Research of Attention-Based Bi-GRU-CRF for Slot Filling

Luwang Zhou*, Zhimin Huang, Ying Nie
*Dept. of Geography Graphics and Image Research and Development, North China Institute of Computing Technology, No.211, North Fourth Ring Road, Beijing 100084, PR China
*luwangchou@foxmail.com

Abstract. Slot Filling (SF) is a critical part of spoken language understanding (SLU) which targets to capture semantic constituents from a specific utterance. It is considered as a sequence labeling issue. Currently, recurrent neural networks have shown promising effectiveness in this issue. Considering the effects of interrelated information within adjacent words and labels, we present that a novel approach consists of bi-directional gate recurrent unit (Bi-GRU), attention mechanism, and conditional random field (CRF). Our model can utilize interrelated information from words in the neighborhood, highlight key information, and exploit the dependencies within labels corresponding to surrounding words. The empirical experiments illustrate that our model significantly boosts the F1 score with around 1% and 5.1% relative enhancement on two public benchmark dataset ATIS and SNIPS separately.

Keywords: Spoken Language Understanding, Slot Filling, Attention Mechanism, Gate Recurrent Unit.

1. Introduction
Slot Filling (SF) plays a vital part in spoken language understanding (SLU) and dialogue system which targets to extract semantic elements from specifically spoken language utterance. It also has been treated as a sequence labeling task which annotates the input tokens \( x = (x_1, x_2, ..., x_T) \) with the corresponding labeled sequences \( y = (y_1, y_2, ..., y_T) \). As is depicted in Tab.1, given the respective utterance of the two benchmarks, there are various slot labels for every input token.

Early work in SF has adopted feature engineering approaches to train statistical models, e.g. conditional random field (CRF). Later, deep learning-based models are becoming an increasing trend in SLU and exceeding the aforementioned models. Recurrent neural networks (RNNs) have obtained significantly promising effects, especially gated recurrent unit (GRU) [1] and long short-term memory (LSTM). Currently, various joint learning approaches have emerged to capture and model the interrelation of intent detection and slot filling[2][3][4][5]. The proposed method in [1] incorporates the contextual information in two different level representation levels and task-specific level. Besides, special networks e.g. Slot-Gated Mechanism[2], Stack-propagation[3], Capsule Neural Networks[4] and pre-training model e.g. the Bidirectional Encoder Representation from Transformer (BERT) model are also applied to further boost the performance in SLU task[3][6][7].
Considering the effects of interrelated information within adjacent words and labels, we propose an effective approach consists of bi-directional gate recurrent unit (Bi-GRU), attention mechanism, and conditional random field (CRF) in this article. More specifically Bi-GRU is used to extract interrelated information within adjacent words. Otherwise, an attention mechanism is used to highlight key information. Finally, the conditional random field (CRF) is used to exploit the dependencies within labels corresponding to surrounding words. The empirical experiments on two benchmark datasets illustrate that our model effectively boosts SLU performance on the ATIS dataset and SNIPS[8] dataset.

| ATIS dataset | Sentence | show | flights | from | Boston | to | New | York | today |
|--------------|----------|------|---------|------|--------|----|-----|------|-------|
| Slot Label   | O        | O    | O       | B-dept | O      | B-arr | I-arr | B-date |
| SNIPS dataset | Sentence | listen | to | westbam | album | allergic | on | google | music |
| Slot Label   | O        | O    | B-artist | B-album | O      | B-service | I-service | O     |

Tab.1 The two utterances of ATIS or SNIPS dataset are annotated with semantic slots using BIO (Begin/In/Out) format

![Diagram](image)

Fig.1 The structure of the proposed Bi-GRU-CRF model

2. Proposed Methods
The structure of our model is depicted in Fig.1. The input example sentence, “what does dfw mean”, is divided into input token sequences \( x = (x_1, x_2, \ldots, x_T) \). The output of the network is the corresponding slot labels \( y = (y_1, y_2, \ldots, y_T) \) of each input token sequence. Next, we elaborate on our model thoroughly.

2.1. GRU Networks
RNNs are treated as an extension of traditional feed-forward neural networks. RNNs are widely-used to take advantage of the influence between past states and future states to deal with the variable-length sequence issues. However, owing to the vanishment or explosion of gradients, it was hard to train RNNs to capture long-term dependencies. Thus, two representative improvements are long short-term
memory (LSTM) and gated recurrent unit (GRU). Compared to LSTM, GRU comparably outperforms and has fewer parameters.

The following describes the definition of GRU. The hidden state $h_t$ of GRU at time $t$ is calculated from the below equations.

$$r_t = \sigma(W_r x_t + U_r h_{t-1})$$  

(1)

$$\tilde{h}_t = \tanh(W x_t + r_t \odot (U h_{t-1}))$$  

(2)

$$z_t = \sigma(W_z x_t + U_z h_{t-1})$$  

(3)

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$  

(4)

Here, $x_t$ stands for input token sequence at time $t$, $\sigma$ denotes as sigmoid function, $\odot$ denotes as the element-wise product of two vectors, $z_t, r_t$ stands for update gate vector and reset gate vector respectively, $W$ and $U$ are transformation matrices. Commonly, we use $h_t = GRU(x_t, h_{t-1})$ as the abbreviation of the above equations.

Bidirectional GRU involves the forward $h_t^{\text{forward}}$ hidden layers and the backward $h_t^{\text{backward}}$ hidden layer.

$$h_t^{\text{forward}} = GRU(x_t, h_{t-1}^{\text{forward}})$$  

(5)

$$h_t^{\text{backward}} = GRU(x_t, h_{t-1}^{\text{backward}})$$  

(6)

Hence, the bidirectional hidden state

$$h_t = [h_t^{\text{forward}}, h_t^{\text{backward}}]$$  

(7)

2.2. Attention-Based Slot Filling
Slot Filling is considered as a sequence labeling task, which is annotating the input tokens $x = (x_1, x_2, \ldots, x_T)$ to the corresponding labeled sequences $y = (y_1, y_2, \ldots, y_T)$. The combination of attention mechanism and neural network is widely-used in SF task and has become an indispensable part of modern neural networks. Its internal mechanism is to focus on important parts of the input sequence by altering weights of certain parts of the hidden state, thus making learning word representation from the given sentence more efficiently.

For each hidden state $h_t$, context vector $c$, slot label of the $i$-th input token $S_i$ can be calculated as follows.

$$c_i = \sum_{j=1}^{T} \alpha_{i,j} h_j$$  

(8)

$$y_i = \text{soft max}(W(h_i + c_i))$$  

(9)
Here, the attention weight vector $\alpha$ is acquired the same way as in [2].

Owing to adjacent slot label predictions are dependent, it is beneficial to consider the interrelate dependencies between the labels in neighborhoods. Therefore, we add the CRF layer on the top of the output labeled sequence layer to jointly decode the best chain of labels of the utterance.

3. Experiments

In this section, we make an evaluation of our proposed model on two English spoken language datasets, namely ATIS (Airline Travel Information System) and SNIPS. The statistics and three-fold division of the dataset are illustrated in Table 2.

| Dataset | Train | Development | Validation | Slots |
|---------|-------|-------------|------------|-------|
| ATIS    | 4478  | 500         | 893        | 120   |
| SNIPS   | 13084 | 700         | 700        | 72    |

**Tab.2** The Division of ATIS and SNIPS dataset

| Models                          | SNIPS Dataset | ATIS Dataset |
|---------------------------------|---------------|--------------|
|                                 | Slot (F1 Score%) | Sentence (Accuracy%) | Slot (F1 Score%) | Sentence (Accuracy%) |
| Joint Sequence[2] *             | 87.30         | 73.20        | 94.30         | 80.70         |
| Attention-Based[2] *            | 87.80         | 74.10        | 94.20         | 78.90         |
| Bi-GRU (no attention)          | 90.45         | 79.29        | 94.85         | 83.06         |
| Bi-GRU (with attention)        | 90.94         | 79.43        | 95.08         | 83.31         |
| Bi-GRU-CRF (no attention)      | 92.68         | 82.29        | 94.98         | 83.29         |
| Bi-GRU-CRF (with attention, proposed) | **92.95** | **83.14** | **95.07** | **84.10** |

**Tab.3** Slot Filling performance of models on ATIS and SNIPS dataset

* indicates the baseline model.

3.1. Data and Parameter

The ATIS dataset is a spoken text dataset for English flight booking. ATIS is widely used in the study of spoken language understanding. SNIPS is a dataset collected from personal voice assistant SNIPS. Compared with the single-domain ATIS dataset, the SNIPS dataset is a more complex and larger size of slot labels. The two benchmarks are divided into three-fold including train, validation, and test dataset.

In the experiments, the size of the hidden vector both of feed-forward and feed-backward hidden state is set to 64, the batch size is 16, the optimizer is Adam with learning rate 0.001, random uniformly distributed embedding with 100 dimensions initialize word embedding, the maximum epoch is set to 10 and 20 on ATIS and SNIPS separately with an early-stop strategy.

The experimental environments are that CPU is Intel Core i5-9600KF 3.7 GHz, GPU is RTX2060 Super with 8 GB memory, the whole computer RAM is 32 GB, Tensorflow version is 1.15 and CUDA version is 7.5.

3.2. Results and Analysis

We make an evaluation of the slot filling performance by using the F1 score and sentence-level semantic frame accuracy. As it is shown in Tab.2 The Division of ATIS and SNIPS dataset, our proposed approach obviously surpasses the baselines on two benchmark datasets. From the aspect of the F1 score, the proposed model can obtain relatively improvement by around 1% on the ATIS dataset while 5.15% on the SNIPS dataset. Furthermore, it may due to the internal mechanism of
AE which concentrates on key parts of the input token sequence by altering weights of specific parts of the hidden state. In addition, from the aspect of sentence-level semantic accuracy, the proposed model has gained better relative improvement on SNIPS than on the ATIS dataset. It may credit to that SNIPS contains much richer contextual information within adjacent words than ATIS. Considering the different complexity of two benchmark datasets, SNIPS involves much more diverse slot labels and larger vocabulary than ATIS. CRF layer plays a positive impact on overall model performance. This is because CRF can capture the maximum possible label sequence on the sentence frame.

In conclusion, the empirical experiments demonstrate that leveraging attention mechanism is able to obviously boost sentence-level semantic frame performance owing to global consideration in SF.

4. Conclusion
In this article, we concentrate on utilizing the effects of interrelated information within adjacent words and labels for slot filling task. We propose that a novel approach consists of attentive Bi-GRU with CRF. The Bi-GRU layer is used to utilize interrelated information within adjacent words while the attention mechanism layer is used to highlight key information. Otherwise, we use the CRF layer to exploit the dependencies within labels corresponding to surrounding words. The empirical experiments illustrate that our model significantly outperforms the baseline models. Although our proposed model has achieved obviously promising results in slot filling tasks on both benchmark datasets, we hope to further boost our model by utilizing the dependencies between intent detection task and slot filling task in the future.

References
[1] Amir P B V, Frank D and Thien H N 2020 Proc. of the 2nd Workshop on Natural Language Processing for Conversational AI(Online) (Stroudsburg PA ACL) pp90-95
[2] Chih-wen G, Guang G, Yun-kai H, Chih-li H, Tsung-chieh C, Keng-wei H and Yun-nung C 2018 Proc. of the Conf. the North American Chapter of the Association for Computational Linguistics (New Orleans) vol 2 (Stroudsburg PA ACL) pp753-757.
[3] Libo Q, Wanxiang C, Yangming L, Haoyang W and Ting L 2019 Proc. of Empirical Methods in Natural Language Processing (Hong Kong) (Stroudsburg: ACL) pp 2078-2087
[4] Chenwei Z, Yaliang L, Nan D, Wei F and Philip Y 2019 Proc. of the 57th Annual Meeting of the Association for Computational Linguistics(Florence) (Stroudsburg ACL) pp 5259-5267
[5] Haihong E, Peiqing N, Zhongfu C and Meina S 2019 Proc. of the 57th Annual Meeting of the Association for Computational Linguistics(Florence) (Stroudsburg PA ACL) pp 5467-5471
[6] Qian C, Zhu W and Wen W 2019 BERT for Joint Intent Classification and Slot Filling ArXiv 1902 10909
[7] Jacob D, Ming-Wei C, Kenton L and Kristina T 2019 Proc. of the Conf. the North American Chapter of the Association for Computational Linguistics(Minneapolis) (Stroudsburg PA ACL) pp 4171-4186
[8] Alice C, Alaa S, Adrien B, Theodore B, Alexandre C, David L, Clement D, Thibault G, Francesco C and Thibaut L 2018 Snips Voice Platform an embedded Spoken Language Understanding system for private-by-design voice interfaces ArXiv 1805 10190