Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 2, 2023

Improving NLP Accuracy with Stack-Propagation and Knowledge Distillation: A Joint Model for Intent Detection and Slot Filling

Yuan Wang*, Zhoumeng Yang, Xiyan Zhang

Department of Computer Science and Information, Beijing Jiaotong University, Beijing,100044, China
* **Corresponding author**: Yuan Wang (Email: bjtu_wangyuan@163.com)

Abstract: Intent detection and slot filling are fundamental tasks in Natural Language Understanding (NLU), constituting important components of intelligent question-answering systems. These tasks are closely interrelated, forming a core aspect of semantic understanding in natural language processing. In this paper, we propose an architecture using Stack-Propagation to improve the accuracy of NLU tasks. In stack propagation, we use a joint model that mainly incorporates token-level intent detection data with sentence word vectors as input for slot filling to capture intent semantic knowledge. Additionally, we use knowledge distillation (KD) to improve model efficiency and enhance the correlation between the two tasks. Our proposed framework significantly differs from existing joint models as it directly leverages intent information in the joint model and adopts token-level intent information for slot filling to ease error propagation. Furthermore, our model can explicitly incorporate intent information for slot filling with Stack-Propagation, making the interaction procedure more interpretable, while other models only interact with hidden states implicitly between the two tasks. We experimentally evaluated our model on two publicly available datasets, and the results demonstrate that it achieves state-of-the-art performance and outperforms previous methods by a significant margin, indicating its superiority in addressing the slot-filling and intent detection tasks. In future research, we will combine the stack-propagation frame with the KD module in the Transform model, which can boost our model performance to move forward a single step in the NLU task.

Keywords: Natural Language Understanding; Slot-filling; Intent Detection; Stack-Propagation; Knowledge Distillation.

1. Introduction

Natural Language Understanding (NLU) is the ability of a computer program to understand human language as it is spoken or written. According to Jurafsky and Martin (2019), NLU is defined as "the process of mapping a given input (usually text or speech) into a structured representation that captures the meaning of the input at different levels of granularity." NLU is an important component of natural language processing (NLP) and is used in various applications such as virtual assistants, chatbots, and intelligent search engines.

Slot filling and intent detection are fundamental tasks in Natural Language Understanding, constituting important components of intelligent question-answering systems. These tasks are closely interrelated, forming a core aspect of semantic understanding in natural language processing. Various joint models have been proposed to capture the correlation between the two tasks, and these models have demonstrated superiority over traditional pipelined approaches.

Table 1. An example with intent and slot annotation

Utterance	Sample Sentence			
Token	Leave On	Wednesday	April	Sixth
Token-level intent	Action	Weekday	Month	Day
Slot-filling	O	depart_date.day_name	depart_date.month_name	depart_date.day_number
Sentence Intent	Depart date information			

As indicated in Table 1, the utterance pertains to obtaining detailed information regarding a flight. By training the datasets of the Airline Travel Information System (ATIS), each token in the utterance is assigned different slot labels, and the whole utterance is assigned an intent label. This enables the model to identify and extract specific information from the utterance, such as the flight number or departure time.

In a general model, intent detection and slot filling are separated into different implemented training models. However, nowadays, more and more studies have shown that the slots and intents are highly correlated and that considering them jointly may improve the performance of natural language understanding systems (Goo et al., 2018; Li et al., 2018; Zhang et al., 2019).

For example, when processing a sentence that contains multiple temporal expressions, it is often necessary to comprehend the sentence as a whole in order to determine whether a particular expression pertains to a departure or arrival. However, using token-level intent detection can help avoid errors in slot filling by providing accurate intent information. For instance, by mapping the slot label "depart_date.day_name" to "arrival_date.day_name", the guiding annotation function of intent information for slot filling can be improved.

Considering the state-of-art performance for both intent detection and slot-filling, It is possible to solve some classical problems of fuzzy words and inclusion words in speech recognition.

In speech recognition emergencies, we often encounter the conflict between the two confusing words "help" and "rescuer", but in the face of such an emergency, we need to

greatly reduce the probability of its triggering by mistake. For such an understanding of semantics, we can use the joint model of slot filling and intent detection to understand token-level intent and finally accurately analyze the true intention of its sentences. This research can provide a new way to solve semantic understanding problems such as fuzzy words.

In this paper, we propose a joint framework to address these issues above. Zhang and Huang (2009) presented a stack-based approach to parsing coordination structures. In this regard, our proposed model adopts the concept of a joint framework for training both intent-detection and slot-filling tasks. Specifically, the output of intent detection is utilized as the token-level intent and serves as input for training the slot-filling task. To enhance the output performance, the framework implements Knowledge Distillation to distill knowledge from a larger deep neural network into a smaller network. The intent and slot-filling encodings are utilized as the Teacher Model, and a student model is obtained through the application of a softmax function via high-temperature knowledge distillation.

2. Analysis

2.1. Guidance

In this section, we will discuss the definitions and approaches of intent detection and slot-filling tasks, which provide a detailed description of our Stack-Propagation framework with Knowledge Distillation. We will also compare our approach with a general multitask framework to highlight the advantages of our proposed method. The structure of the Stack-Propagation with Knowledge Distillation model has been shown in Figure 1, and the following paragraph will explain each part of the model species.

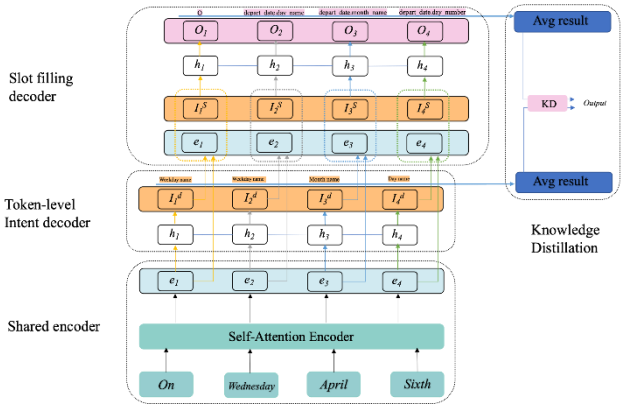

Figure 1. Stack-Propagation Framework with Knowledge Distillation

The Stack-Propagation with Knowledge Distillation model for correlative intent detection and slot filling is illustrated in this figure. The model comprises a shared encoder and two decoders. The output distribution of the intent detection network and the representations obtained from the encoder are concatenated to serve as input for slot filling. Both of the decoder outputs are transmitted to the Knowledge Distillation module for better performance.

2.2. Shared Encoder

The proposed Stack-Propagation framework with KD employs a shared encoder for both intent detection and slot-filling tasks, which is equipped with a Bidirectional Long Short-Term Memory (BiLSTM) and a self-attention mechanism. The BiLSTM reads the input tokens in both forward and backward directions, allowing for the production of a context-sensitive hidden state. The self-attention mechanism captures the hidden state and utilizes them to represent the Queries, Keys, and Values (QKV) matrices using different linear projections. The output, e, is then formulated as:

$$\mathbf{e} = \text{softmax}(\frac{\mathbf{QK}^\top}{\sqrt{d_k}})\mathbf{V}.$$

2.3. Token-level Intent Decoder

In this framework, we perform token-level intent detection. Through training a multitude of label features, the Stack-Propagation framework divides the linguistic meaning between various words and selects the most probable utterance-level intention via a voting process. Subsequently, the resulting semantic information is utilized to correct the token-level semantic output. The corrected token-level result is then input to the slot-filling decoder to complete the stack-propagation framework task.

2.4. Slot-filling Decoder

This paper highlights one of the advantages of the proposed Stack-Propagation framework, which is the direct utilization of explicit intent information to restrict the slots to specific intent and thereby reduce the workload of the slot-filling decoder. Moreover, to serve as the slot-filling decoder, we utilize another unidirectional LSTM in a similar manner. To obtain a more accurate output for slot filling, our proposed Stack-Propagation framework conducts bidirectional training, taking into account both the intent of the word vector and the token-level information in the original sentence. This allows the framework to obtain the instruction of the utterance intent in slot filling, resulting in a better output through the linkage effect between the two tasks.

2.5. Joint training with Knowledge Distillation

In another innovative model, we use a Knowledge distillation module with both the slot filling and intention detection results for further training. In this way we can not only convert the sentence-level classification task into token-level prediction to directly leverage token-level intent information for slot filling but can also combine the parameters between the original teacher model through knowledge transformation, and finally obtain the number of parameters of the student model is greatly reduced compared with that of the teacher model, thus greatly reducing the difficulty and training time of the training.

The intent detection objection can be formulated as:

$$\mathcal{L}_I \triangleq -\sum_{j=1}^{m}\sum_{i=1}^{n_I} \hat{\mathbf{y}}_j^{i,I}\log(\mathbf{y}_j^{i,I})$$

Thus, the slot-filling task output can be defined as:

$$\mathcal{L}_S \triangleq -\sum_{j=1}^{m}\sum_{i=1}^{n_S} \hat{\mathbf{y}}_j^{i,S}\log(\mathbf{y}_j^{i,S})$$

To achieve simultaneous slot filling and intent detection, the final joint objective combined with the knowledge distillation module $\mathcal{L}_K$ is expressed as follows:

$$\mathcal{L}_{final} = \mathcal{L}_I + \mathcal{L}_S + \mathcal{L}_K$$

The loss function is proposed by Liu and Lane, 2016, which is known as the pipeline approach, which separates intent detection and slot filling into two distinct processes. but it

suffers from error propagation, where errors in earlier stages can propagate and affect later stages.

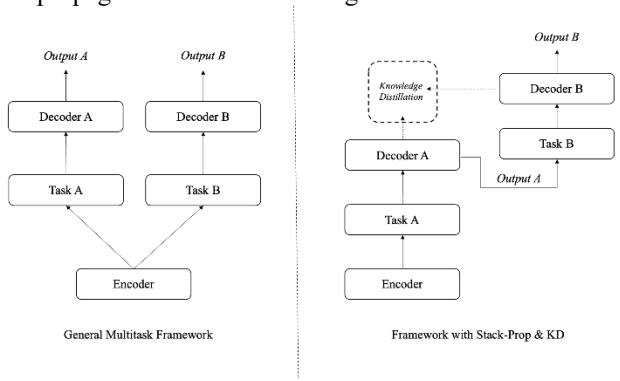

Figure 2. Comparison of General and Evaluated Multitask Framework

2.6. Framework with Stack-propagation and KD v.s. General Multitask Framework

Figure 2, illustrates the General Multitask Framework, which involves the execution of distinct tasks, namely Task A and Task B. However, the framework lacks correlation during the training stage, and the only shared component is the encoder. This results in a lack of guidance from the upstream task to the downstream task within the General Multitask Framework. Furthermore, in the event of an error in the upstream task, there is no provision for negative feedback from the downstream task to modify and correct the error promptly, leading to the propagation and escalation of errors. However, Stack-Propagation with knowledge distillation can mitigate the shortcoming. First, the output of Task A can guide the training results of Task B, and Task B can also adjust the input results to the training model. In the final KD module, the joint model can also reduce the training amounts of parameters, so that the model can be optimized.

3. Experiment

In order to test the training efficiency of our model, we used two open-source data sets to obtain the advantages of this model in training. The ATIS (Airline Travel Information System) and SNIPS (Smart Home Interaction) datasets are two of the most widely used benchmarks for natural language understanding (NLU) tasks. The ATIS dataset comprises a collection of queries related to airline travel, while the SNIPS dataset consists of a range of queries related to smart home devices. Both datasets contain a large number of user queries, annotated with intent labels and slot labels. The intent labels represent the overall intention of the query, while the slot labels identify specific pieces of information within the query. These datasets have been used extensively in research on NLU, which have been instrumental in the development of many state-of-the-art models for intent detection and slot-filling. The parameters in our model are optimized using Adam (Kingma and Ba, 2014), which is a popular optimization algorithm in deep learning.

3.1. Baselines

We analyze the baselines that are relevant to this model, including:

Joint Sequential model (Joint Seq.): It is a popular baseline for the task of intent detection and slot filling. It involves training a single model to perform both tasks simultaneously, with the output of the model consisting of both the intent and the slot labels for each input sentence. This approach has been shown to achieve good performance on benchmark datasets, and has been widely used in research on natural language understanding (NLU) tasks.

Attention-based model (Attention-based.): The model utilizes a self-attention mechanism to capture contextual information and enhance the model's ability to attend to relevant parts of the input sentence. This approach has been shown to achieve state-of-the-art performance on various NLU tasks, including intent detection and slot filling.

Transformer-based model (Transformer-based.): The Transformer-based model is a neural network model that has been used for various natural language processing tasks, including intent detection and slot filling. It was first introduced in a paper by Vaswani et al. (2017) and has since become a popular approach in the field of NLP. The Transformer model utilizes self-attention to capture contextual information and has been shown to achieve state-of-the-art performance on a range of benchmark datasets.

3.2. Overall Result

Table 2. Performance of related model

Model	SNIPS			ATIS		
	Slot	***Intent***	***Overall***	***Slot***	***Intent***	***Overall***
Joint Seq. Model	87.3	96.9	73.2	94.3	92.6	80.7
Att.based Model	87.8	96.7	74.1	94.2	91.1	78.9
Trans. Model	90.0	97.5	81.0	95.1	96.8	82.8
Our model without K.D.	94.2	98.9	86.9	95.9	96.9	86.5
Our model	95.7	96.7	86.7	95.4	96.9	87.1

The final line shows the performance of Stack-Prop. With KD model

As can be seen from the table, the performance of the first three models has gradually increased. In the model proposed in this paper, state-of-the-art performance is obtained, and the training results and training difficulty are reduced after using knowledge distillation.

This characteristic of ours states that our architecture directly incorporates the intent information to improve the accuracy of NLU tasks by using stack propagation. In stack propagation, a joint model is used that mainly incorporates token-level intent detection data with sentence word vectors as input for slot filling to capture intent semantic knowledge. Additionally, knowledge distillation is used to improve model efficiency and enhance the correlation between the two tasks. To visualize the role of intent information in NLU tasks, we present the results in Table 2, where the output of the joint model with intent information used in slot filling is compared to the output when only sentence-level information is utilized. The results demonstrate that the joint model with intent information achieves better performance in slot filling by capturing better intent information. Moreover, the results confirm our claim that intent information can be used to guide slot filling.

4. Related Work

Slot filling is commonly considered a task of sequence labeling. Popular techniques used for slot filling are conditional random fields (CRF) and recurrent neural networks (RNN) (Yao et al., 2014).

On the other hand, intent detection is approached as an

utterance classification problem. Various classification methods such as support vector machine (SVM) and RNN (Sarikaya et al., 2011) have been proposed to tackle this problem.

To address the issue of error propagation caused by pipelined approaches, researchers have proposed joint models in recent times. Zhang and Wang (2016) were the first to use Recurrent Neural Networks (RNNs) in a joint model to capture the correlation between intent and slots. Liu and Lane (2016) proposed an attention-based neural network that jointly modeled the two tasks, leading to mutual enhancement between them. These joint models have demonstrated superiority over pipeline models by considering the correlation between the two tasks. However, they did not explicitly model the intent information for slots, only considering their correlation by sharing parameters. In addition to these works, some recent studies have proposed incorporating intent information into the joint model. For instance, Chen et al. (2019) proposed an intent-aware attention-based model that takes into account the intent information while performing slot filling. They demonstrated that their approach outperformed previous joint models that did not consider explicit intent information. Recently, there has been a growing interest in joint models to address the issue of error propagation caused by pipelined approaches in slot-filling and intent detection tasks. Researchers have proposed various joint models to capture the correlation between the two tasks. However, most of these models did not explicitly model the intent information for slots and only considered their correlation by sharing parameters.

Several joint models have been proposed to incorporate the intent information for slot filling. Li et al. (2018) proposed the intent-augmented gate mechanism to utilize the semantic correlation between slots and intent. Wang et al. (2018) proposed the Bi-model to consider the cross-impact between the intent and slots and achieve state-of-the-art results. Zhang et al. (2019) proposed a hierarchical capsule neural network to model the hierarchical relationship among word, slot, and intent in an utterance. Chen et al. (2019) proposed an intent-aware attention-based model that takes into account the intent information while performing slot filling. They demonstrated that their approach outperformed previous joint models that did not consider explicit intent information.

5. Conclusion

Our proposed framework is significantly different from these models in several aspects. Firstly, our model directly leverages the intent information in the joint model by feeding the predicted intent information directly into slot filling with a Stack-Propagation framework. Secondly, we can directly incorporate the intent information for slot filling explicitly with Stack-Propagation, making the interaction procedure more interpretable, while their models just interact with hidden states implicitly between two tasks. Finally, for a better performance of training the model and improving the correlation between the intent detection and slot-filling tasks, we put the results generated by the two tasks into the knowledge distillation module, and the final results show that the accuracy is also improved. For the experiment section, the results on the two datasets show the state-of-art level beyond the former experiment, which shows the effectiveness of the stack propagation framework and knowledge distillation module.

References

[1] Jurafsky, D., & Martin, J. H. (2019). Speech and language processing (3rd ed.). Cambridge University Press.

[2] Chen, L., Zhang, Y., Du, N., Liu, X., & Sun, M. (2019). Multi-task learning for joint language understanding and dialogue state tracking. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 6901-6908).

[3] Li, X., Chen, Q., Li, X., Du, N., & Zhou, D. (2018). A self-attentive model with gate mechanism for spatiotemporal slot filling in spoken language understanding. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 4019-4028).

[4] Zhang, Y., Chen, Y. N., & Chen, W. (2019). Joint extraction of entities and relations based on a novel graph scheme in natural language processing. IEEE Access, 7, 47493-47505.

[5] Liu, B., & Lane, I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. In Proceedings of the Association for Computational Linguistics (ACL) (pp. 1-10).

[6] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).