Intent Detection and Slot Filling from Dependency Parsing Perspective: A Case Study in Vietnamese

Phu-Thinh Pham, Duy Vu-Tran, Duc Do, An-Vinh Luong, Dien Dinh
University of Science, Ho Chi Minh city, Vietnam
Vietnam National University, Ho Chi Minh city, Vietnam
{phpthinh18, vtduy18}@apcs.fitus.edu.vn
{dotrananhduc, anvinhluong}@gmail.com
ddien@fit.hcmus.edu.vn

Abstract

Spoken language understanding (SLU) systems using deep learning techniques require effective intent detection and slot filling models. Previous studies in this field taking advantage of sequence to sequence models have achieved good results. However, they focus on the locality of words and thus are sensitive to surrounding terms. In this paper, we introduce a new approach for this problem inspired by the dependency parsing techniques via a biaffine model to give the system a global view of the input. The experiments on PhoATIS dataset for Vietnamese have shown that our joint model for intent detection and slot filling obtains potential results.

1 Introduction

Spoken language understanding (SLU) has been applied to many chatbot applications in recent years. Intent detection and slot filling are two main tasks in this field for building task-oriented dialog systems. The purpose of intent detection task is to classify users' intent and that of slot filling task is to extract semantic constituents from the natural language utterances (Tur and De Mori, 2011). The most common approach for intent detection task is using a classifier based on [CLS] context representation. In parallel, the slot filling task is usually considered as a sequence to sequence problem, with the help of conditional random fields (CRFs) and recurrent neural network (RNN). Normally, these two tasks are considered as two distinct tasks, thus implemented separately, although the slots intuitively depend on the intent (Goo et al., 2018). Hence, some studies have proposed joint models based on the correlation between two tasks, enhancing the performance of each other (Goo et al., 2018; Dao et al., 2021; Wu et al., 2020; Wang et al., 2018; Chen et al., 2019).

Briefly summarized, most of the previous studies take advantage of autoregressive model or sequence to sequence architecture to solve the problems (Wu et al., 2020). For example, conditional random field (CRF) is a common approach for slot filling tasks since it considers the correlations between tags. However, we argue that this approach heavily relies on the locality of words and we need to provide the model with a global view of the input.

In this study, we propose a joint model for intent detection and slot filling tasks inspired by the dependency parsing technique. For slot filling task, we reformulate it as the task of identifying the span of a slot and assigning its category, following the study of Yu et al. (2020). In parallel, we consider the intent detection task as the task of classifying the intent labels of the span from the beginning to the end of an utterance. Furthermore, we incorporate the intent context information with an intent-slot attention layer into slot filling, following Dao et al. (2021). Our system uses two biaffine modules (Dozat and Manning, 2016) for the two tasks to estimate the scores to all spans in an utterance. After that, the logits are decoded to return the final results to satisfy the constraints.

We evaluate our system on the PhoATIS dataset (Dao et al., 2021), the first public dataset for Vietnamese intent detection and slot filling. In spite of being the 17th most spoken language in the world (Eberhard et al., 2019), the research attention in this field for Vietnamese has not gained any consideration until the appearance of PhoATIS. The experiments show that our system achieved competitive results and set a new benchmark for this corpus and this language.

In summary, we: (1) introduce a new approach for intent detection and slot filling system inspired by the graph-based dependency parsing technique; (2) propose a joint model that obtains better performance on the Vietnamese dataset.

2 Related work

The introduction of ATIS dataset (Hemphill et al., 1990) has motivated research studies in natural langugae understanding (NLU) and there are efforts to conquer this field. (Chen et al., 2019) has explored the influence of BERT (Devlin et al., 2018) on SLU systems by proposing a joint intent classification and slot filling model based on BERT. With the help of such powerful architecture, they obtain significant improvement in intent classification accuracy, slot F1 score, and sentence-level semantic frame accuracy.

For Vietnamese, PhoATIS (Dao et al., 2021) has been introduced as the first public intent detection and slot filling dataset, setting a starting point for future Vietnamese SLU research. In addition, they also propose a joint model based on the work of Devlin et al. (2018), extending the model by integrating an intent-context attention layer. It helps the model to recognize slots in an utterance more effectively with intent context information. With this architecture, they achieve potential results on the Vietnamese dataset, significantly outperforming the original work.

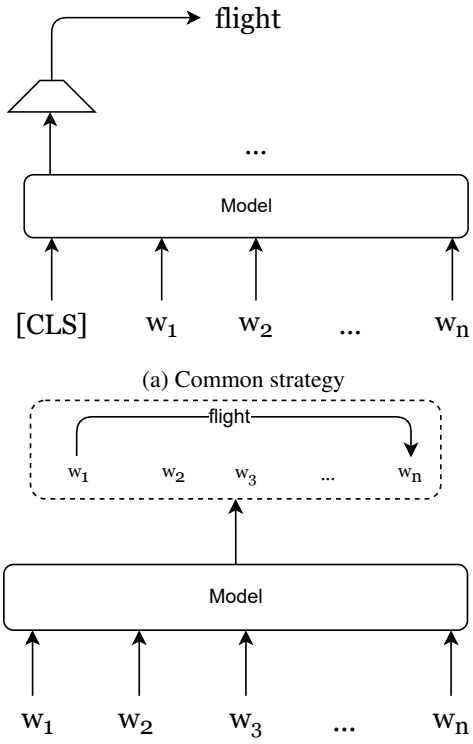
The study of Yu et al. (2020) gives a novel view to named entity recognition (NER) task, as well as sequence labeling problems in general, by applying the ideas from graph-based dependency parsing. It uses a biaffine model (Dozat and Manning, 2016) to score all possible spans in a sentence, enabling the model to predict named entities more accurately. From scores of all pairs of start and end tokens, it ranks the candidate spans based on their scores and selects top-ranked spans satisfying the constraints for flat or nested entities. The experimental results show that the model can handle nested entities well and gain competitive performance on both nested and flat NER.

3 Method

In this section, we first briefly introduce our novel approach for both intent detection and slot filling tasks. Thus, we describe the proposed joint model based on the dependency parsing technique.

3.1 Intent detection

In general, the common strategy for the intent detection task is to predict the intent based on the hidden state of the first special token ([CLS]). In this paper, we reformulate it as the task of classifying the whole sentence, represented by a span from the beginning to the end of the sentence. Figure 1 compares our proposed approach and the common approach.



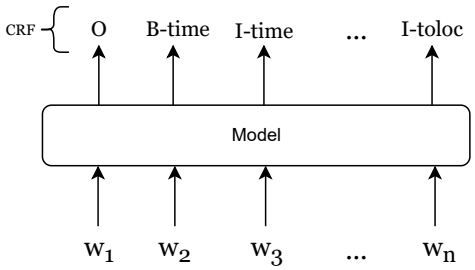
(a) Common strategy

(b) Our new approach

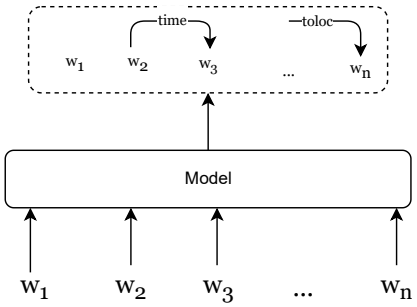
Figure 1: Comparison between the common method and our proposed approach

3.2 Slot filling

Previous studies consider slot filling task as a sequence labeling problem, with a CRF layer for prediction. In our approach, to provide a more general view, we adopt ideas from the graph-based dependency parsing model inspired by the study of Yu et al. (2020). In detail, we reformulate slot filling as the task of identifying the start and end indices of a slot, as well as classifying its category. Table 2 illustrates the difference between our approach and the previous approach. By using the biaffine model of Dozat and Manning (2016), our model scores all possible spans that could form a slot in an utterance. Thus, our system ranks these spans based on the logits predicted by the biaffine model and accordingly selects top-ranked spans complying with constraints that no two slots are overlapped. Formally, given an $n \times n \times c$ tensor T from our model, where n is the length of the utterance and c is the number of slot types $+1$ (for non-slot), each span i with the start and end indices $s_i \leq e_i$ is assigned the category c with the highest



(a) Common approach using CRF layer



(b) Our approach using Biaffine layer

Figure 2: Comparison between previous and our approaches

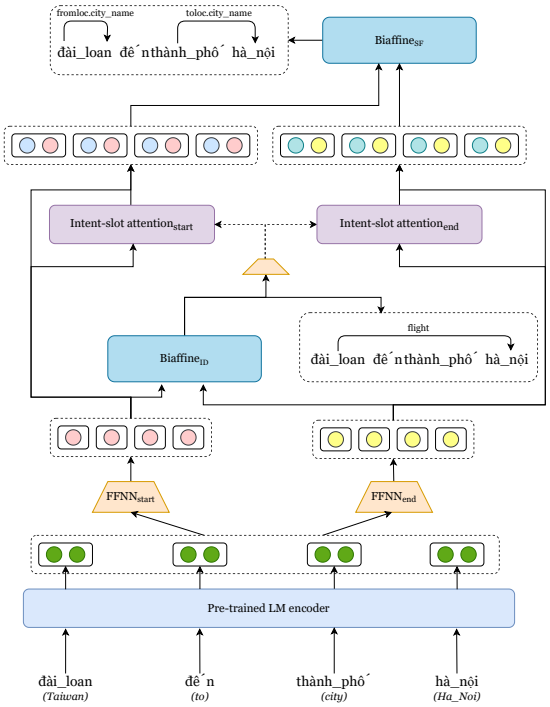


Figure 3: Illustration of our proposed model

score:

$$y'(s_i, e_i) = \arg\max_c T(s_i, e_i, c) \quad (1)$$

Finally, all spans whose category is different from non-slot are ranked based on their scores in descending order. A slot i will be selected if there is no higher-ranked slot j such that $s_i \leq s_j \leq e_i$ or $s_j \leq s_i \leq e_j$.

3.3 Model architecture

The architecture of our joint model is illustrated in Figure 3, consisting of 7 layers: an encoding layer, two feed-forward neural network (FFNN) layers, two intent-slot attention layers, and two biaffine layers.

Encoding layer

In the encoding layer, we employ a pre-trained Transformer-based language model (LM) to generate context-dependent sentence representations of an utterance. Here, we utilize XLM-R (Conneau et al., 2019) as the encoder for the syllable-level dataset and PhoBERT (Nguyen and Nguyen, 2020) for its automatically word-segmented variant. Given an n-length input token sequence $\boldsymbol{w} = (w_1, w_2, ..., w_n)$, the output produced by the encoding layer is feature embeddings $\mathbf{c}_i$ representing the i^{th} token.

FFNN layers

Following the encoding layer, two separate FFNNs are used to extract different representations for the start and end of the spans. This allows the model to distinguish different contexts of start and end of spans and reduce the dimensions of the encoder's output. In particular, each layer feeds $\mathbf{c}_i$ into a single-layer feed-forward network to obtain the start or end representation of token i:

$$\mathbf{v}_i^{start} = \text{FFNN}_{start}(\mathbf{c}_i) \quad (2)$$

$$\mathbf{v}_i^{end} = \text{FFNN}_{end}(\mathbf{c}_i) \quad (3)$$

Intent-slot attention layers

Following the architecture presented in (Dao et al., 2021), we use an attention mechanism to take advantage of intent information for better slot filling performance. In particular, each intent-slot attention layer takes the input from the start/end (s/e) FFNN layer and the probability vector $\mathbf{p}$ from the output of the intent detection module to produce s/e intent-specific vectors. In details, the layer creates an intent label embedding $\mathbf{r}$ via a label weight matrix $\mathbf{R}$ and uses it to give the intent-specific vector $\mathbf{h}_i$:

$$\mathbf{r}^{s/e} = \mathbf{R}^{s/e}\mathbf{p} \quad (4)$$

$$\alpha_i^{s/e} = \frac{\exp((\mathbf{r}^{s/e})^T \mathbf{v}_i^{s/e})}{\sum_{j=1}^{n} \exp((\mathbf{r}^{s/e})^T \mathbf{v}_j^{s/e})} \quad (5)$$

$$\mathbf{h}_i^{s/e} = \alpha_i^{s/e} \mathbf{r}^{s/e} \quad (6)$$

After that, a sequence of vectors $\mathbf{s}_{1:n}$ is created, where s_i is the concatenation of intent-specific vector and the corresponding start/end representation from the FFNN layer:

$$\mathbf{s}_i^{s/e} = \mathbf{h}_i^{s/e} \circ \mathbf{v}_i^{s/e} \quad (7)$$

Biaffine layers

Our model consists of two biaffine layers of Dozat and Manning (2016), one for intent detection and the other for slot filling task. To be more specific, the layer for intent detection takes two sequences of vectors $\mathbf{v}_{1:n}^{start}$ and $\mathbf{v}_{1:n}^{end}$ as the inputs while the other one feeds $\mathbf{s}_{1:n}^{start}$ and $\mathbf{s}_{1:n}^{end}$. Each layer returns an $n \times n \times c$ tensor, where c is the number of intent labels for intent detection task and the number of slot types $+1$ for slot filling task as explained in 3.2.

3.4 Joint training

The learning objective of our model is to classify the correct intent and correct slot type for each valid span. Therefore, we consider them as two multi-class classification problems and optimize our models for both tasks with softmax cross-entropy. Given the tensor T_{ID} produced by the biaffine layer for intent detection, the probability vector $\mathbf{p}$ is calculated via a softmax function:

$$p_i = \frac{\exp(T_{ID}(1, n, i))}{\sum_{j=1}^{k} \exp(T_{ID}(1, n, j))} \quad (8)$$

where k is the number of intent classes. Based on the vector $\mathbf{p}$, a loss $\mathcal{L}_{ID}$ for intent classification is then computed:

$$\mathcal{L}_{ID} = -\sum_{i=1}^{k} y_i \log(p_i) \quad (9)$$

For slot filling, a cross-entropy objective loss $\mathcal{L}_{SF}$ is calculated from the output T_{SF} of biaffine layer:

$$p'(s, e, i) = \frac{\exp(T_{SF}(s, e, i))}{\sum_{j=1}^{c} \exp(T_{SF}(s, e, j))} \quad (10)$$

$$\mathcal{L}_{SF} = -\sum_{s=1}^{n} \sum_{e=s}^{n} \sum_{i=1}^{c} y(s, e, i) \log(p'(s, e, i)) \quad (11)$$

The final loss $\mathcal{L}$ is the weighted sum of the intent detection loss $\mathcal{L}_{ID}$ and slot filling loss $\mathcal{L}_{SF}$.

$$\mathcal{L} = \delta \mathcal{L}_{ID} + (1 - \delta) \mathcal{L}_{SF} \quad (12)$$

where $0 < \delta < 1$ is the mixture weight.

Model	Intent	Slot	Sent.
Syllable-level			
JointBERT+CRF	97.42	94.62	85.39
JointIDSF	97.56	94.95	86.17
Our model	97.61	**95.05**	85.89
Word-level			
JointBERT+CRF	97.40	94.75	85.55
JointIDSF	97.62	94.98	86.25
Our model	97.80	**95.43**	**87.05**

Table 1: Results on the test set. Numbers written in bold indicate that the improvement of our model is statistically significant with $p\text{-}value < 0.05$ under t-test.

4 Experiments and Results

4.1 Experimental setup

We evaluate our models on the PhoATIS dataset (Dao et al., 2021) and conduct the experiments on both word and syllable levels. The dataset consists of 4478, 500 and 893 utterances for train, validation and test set, respectively with 28 intent labels and 82 slot types. For hyper-parameters, we follow the same configuration in the original work of Dao et al. (2021). To optimize the model, we use AdamW optimizer (Loshchilov and Hutter, 2017) and test on different δ in $\{0.05, 0.1, 0.15, ..., 0.95\}$ to select the optimal value. The batch size is set to 32 and the number of Transformer layers, attention heads and hidden sizes are 12, 12 and 768 respectively.

The metrics used for evaluation are the intent accuracy for intent detection, the F_1-score for slot filling and the overall sentence accuracy (Louvan and Magnini, 2020; Weld et al., 2021). During training, we compute the average score of intent accuracy and F_1 score at each epoch to select the checkpoint achieving the best performance on the validation set. We train the model for 100 epochs with the early stopping strategy. All results are reported on average over 3 runs with 3 different random seeds.

4.2 Results

Table 1 gives information about the results on the test set of our models, in comparison to the baseline JointBERT+CRF and JointIDSF reported in (Dao et al., 2021). Since we evaluate our models using the syllable-level dataset and its word-segmented variant, the results are presented in two comparable settings.

In syllable level, our model achieves 97.61%, 95.05% and 85.89% for intent accuracy, slot F_1

score and sentence accuracy, respectively. Especially, the slot F_1 score improvement over JointIDSF is statistically significant with $p\text{-}value < 0.05$. On the other hand, our model obtains better results in word level, with 97.80%, 95.43% and 87.05% for intent accuracy, slot F_1 and sentence accuracy respectively. In comparison to JointIDSF baseline, the slot F_1 score and sentence accuracy are statistically significant with $p\text{-}value < 0.05$.

From the results, we find that our models achieve better performance, except for the sentence accuracy on the syllable-level dataset. This can be explained by the fact that representing Vietnamese tokens at the syllable level cannot capture the whole meaning compared to word-segmented tokens. Thus, employing such information-lost token representations to compute the scores for all possible spans has a negative impact on model performances, leading to low sentence accuracy although the intent accuracy and slot F_1 are higher than the JointIDSF baseline. The significant difference in the sentence accuracy between the syllable-level dataset and its automatically word-segmented variant, 85.89% and 87.05% respectively, is the strong evidence for our explanation.

4.3 Ablation study

To evaluate the effectiveness of individual components in our system, we do an ablation study using the word-level setup because of its better performance. In particular, we remove selected components in our model and train them for evaluation.

To verify the contribution of two intent-slot attention layers in our proposed architecture, we sequentially remove one and then both of them. With only one attention layer, our system creates the s/e intent-specific vectors by sharing common parameters (using $\mathbf{r} = \mathbf{Rp}$ in equation 4 and replacing $\mathbf{r}^{s/e}$ in equations 5 and 6 by $\mathbf{r}$). Meanwhile, when two attention layers are removed, our model becomes a joint model consisting of two biaffine modules with the same s/e representations (using $\mathbf{s}_i^{s/e} = \mathbf{v}_i^{s/e}$ in equation 7). Finally, to confirm the influence of our new approach for intent detection task, we replace the biaffine layer responsible for classifying intents by a linear prediction layer using the [CLS] token.

Table 2 clearly shows that removing any components from our full model has a negative impact on its performance in all three metrics. When we completely remove the intent-slot attention layers, the performance witnesses a significant drop by 2.36% in sentence accuracy (from 87.05% to 84.69%). Adding an attention layer helps our model improve 1.8% score from 84.69% to 86.49%, 0.56% lower than the full model, clearly proving the contribution of this component in our architecture. Besides, when we replace the biaffine layer for intent detection with a single-layer feed-forward network based on the contextualized embedding of the classification token [CLS], the performance of our full model is reduced by 1.05% to 86.00%.

	Intent	**Slot**	**Sent.**
Our model	97.80	95.43	87.05
- One attention	97.76	95.23	86.49
- No attention	97.46	94.84	84.69
- [CLS] token	97.65	95.10	86.00

Table 2: Ablation study results on the test set.

5 Conclusion

In this paper, we have presented our work for Vietnamese intent detection and slot filling tasks. By proposing an effective architecture for jointly training intent detection and slot filling, we achieve better performance than the previous work JointIDSF. In particular, we adopt the ideas and techniques from dependency parsing to apply to our models, along with taking advantage of the intent-slot attention layer to integrate intent context information for better slot filling. In addition, we find that our proposed architecture works better at the word level compared to the syllable level. Furthermore, we empirically conduct experiments on the dataset to verify the contribution of each component in the architecture.

Acknowledgments

This research is supported by research funding from Advanced Program in Computer Science, University of Science, Vietnam National University - Ho Chi Minh City.

References

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised

cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2021. Intent detection and slot filling for vietnamese. *arXiv preprint arXiv:2104.02021*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

David Eberhard, Gary Simons, and Chuck Fennig. 2019. *Ethnologue: Languages of the World, 22nd Edition*.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. *arXiv preprint arXiv:2011.00564*.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based rnn semantic frame parsing model for intent detection and slot filling. *arXiv preprint arXiv:1812.10235*.

HENRY Weld, Xiaoqi Huang, SIQU Long, Josiah Poon, and SOYEON CAREN Han. 2021. A survey of joint intent detection and slot-filling models in natural language understanding. *arXiv preprint arXiv:2101.08091*.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020. Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling. *arXiv preprint arXiv:2010.02693*.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. *arXiv preprint arXiv:2005.07150*.