A Multiple Utterances based Neural Network Model for Joint Intent Detection and Slot Filling

Lingfeng Pan, Yi Zhang, Feiliang Ren, Yining Hou, Yan Li, Xiaobo Liang, Yongkang Liu
School of Computer Science and Engineering
Northeastern University
Shenyang, 110169, China
{renfeiliang@mail.neu.edu.cn}

ABSTRACT: *Spoken language understanding(SLU), which usually involves slot filling and intent detection, is an important task in natural language processing. Most of the state-of-the-art methods are usually take single utterance as input, which would introduce much ambiguity because the loss of context information. To address this issue, we propose a new neural network based joint intent detection and slot filling model which takes multiple utterances as input. In our method, we use an utterance2utterance attention mechanism to combine the information of multiple continuous utterances. We also combine the intent information to the slot filling process with a gating mechanism. Using this proposed model, we participated in the task2 of CCKS2018. Finally, our model ranks NO.2 among the hugely competitive models.*

Keywords: Intent Detection, Slot Filling, Neural Network, Attention Mechanism

Received: 16 November 2019, Revised 17 January 2020, Accepted 29 January 2020

DOI: 10.6025/jisr/2020/11/2/54-60

1. Introduction

Spoken language understanding (SLU) is a key component in spoken dialogue systems, which typically involves two tasks: intent detection and slot filling [1]. Given speakers?? several utterances as input, intent detection is to predict the last utterance's intent and slot filling is to extract some semantic slots for the last utterance. For example, given a utterance *放一个周杰伦的晴天* (Play a Jay Chou's Sunny day.)", a SLU system aims to identify the intent of this utterance is about "music", and further to tag the "*周杰伦*((Jay Chou) "artist" slot and tag "*晴天* (Sunny day)" "song" slot.

A high performance SLU system would benefit lots of natural language processing(NLP) tasks, such as Q&A, information extract, and so on What's more.

SLU is an important foundation for human-machine spoken dialogue systems. Thus, lots of researchers spare no effort to study in this research field.

Previous research usually treats intent detection and slot filling as two separated tasks. Intent detection is often treated as an utterance semantic classification task, in which many popular classifiers can be applied, such as support vector machines (SVM) [2] and deep neural network, etc. Slot filling is usually viewed as a sequence labeling task, in which lots of approaches can be used: such as maximum entropy Markov models (MEMMs) [3], conditional random fields (CRFs)[4], and recurrent neural networks (RNNs)[5], among others.

Recently, researchers begin to pay attention to the joint models that train intention detection and slot filling synchronously. For example, Liu and Lane (2016) [6] propose a joint model where the parameters for intent detection and slot filling are learned in a single model with a shared framework, and achieved state-of-the-art experimental results on some benchmark datasets for SLU.

However, most of these existing SLU models are evaluated by English datasets. It is unclear whether they could still perform well for other languages like Chinese which word boundaries are not readily identified. Besides, most of these methods usually take single utterance as input, which would introduce much ambiguity because the loss of context information. For example, given three utterances 唱一首 我们不一样 *(Please sing the song that "we are different")* 。/来一首歌 *(Please sing a song)* /我们不一样 *(We are different)*。" , we could hardly identify the correct intention of the last utterance without the former two utterances. On the contrary, we could recognize that "我们不一样。(*We are different*)" as "song" easily with the semantic information from the former two utterances.

To address these issues, we propose a new joint neural network based intent detection and slot filling method which takes multiple utterances as input. In our method, we apply a double-layer attention mechanism to combine the semantic information of the contextual utterances into the target utterance. Specially, we first use an *utterance2utterance* attention mechanism to combine the information of multiple continuous utterances. Then we combine the intent information to the slot filling process with a gating mechanism. With this model, we participated in the task2 of CCKS2018 (2018 China Conference on Knowledge Graph and Semantic Computing), and our model ranks NO.2 among the hugely competitive models.

2. Related Work

SLU is an important task in NLP, which usually involves two tasks: intent detection and slot filling. Intent detection is usually regarded as a classification task. And slot filling is usually taken as a sequence labeling task. In recent years, SLU has received extensive attention from academics and industry due to an increase in publicly available huge datasets.

For intent detection, early research [2] mainly focuses on choosing appropriate features, such as dates, locations, etc. Then, the selected features will be fed into a classifier. For slot filling, CRF [4] is widely used because of its great ability in handling sequence labeling task. Various deep learning models are also widely applied in both intent detection and slot filling. For example, deep belief nets [7] and deep convex networks [8] are used in intent detection. RNNs [5] perform well for slot filling and outperform some traditional models like CRF [4].

Recently, many neural networks based joint training models [9] have been proposed which aim to take full advantage of the correlative information between intent detection and slot filling. These joint models achieved great success and obtained state-of-the-art performance for SLU. Besides, the attention mechanism [10] is another widely used technique for SLU. For example, Liu and Lane [6] used the attention mechanism to train the network so that it can focus on the important components of a input sequence. Finally, their method achieved better experimental results than compared baselines. Hierarchical structure is viewed as another kind of useful information for joint modeling. For example, Contextual Hierarchical Joint (CHJ) Model [11] makes use of both hierarchical and contextual features when jointly modeling intent detection and slot filling. Gating mechanism can control the flow of information, which means it can retain important information and abandon redundant information. Thus it is also widely used for SLU. For example, slot-gated models [12] use a slot gate to focus on learning relationship between intent and slot attention to get better results.

However, different from the proposed method of this paper, none of these mentioned methods take multiple utterances as input for SLU.

3. Model

Figure 1 demonstrates the architecture of our method. There are 3 major components in our method, which are utterance2utterance attention component, slot attention and intention attention component, and the output component.

3.1. Utterance2utterance Attention

This component aims to encode the utterances in a sample with BiGRU, and then fuse the generated representations of these utterances with an *utterance2utterance* attention mechanism. There are two steps in this component. First, the input utterances are encoded into real-valued vectors, and then these vectors are fused together with an *utterance2utterance* attention mechanism.

Encode Layer. Bidirectional Gated Recurrent Unit(BiGRU) is useful to address the long-term dependency issue. Moreover, BiGRU can alleviate the gradient vanishing/exploding problems. Compared with bidirectional long short-term memory(BiLSTM), BiGRU usually performs better with less parameters. So here we use BiGRU to encode the input utterances.

The input of our model is m utterances (in our experiments, $m = 3$), each of which has a form of $\{w_t^i\}_{t=1}^{T_i}$, where w_t^i represents the word embedding of word w_t in utterance i, and T_i is the length of utterance i. Then we encode w_t^i with equation (1).

$$h_t^i = BiGRU_i(h_{t-1}^i, w_t^i) \tag{1}$$

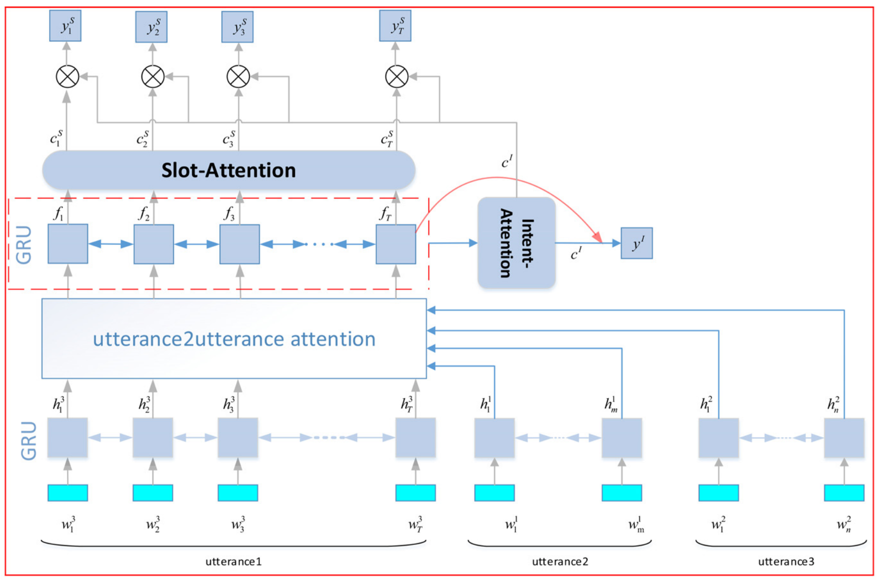

Figure 1. Architecture of our Model

Information Fusing. Taking the representations $\{h\}_{t=1}^{T_i}$ computed by (1) as input, we will fuse the semantic information of the m utterances into one vector representation with an *utterance2utterance* attention mechanism. The final representation is computed with equation (2).

$$f_t = BiGRU(f_{t-1}, [h_t^3; s_t^{1,3}; s_t^{2,3}]) \tag{2}$$

where $s_t^{p,q}$ is the result of combining the features from *utterancep* into the representation of *utteranceq*, and is computed with equation (3).

$$s_t^{p,q} = \sum_{j=1}^{T_p} \beta_{t,j} h_j^p$$

$$\beta_{t,j} = exp(x_{j,t}) / \sum_{j=1}^{T_p} exp(x_{j,t})$$

$$x_{j,t} = \varphi(h_j^p, h_t^q) \tag{3}$$

where φ is a function to compute the similarity between h_j^p and h_t^q, and we use dot product here.

3.2. Slot Attention and Intent Attention

We denote the output of previous step as $\{f_t\}_{t=1}^{T_3}$. For each f_i, we compute its slot context vector c_i^S as a weighted sum of $\{f_t\}_{t=1}^{I_3}$ multiplied by the learned attention weights $\alpha_{i,j}^S$:

$$c_t^S = \sum_{j=1}^{T_3} \alpha_{t,j}^S f_t \qquad (4)$$

The attention weights are computed with equation (5).

$$\alpha_{t,j}^S = exp(e_{j,t}) / \sum_{k=1}^{T_3} exp(e_{k,t})$$

$$e_{k,t} = g(f_k, f_t) \qquad (5)$$

With the same way, we can compute c^I which will be used to detect the intention.

3.4. Output layer

With a slot-gated mechanism, both $\{c_t^S\}_{t=1}^{T_3}$ and c^I are used to predict the slot of each word in the last input utterance. Besides, c^I will be utilized to predict the intention of the last utterance directly.

Slot-gated mechanism. We apply a gate that can leverage intent context vector for modeling slot-intent relationships to improve the performance of slot filling. The gate g is computed with equation (6).

$$g = \sum v \cdot \tanh(c_t^S + W \cdot c^I) \qquad (6)$$

where v and W are trainable vector and matrix respectively. Here g can be viewed as the weighted feature of the c_t^S and c^I.

Slot prediction. With g, we can predict the slots of last utterance with equation (7).

$$y_t^S = softmax(W^S(f_t + c_i^S \cdot g)) \qquad (7)$$

Intention Detection. The intention can be computed with equation (8), where f_r is the last hidden state of BiGRU.

$$y^I = softmax(W^I(f_T + c^I)) \qquad (8)$$

3.5. Training

The loss function of our method is defined as followings

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}(\log(y_I^I|\theta) + \sum_{t=1}^{T_3^i}\log(y_t^S|\theta)) \qquad (9)$$

where N is the total number of training samples and θ indicates all parameters of our models.

We adopt Adam to minimize the loss function and apply the way of mini-batch to train our model. We apply the dropout operation at encode layer during training.

Specially, the dropout operation is used at the output of equation (1).

4. Experiments

4.1. Dataset

We use the dataset provided by the task2 of CCKS2018 to evaluate our model. This dataset is used for intent detection and slot filling in Chinese music field. There are 3 utterances for each sample in this dataset, and we need to detect the intention of the last utterance with the help of the first two utterances. Figure2 shows the architecture of a sample in this dataset, where each word in the third utterance possesses a slot label, and the whole third utterance own a specific intention. And all of these samples come from real user logs in spoken dialogue system, and have been selected and processed by human. Totally there are 12000 samples in this dataset. We use 10000 of them as training set and 2000 of them as dev set.

	Utterance 1	utterance2	utterance3
Utterances	唱 一首 歌 。	来 一个 。	放 一个 周杰伦 的 稻香 。
Slots			O O artist O song O
Intent	Music		
Final answer	{“artist”: “周杰伦”, “song”: “稻香”}		

Figure 2. An sample(three utterances) with annotations of semantic slots, Intent and final answer provided by the Task2 of CCKS2018

4.2. Evaluation and Experiments Setup

We take F1 and accuracy(Acc) as evaluation metrics just as the task requires. Specifically, we evaluate the performance of intention detection with F1_I. When the last utterance of a sample has a music intention and some music entities, we use F1_E to evaluate the performance of slot filling. And when the last utterance of a sample have a music intention but hasn’t music entities, we use Acc to evaluate the performance of slot filling. The organizer of this competition task also uses a unified score to evaluate the whole performance of intention detection and slot filling. And the score is computed with equation (10).

$$\mathrm{Score} = \mathrm{F1\_I} + \alpha \cdot \mathrm{F1\_E} + \beta \cdot \mathrm{Acc} \tag{10}$$

where α represents the proportion of the samples that possess music intention with music entities; β represent the proportion of the samples that possess music intention without music entities.

In all experiments, we set the number of units in BiGRU cell as 128. Word embeddings, the dimension of which is set to 128, are pre-trained by word2vec tool with all the 12000 samples. Other hyper parameters are set as follows: the batch size is set to 16, the dropout rate is set to 0.5, and the learning rate is set to 0.001.

4.3. Experimental Results and Discussion

Comparisons with other methods. In the first part of our experiments, we compare our method with other two DNN based SLU methods. One is LSTM+CRF, which is widely used in named entity recognition (NER) task. The other is an attention-based Encoder-Decoder model. The comparison results are shown in Table 1.

Models	F1_I	F1_E	Acc	Score
LSTM+CRF [13]	81.5	50.7	86.4	1.170
Encoder-Decoder [6]	84.1	76.1	84.2	1.365
Our model	87.2	81.3	84.4	1.41

Table 1. Off line performance of CCKS task2 with different models

From Table 1, we can see that the performance of LSTM+CRF is not as good as other models. We argue the possible reason is that the dataset used here is not large enough to learn the transition probability between slot tags. Besides, the utterances in the dataset are always short and the categories of slots are much diverse where the connections between slots may be too weak to learn.

Effectiveness of different model components. In the second part of our experiments, we carry out ablation experiments to illustrate the contributions of each component in our model. The experimental results are shown in Table 2.

Models	**F1_I**	**F1_E**	**Acc**	**Score**
Attention-based RNN	85.9	79.8	82.7	1.398
+slot-gate	86.3	81.1	84.6	1.409
+Full sentence	87.2	81.3	84.4	1.417

Table 2. Off line performance of each part in our model

From Table 2 we can see that F1_E is improved about 2% when the slot gating mechanism is used. In other words, with a slot-gated mechanism, our model can learn the relations between slots and intent sufficiently. Thus much useful information is provided for accurately filling the slot. what's more, when we add the feature information from utterance1 and utterance2 to utterance3 with the attention mechanism, the F1_I is improved about 1%. We argue the possible reason is that: for the short utterances, the model is difficult to judge their intentions without context information. When adding the semantic information of contextual utterances (here are utterance1 and utterance2), the model could more easily detect an utterance's intention. Take following sample *唱一首我们不一样(Please sing the song "we are different".)* 。/ *来一首歌(Sing a song)*。/ *我们不一样 (We are different.)*。" the last utterance will be classified as **NO** music intention before fusing the first two utterances, but it is classified correctly when the semantic information of the first two utterances is added.

5. Conclusion

In this paper, we propose a new joint neural network model for intent detection and slot filling in music domain. The main contributions of our method are listed as follows.

First, we take multiple utterances as input and combine all utterances in a sample with an utterance2utterance attention mechanism.

Second, we apply a slot gated mechanism to add the features from the intent detection task into the slot filling task, which is effective for further improving the performance of slot filling.

In the future, we will further explore the deeper connections within multiple continuous utterances. In addition, we will also explore using input utterances syntactic structure information to further improve the performance of input representations.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC No. 61572120, 61672138 and 61432013).

References

[1] Tur G, Hakkani-Tür, D., Heck, L., et al. (2011). Sentence simplification for spoken language understanding[C]// *IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2011:5628-5631.*

[2] Haffner, P., Tur, G., Wright, J. H. (2003). Optimizing SVMs for complex call classification[J]. Icassp, 2003, 1:I-632-I-635 vol.1.

[3] Mccallum, A., Freitag, D., Pereira, F C N. (2000). *Maximum entropy markov models for information extraction and segmentation[J]. Icml*, 2000:591—598.

[4] Puyang Xu., Ruhi Sarikaya. (2013). Convolutional neural network based triangular crf for joint intent detection and slot filling. In ASRU, 2013: 78-83.

[5] Mesnil, G., Dauphin, Y., Yao, K., et al. (2015). Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding[J]. *IEEE/ACM Transactions on Audio Speech & Language Processing*, 2015, 23(3):530-539.

[6] Bing Liu., Ian Lane. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. In Proceedings of INTERSPEECH, 2016

[7] Sarikaya, R., Hinton, G E., Ramabhadran, B. (2011). Deep belief nets for natural language callrouting[C]// *IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE*, 2011:5680-5683.

[8] Tur, G., Deng, L., Hakkani-Tür, D., et al. (2012). Towards deeper understanding: Deep convex networks for semantic utterance classification[C]// IEEE International Conference on Acoustics, *Speech and Signal Processing. IEEE*, 2012:5045-5048.

[9] Daniel Guo., Gokhan Tur., Wen-tau Yih., and Geoffrey Zweig. (2014). Joint semantic utterance classification and slot filling with recursive neural networks. *In: IEEE Spoken Language Technology Workshop* (SLT), 2014: 55-59

[10] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine translation by jointly learning to align and translate, " arXiv preprint arXiv; 1409.554-559

[11] Liyun Wen., Xiaojie Wang., Zhenjiang Dong., and Hong Chen. (2017). *Jointly Modeling Intent Identification and Slot Filling with Contextual and Hierarchical Information*. In NLPCC 2017: 3-15.

[12] Chih-Wen Goo., Guang Gao., Yun-Kai Hsu., Chih-Li Huo., Tsung-Chieh Chen., Keng-Wei Hsu., Yun-Nung Chen. (2018). Slot-Gated Modeling for Joint Slot Filling and Intent Prediction. In NAACL-HLT, 2018: 753-757.

[13] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. *Computer Science*, 2015.