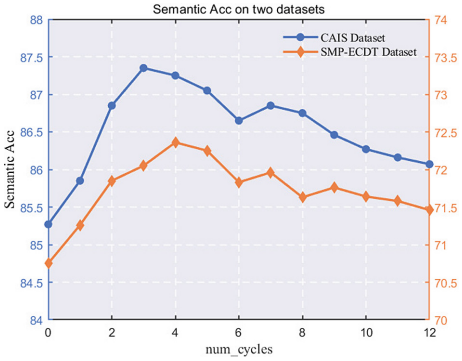


Fig. 5. Semantic Acc of MSIM with varying interaction levels on two datasets.

The impact of deep interactions between ID and SF in the MSIM model on performance was studied. According to the results in Fig. 5, when $num\_cycles = 0$, there is no explicit joint learning and no interaction between intent and slot information. The results indicated that as the number of interactions increased, Semantic Acc gradually improved. The CAIS dataset achieved the best performance when $num\_cycles = 3$, while the SMP-ECDT dataset achieved its best when $num\_cycles = 4$. This demonstrates the effectiveness of deep interaction between SF and ID. Increasing interactions strengthened the connection between SF and ID, resulting in performance improvement. Although there was a slight decrease in Semantic Acc beyond a certain depth of interaction, all models with interactions outperformed the model without interactions. These findings emphasize the significance of deep interaction between ID and SF in enhancing model performance and validating the mutual reinforcement of SF and ID tasks.

5 Conclusion and Future Work

This paper introduces the MSIM-SOG model to address the challenges of fusing Chinese word and character information in the Chinese SLU domain while studying the deep interaction between ID and SF. The model consists of two modules:

MSIM enables deep bi-directional interaction between ID and SF by updating multi-scale slots and intent information cyclically. The SOG module enhances fusion by selecting the first-order gate output and performing second-order gating calculation. Experimental results on Chinese SLU datasets demonstrate significant performance improvement compared to existing models, achieving state-of-the-art results. Future work includes applying the MSIM-SOG model to multi-intent Chinese datasets to assess its generalization ability, as well as exploring the applicability of the SOG fusion mechanism in other NLP tasks such as sentiment analysis, recommendation systems, and semantic segmentation.

Acknowledgements. This work was supported in part by the National Science Foundation of China under Grant 62172111, in part by the Natural Science Foundation of Guangdong Province under Grant 2019A1515011056, in part by the Key technology project of Shunde District under Grant 2130218003002.

References

1. Goo, C.W., Gao, G., Hsu, Y.K., Huo, C.L., Chen, T.C.: Slot-gated modeling for joint slot filling and intent prediction. In: NAACL, pp. 753–757 (2018)
2. Haffner, P., Tür, G., Wright, J.H.: Optimizing SVMs for complex call classification. In: ICASSP (2003)
3. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
4. Kim, S., D'Haro, L.F., Banchs, R.E., Williams, J.D., Henderson, M.: The fourth dialog state tracking challenge. In: Jokinen, K., Wilcock, G. (eds.) Dialogues with Social Robots. LNEE, vol. 999, pp. 435–449. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-2585-3_36
5. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR, pp. 1–11 (2014)
6. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: AAAI 2015, pp. 2267–2273 (2015)
7. Li, C., Zhou, Y., Chao, G., Chu, D.: Understanding users' requirements precisely: a double bi-LSTM-CRF joint model for detecting user's intentions and slot tags. Neural Comput. Appl. **34**, 13639–13648 (2022)
8. Liu, L., et al.: On the variance of the adaptive learning rate and beyond. ArXiv, pp. 1–13 (2019)
9. Liu, Y., Meng, F., Zhang, J., Zhou, J.: CM-net: a novel collaborative memory network for spoken language understanding. In: EMNLP-ICJNLP, pp. 1051–1060
10. Ma, Z., Sun, B., Li, S.: A two-stage selective fusion framework for joint intent detection and slot filling. IEEE Trans. Neural Netw. Learn. 1–12 (2022)
11. Ni, P., Li, Y., Li, G., Chang, V.I.: Natural language understanding approaches based on joint task of intent detection and slot filling for IoT voice interaction. Neural Comput. Appl. 1–18 (2020)
12. Qin, L., Che, W., Li, Y., Wen, H., Liu, T.: A stack-propagation framework with token-level intent detection for spoken language understanding. In: EMNLP-IJCNLP, pp. 2078–2087 (2019)
13. 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 (2021)

14. Tang, H., Ji, D.H., Zhou, Q.: End-to-end masked graph-based CRF for joint slot filling and intent detection. Neurocomputing **413**, 348–359 (2020)
15. Teng, D., Qin, L., Che, W., Liu, T.: Injecting word information with multi-level word adapter for Chinese spoken language understanding. In: ICASSP, pp. 8188–8192 (2021)
16. Wei, P., Zeng, B., Liao, W.: Joint intent detection and slot filling with wheel-graph attention networks. J. Intell. Fuzzy Syst. **42**, 2409–2420 (2021)
17. Weld, H., Huang, X., Long, S., Poon, J., Han, S.C.: A survey of joint intent detection and slot filling models in natural language understanding. ACM Comput. Surv. **55**, 1–38 (2021)
18. Xu, C., Li, Q., Zhang, D., Cui, J.: A model with length-variable attention for spoken language understanding. Neurocomputing **379**, 197–202 (2020)
19. Xu, P., Sarikaya, R.: Convolutional neural network based triangular CRF for joint intent detection and slot filling. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 78–83 (2013)
20. Yao, K., Peng, B., Zhang, Y., Yu, D., Zweig, G.: Spoken language understanding using long short-term memory neural networks. In: IEEE-SLT, pp. 189–194 (2014)
21. 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: ICASSP, pp. 7617–7621 (2022)