SENTIMENT INJECTED ITERATIVELY CO-INTERACTIVE NETWORK FOR SPOKEN LANGUAGE UNDERSTANDING

Zhiqi Huang[1], *Fenglin Liu*[1], *Peilin Zhou*[1], *Yuexian Zou*[1,2,*]

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Spoken Language Understanding (SLU) is an essential part of the spoken dialogue system, which typically consists of intent detection (ID) and slot filling (SF) tasks. During the conversation, most utterances of people contain rich sentimental information, which is helpful for performing the ID and SF tasks but ignored to be explored by existing works. In this paper, we argue that implicitly introducing sentimental features can promote SLU performance. Specifically, we present a Multi-task Learning (MTL) framework to implicitly extract and utilize the aspect-based sentimental text features. Besides, we introduce an Iteratively Co-Interactive Network (ICN) for the SLU task to fully utilize the comprehensive text features. Experimental results show that with the external BERT representation, our framework achieves new state-of-the-art on two benchmark datasets, i.e., SNIPS and ATIS.

Index Terms— Spoken Language Understanding, Multi-task Learning, Iteratively Co-Interactive Network, BERT, Aspect-based Sentiment Information

1. INTRODUCTION

Spoken Language Understanding (SLU) is a critical component in task-oriented dialogue systems and has been widely exploited. It typically involves intent detection (ID) and slot filling (SF) tasks. Examples of the SLU task are shown in Table 1, given an utterance: *"reserve for highly rated restaurant in seychelles"* from the SNIPS dataset, there are different slot labels for each token and an intent for the whole utterance.

In the spoken dialogue system, we find that some utterances contain rich sentimental information, and more importantly, such sentimental information is highly relevant to the SLU task [1, 2, 3, 4, 5, 6, 7, 8]. For example, as can be seen in Table 1, the utterances sampled from the SNIPS dataset contain sentimental related tokens (i.e., *highly rated* and *most popular*) which are also the key slots in the SF task. In addition, as for the End-to-End Aspect-based Sentiment Analysis (E2E-ABSA) task [9, 10], such tokens are usually treated as users sentiment. For example, *highly rated* and *most popular* are treated as positive sentiment in the E2E-ABSA task. Intuitively, the acquisition of sentimental information is effective for identifying some special slots with sentimental tendencies. In this paper, we argue that the SLU system can perform better if we can effectively extract and utilize the sentimental information as additional features. Specifically, we explore the effect of external aspect-based sentimental features on the SLU system by introducing an MTL framework, aiming to extract implicit information from the E2E-ABSA task.

Table 1: Utterances with intents and slots annotation (BIO format) sampled from the SNIPS dataset.

Utter.1	reserve	for	highly	rated	restaurant	in	seychelles
Senti.	O	O	$B-POS$	$I-POS$	O	O	O
Slots	O	O	$B-sort$	$I-sort$	$B-restaurnt\_type$	O	$B-country$
Intent	$BookRestaurant$						
Utter.2	play	the	most	popular	album	on	GoogleMusic
Senti.	O	O	$B-POS$	$I-POS$	O	O	O
Slots	O	O	$B-sort$	$I-sort$	$B-music\_item$	O	$B-service$
Intent	$PlayMusic$						

To make better use of the enriched representation, we introduce an Iteratively Co-Interactive Network (ICN) for ID and SF tasks, which can explicitly model relation and interaction between the two tasks in the decoder stage with co-interactive layers, aiming to force the two tasks to make full use of the decoder information from the other one. We conduct experiments on two benchmark datasets including SNIPS [11] and ATIS [12]. The results show the effectiveness of our framework by outperforming many existing methods by a large margin. Moreover, BERT [13], a pre-trained language model, is employed to further boost the performance of our ICN and our proposed MTL framework, which outperforms the state-of-the-art model on SNIPS and ATIS by 0.7% and 1.6% in terms of accuracy on ID task, 0.6% and 0.3% in terms of F1 score on SF task, respectively.

To summarize, our contributions are as follows: 1) We propose an MTL framework to extract sentimental information as external features for the SLU model, which outperforms the existing SLU models; 2) We introduce an ICN for the SLU task, which simultaneously models the relation and interaction within SLU task in an explicit way; 3) Experimental results show that our proposed framework achieves signifi-

This paper is partially supported by Shenzhen Municipal Research Funding Project "Research on Key Technologies of Spoken Language Understanding via Representation Learning and Knowledge Assisting". Special acknowledgements are given to AOTO-PKUSZ Joint Lab for its support.

* Corresponding author: zouyx@pku.edu.cn

cant and consistent improvement on two benchmark datasets.

2. APPROACH

As can be seen in Figure 1, the MTL framework is employed on the two modules: ICN and E2E-ABSA. In this section, we first talk about the detail of ICN, then we discuss and show the difference between our MTL strategy and others.

2.1. Iteratively Co-Interactive Network

Considering the traditional jointly training methods just model the relationship between ID and SF by sharing parameters, [14] proposed to use the joint model with stack-propagation framework, achieving significant results on the SLU task. However, during forward propagation, only the feature of ID is sent to SF, which reduces the interaction of the two tasks. Thus, we proposed to model SLU task with Iteratively Co-Interactive Layer (ICL) to mitigate such shortcoming. In this way, the ICN consists of Self-Attentive Encoder and the ICL.

2.1.1. Input Layer

We employ word embedding $\mathbf{emb}_i^{word}$ to the input sentence as the model input layer. Besides, we explore $\mathbf{emb}_i^{MTL}$ and $\mathbf{emb}_i^{BERT}$ to boost the performance of the baseline model. Formally, the input representation of the i^{th} token is:

$$\mathbf{x}_i = \mathbf{emb}_i^{word} \oplus \mathbf{emb}_i^{MTL} \oplus \mathbf{emb}_i^{BERT}$$

where $\oplus$ is concatenation operator. The $\mathbf{emb}_i^{MTL}$ depends on whether we use the proposed MTL framework or not. The $\mathbf{emb}_i^{BERT}$ depends on whether we use BERT representation.

2.1.2. Self-Attentive Encoder

For easier comparison of effects, we follow [14] where the ID and SF tasks share the same Self-Attentive encoder. We use the Bi-LSTM [15] to capture the temporal features $\mathbf{H}$ and self-attention mechanism [16] to extract the contextual information $\mathbf{C}$. Then, we concatenate these two representations $\mathbf{H} \oplus \mathbf{C}$ as the final encoding representation $\mathbf{E}$.[1]

2.1.3. Iteratively Co-Interactive Layer

Intent/Slot Decoder. We use unidirectional LSTM as the *Intent Decoder L* and *Slot Decoder L*. The hidden state and the output vector at each decoding time step t is calculated as:

$$\mathbf{h}_t^i = \begin{cases} LSTM\left(\mathbf{h}_{t-1}^i, \mathbf{y}_{t-1}^i \oplus \mathbf{e}_t\right), & L = 1 \\ LSTM\left(\mathbf{h}_{t-1}^i, \mathbf{y}_{t-1}^i \oplus \mathbf{y}_t^j \oplus \mathbf{e}_t\right), & L > 1 \end{cases}$$
$$\mathbf{y}_t^i = \text{ softmax}\left(\mathbf{W}_h^i \mathbf{h}_t^i\right), \quad L >= 1$$

where $(i, j) \in \{(slot, intent), (intent, slot)\}$, $\mathbf{h}_{t-1}^i$ is hidden state in the previous time step i, $\mathbf{e}_t$ is the aligned encoder hidden state, $\mathbf{y}_t^i$ is the i decoder output distribution of the t^{th} token in the utterance, $\mathbf{y}_t^j$ is the output of j decoder layer, $\mathbf{W}_h^i$ are trainable parameters of the model, and $\mathbf{y}_{t-1}^i$ is the previously predicted i decoder output distribution.

Label Prediction. Finally, the label of the t^{th} token in the utterance are predicted by: $o_t^i = \operatorname{argmax}\left(\mathbf{y}_t^i\right)$, where $i \in \{intent, slot\}$, $\mathbf{y}_t^i$ is the i output distribution of the t^{th} token in the utterance, o_t^i represents the i label of the t^{th} token in the utterance. The final intent o^I is generated by voting from all token intent results [14]: $o^I = \operatorname{argmax} \sum_{t=1}^{T} \sum_{j=1}^{N_I} \alpha_j \mathbb{1}\left[o_t^I = j\right]$, where T is the utterance length, N_I is the number of intent labels, α_j denotes a 0-1 vector $\alpha \in \mathbb{R}^N$ of which the j^{th} unit is 1 and the others are 0, $argmax$ returns the indices of the maximum values in α.

2.1.4. Multi-Level Supervision

Previous joint model [21] shows that when there are too many decoder layers, only little available information may be obtained at the bottom, which makes the interaction at the bottom of the model insufficient and not conducive to the task. Thus, we propose to use the multi-level supervision training method in our ICN. Specifically, as can be seen in the Figure 1b, instead of calculating the cross entropy loss only in the last layer, we give each decoder supervision by calculating loss at each decoder layer. In this way, the exchanged information is explicit during the interaction and each layer can generate the output of the task, which cannot be done by the previous models; 3) In the inference phase, we can further propose various inference methods including, directly using the output of the last layer as the result, using the average or max-pooling value of each layer as the result.

The ID objective in the i^{th} decoder layer is defined as:

$$\mathcal{L}_i^I = -\sum_{t=1}^{T} \sum_{j=1}^{N_I} \hat{\mathbf{y}}_t^{j,I} \log\left(\mathbf{y}_t^{j,I}\right)$$

where T is the utterance length, N_I is the number of intent labels, $\hat{\mathbf{y}}_t^{j,I}$ is the gold intent label. We get the SF objective $\mathcal{L}_i^S$ in the similar way.

Thus, given the number of iterative decoder layers L, the final objective is computed as: $\mathcal{L} = \sum_{i=1}^{L}\left(\mathcal{L}_i^I + \mathcal{L}_i^S\right)$.

2.2. E2E-ABSA Module

Inspired by [9], our E2E-ABSA module consists of the self-attentive encoder and E2E-ABSA layer from which we extract features, as can be seen in Figure 1a. Empirically we employ a unidirectional LSTM as the E2E-ABSA layer.

2.3. Implementations of MTL

There are many MTL frameworks to train tasks simultaneously. We aim to utilize them to effectively exploit the sentimental representation to improve the SLU performance. As can be seen in Figure 1e, compared to other MTL methods, we regard E2E-ABSA as the auxiliary task and extract

[1] For more details of the Self-Attentive Encoder, please refer to [14].

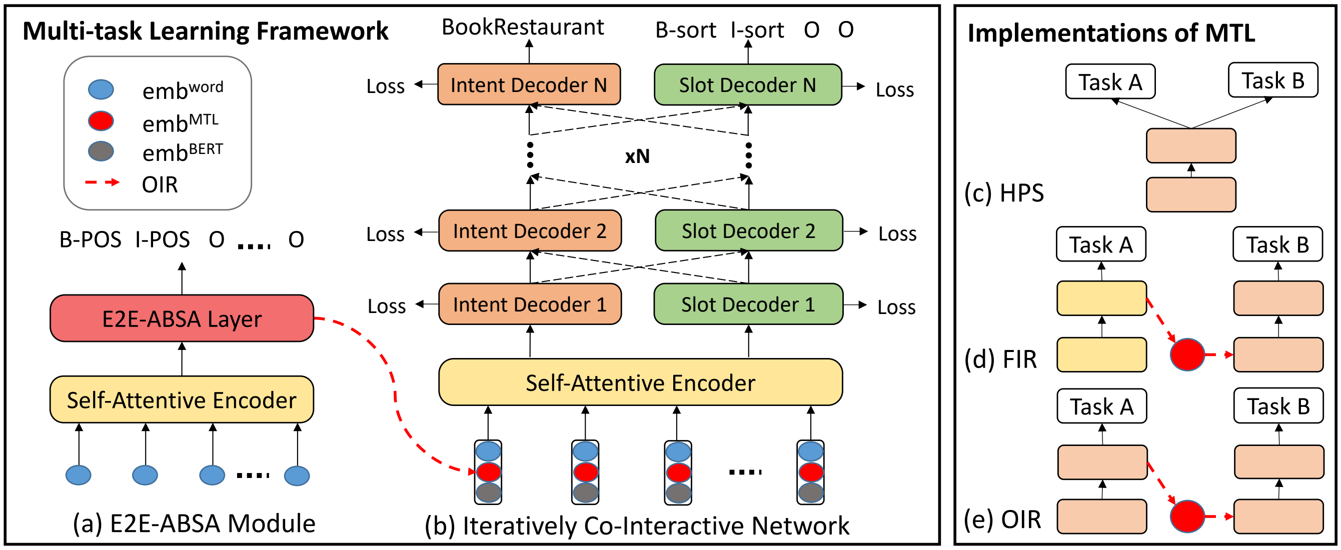

Fig. 1: ab) Architecture of the proposed MTL framework with the E2E-ABSA module and the ICN; cde) Illustration of different MTL implementations. Task A and Task B stand for the E2E-ABSA and the SLU task.

Table 2: Performance of different models and frameworks on the SNIPS and ATIS datasets.

Model	**SNIPS**			**ATIS**		
	Slot (*F1*)	Intent (*Acc*)	Overall (*Acc*)	Slot (*F1*)	Intent (*Acc*)	Overall (*Acc*)
Joint Seq (INTERSPEECH 2016) [17]	87.3	96.9	73.2	94.3	92.6	80.7
Attention BiRNN (SIGDIAL 2016) [2]	87.8	96.7	74.1	94.2	91.1	81.9
Slot-Gated Full Atten (NAACL 2018) [18]	88.8	97.0	75.5	94.8	93.6	82.2
Self-Attentive Model (EMNLP 2018) [19]	90.0	97.5	81.0	95.1	96.8	82.2
Bi-Model (NAACL 2018) [20]	93.5	97.2	83.8	95.5	96.4	85.7
SF-ID Network (ACL 2019) [21]	90.5	97.0	78.4	95.6	96.6	86.0
Joint BERT (arXiv 2019) [22]	97.0	98.6	92.8	96.1	97.5	88.2
Baseline (Stack-Propagation)	94.2	98.0	86.9	95.9	96.9	86.5
ICN	94.5	99.1	88.0	95.9	97.2	87.1
ICN + MTL (HPS)	95.9	99.1	89.7	96.1	97.6	87.9
ICN + MTL (FIR)	96.0	99.0	89.4	96.0	97.6	87.6
ICN + MTL (OIR)	96.2	99.1	90.2	96.2	98.0	88.2
Full model: ICN + MTL (OIR) + BERT	**97.6**	**99.3**	**93.0**	**96.4**	**99.1**	**88.5**

the hidden output of it as the sentimental representation $\boldsymbol{emb}^{MTL}$ which is fed into the input layer of the ICN to be concatenated with the original SLU input. The details of integration can be concluded as follows:

Hard parameter sharing (HPS) is the most commonly used approach to MTL in neural networks. It is applied by sharing the embeddings of input sentences and self-attentive encoder between the E2E-ABSA module and the ICN.

Fixed implicit representations (FIR) means extracting the outputs of the E2E-ABSA layer from a fixed pre-trained model, then train the ICN with the extracted representations.

Online implicit representation (OIR) requires to give E2E-ABSA module an initialization firstly by train it with the E2E-ABSA dataset, then the parameters of both E2E-ABSA module and ICN are updated for the same SLU objective.

3. EXPERIMENTS

3.1. SLU Datasets and E2E-ABSA Dataset

To evaluate the efficiency of our proposed MTL framework, we conduct experiments on two benchmark datasets, the widely used ATIS dataset [12] and custom-intent-engine dataset called the SNIPS [11], which is collected by snips personal voice assistant. Both evaluated datasets used in our paper follow the same format and partition as in [14], and three evaluation metrics are used to measure the performance of our proposed model. Following the setup in [9], in the E2E-ABSA module, we employ a commonly used review dataset, LAPTOP, from the laptop domain in SemEval ABSA challenge 2014 [23] but re-prepared in [24]. The statistics of the datasets can be referred in [24, 14].

Table 3: Employ MTL framework to the SLU task with (w/) and without (w/o) the ICN. Δ stands for relative value where we use the results w/ ICN to subtract the results w/o ICN.

Δ = w/ ICN - w/o ICN	SNIPS		ATIS	
	Slot (Δ)	Intent (Δ)	Slot (Δ)	Intent (Δ)
Baseline	+0.3	+0.9	+0	+0.3
+ MTL (HPS)	+0.2	+1.2	+0.3	+0.3
+ MTL (FIR)	+0.3	+1.1	+0.2	+0.4
+ MTL (OIR)	+0.4	+1.2	+0.2	+0.6

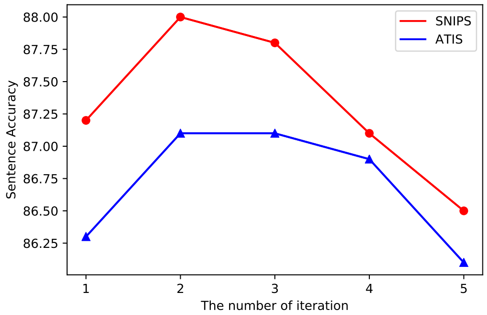

Fig. 2: Effect of the iteration number of decoder layer on two datasets on ICN with multi-level supervision setting.

3.2. Settings

We adopt the Adam optimizer for optimizing the parameters, with a mini-batch size of 16 and initial learning rate of 0.001. The word embedding dimensionality d is set as 128 and 256 for the SNIPS and ATIS, respectively. According to the performance on the validation set, we consistently set 2 as the iteration number of the decoder layer in the SLU module unless otherwise specified.

3.3. Main Results

Table 2 shows the experimental results of our approach on the SNIPS and ATIS. We can see that the improvement of the SLU task is mainly due to the ICN and MTL. Compare with the baseline, where we employ the stack-propagation framework with the self-attentive encoder, the ICN and all the MTL implementations based on ICN have improvements on most of or all the metrics, while our proposed OIR brings the most, which confirms the effectiveness of the ICN and indicates that the implicit sentimental information has the potential to improve the ID and SF tasks. With BERT, our method achieves the best results on SLU task.

3.4. Effect of the MTL and ICN

Table 4: Experiment on whether to use multi-level supervision or not with 2 and 3 interative decoder layers.

SNIPS	Slot (*F1*)	Intent (*Acc*)	Overall (*Acc*)
L=2, ICN	94.5	99.1	88.0
L=2, ICN w/o multi.	94.2	98.8	87.7
L=3, ICN	94.2	99.0	87.8
L=3, ICN w/o multi.	93.8	98.5	87.3

Table 5: Effect of using BERT as external features for SLU.

Δ = w/ BERT - w/o BERT	SNIPS		ATIS	
	Slot (Δ)	Intent (Δ)	Slot (Δ)	Intent (Δ)
ICN	+0.1	+0.2	+0	+0.2
+ MTL (HPS)	+0.9	+0.2	+0.1	+0.1
+ MTL (FIR)	+1.2	+0.1	+0.3	+0.2
+ MTL (OIR)	+1.4	+0.2	+0.2	+0.2

To explore the effectiveness of the ICN and MTL further, we show the resulting variation after removing the ICN in Table 3 on different MTL implementations. The baseline model employs only one decoder layer to predict the intents and slots without any MTL framework. We can see that:1) after employing the ICN, most of the tasks have improvement; 2) nearly all the metrics have some promotions with the MTL framework, and the proposed OIR seems to be a bit better than the HPS and FIR. Thus, the ICN has the potential to improve the SLU task through better utilizing the implicit aspect-based sentimental features, especially under our proposed MTL framework. And the ICN can enhance the information interaction capabilities of ID and SF so as to make fuller use of the extracted enriched representations.

3.5. Effect of the iteration number

Selecting a proper iteration number of ICL is important to the SLU task. We explore the best number of decoder layers under the multi-level supervision setting without utilizing the sentimental information. Sentence accuracy is applied as the performance measure because it can reflect the model ability from the overall perspective. As can be seen in Figure 2, the two lines quickly reach the top when the iteration number is 2, then they decrease gradually.

3.6. Effect of multi-level supervision

In order to further explore the architecture of ICN, we conduct experiments to see the model performance without employing the multi-level supervision method. From Table 4, we can see that, for the model with more than one decoder layer, the multi-level supervision training method is useful.

3.7. Effect of BERT

As a successful pre-trained model, Bidirectional Encoder Representation from Transformer [13, BERT] is employed to boost the performance of our proposed MTL framework by providing additional input features to the ICN. From Table 5, all model settings have a certain degree of performance improvement with BERT, and the model trained using OIR with BERT performs remarkably well on both two datasets. We attribute this to the fact that BERT can provide great feature representations which may contain a wealth of sentimental features that are demonstrated to be helpful for SLU.

4. CONCLUSION

In this paper, we propose an MTL framework for the SLU task. Our approach achieves the improvement by considering the implicit sentimental information in the SLU task. Specifically, we design a novel ICN to model the relationship between ID and SF tasks. Our proposed MTL approach is a generic framework for leveraging sentimental information. It is also extensible and can be adapted to promote the performance of other NLP tasks with minimum modifications to model implementations. Detail experiments and analysis are performed on SNIPS and ATIS datasets to demonstrate the effectiveness of our MTL framework and the ICN.

5. REFERENCES

[1] Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig, and Yangyang Shi, "Spoken language understanding using long short-term memory neural networks," in *SLT*, 2014.

[2] Bing Liu and Ian Lane, "Joint online spoken language understanding and language modeling with recurrent neural networks," in *SIGDIAL*, 2016.

[3] Chenwei Zhang, Wei Fan, Nan Du, and Philip S. Yu, "Mining user intentions from medical queries: A neural network based heterogeneous jointly modeling approach," in *WWW*, 2016.

[4] Yun-Nung Chen, Dilek Hakkani-Tür, Gökhan Tür, Jianfeng Gao, and Li Deng, "End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding," in *INTERSPEECH*, 2016.

[5] Baolin Peng, Kaisheng Yao, Jing Li, and Kam-Fai Wong, "Recurrent neural networks with external memory for spoken language understanding," in *NLPCC*, 2015.

[6] Ngoc Thang Vu, "Sequential convolutional neural networks for slot filling in spoken language understanding," in *INTERSPEECH*, 2016.

[7] Chenwei Zhang, Nan Du, Wei Fan, Yaliang Li, Chun-Ta Lu, and Philip S. Yu, "Bringing semantic structures to user intent detection in online medical queries," in *BigData*, 2017.

[8] Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu, "Zero-shot user intent detection via capsule neural networks," in *EMNLP*, 2018.

[9] Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam, "Exploiting BERT for end-to-end aspect-based sentiment analysis," in *EMNLP*, 2019.

[10] Zheng Li, Xin Li, Ying Wei, Lidong Bing, Yu Zhang, and Qiang Yang, "Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning," in *EMNLP/IJCNLP*, 2019.

[11] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *CoRR*, vol. abs/1805.10190, 2018.

[12] Charles T. Hemphill, John J. Godfrey, and George R. Doddington, "The ATIS spoken language systems pilot corpus," in *HLT*, 1990.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[14] Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu, "A stack-propagation framework with token-level intent detection for spoken language understanding," in *EMNLP/IJCNLP*, 2019.

[15] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, 1997.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[17] Dilek Hakkani-Tür, Gökhan Tür, Asli **c**Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang, "Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM," in *INTERSPEECH*, 2016.

[18] Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen, "Slot-gated modeling for joint slot filling and intent prediction," in *NAACL-HLT*, 2018.

[19] Changliang Li, Liang Li, and Ji Qi, "A self-attentive model with gate mechanism for spoken language understanding," in *EMNLP*, 2018.

[20] Yu Wang, Yilin Shen, and Hongxia Jin, "A bi-model based RNN semantic frame parsing model for intent detection and slot filling," in *NAACL-HLT*, 2018.

[21] Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song, "A novel bi-directional interrelated model for joint intent detection and slot filling," in *ACL*, 2019.

[22] Qian Chen, Zhu Zhuo, and Wen Wang, "BERT for joint intent classification and slot filling," *CoRR*, vol. abs/1902.10909, 2019.

[23] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar, "Semeval-2014 task 4: Aspect based sentiment analysis," in *SemEval@COLING*, 2014.

[24] Xin Li, Lidong Bing, Piji Li, and Wai Lam, "A unified model for opinion target extraction and target sentiment prediction," in *AAAI*, 2019.