INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Intent Detection and Slot Filling for Vietnamese

Mai Hoang Dao, Thinh Hung Truong*, Dat Quoc Nguyen*

VinAI Research, Hanoi, Vietnam
{v.maidh3, v.thinhth88, v.datnq9}@vinai.io

Abstract

Intent detection and slot filling are important tasks in spoken and natural language understanding. However, Vietnamese is a low-resource language in these research topics. In this paper, we present the *first* public intent detection and slot filling dataset for Vietnamese. In addition, we also propose a joint model for intent detection and slot filling, that extends the recent state-of-the-art JointBERT+CRF model [1] with an intent-slot attention layer to explicitly incorporate intent context information into slot filling via "soft" intent label embedding. Experimental results on our Vietnamese dataset show that our proposed model significantly outperforms JointBERT+CRF. We publicly release our dataset and the implementation of our model at: https://github.com/VinAIResearch/JointIDSF.

Index Terms: Intent detection, Slot filling, Vietnamese language understanding, Joint learning

1. Introduction

Spoken language understanding (SLU) is a crucial component of task-oriented dialogue systems, that typically handles natural language understanding tasks including intent detection and slot filling. In particular, intent detection aims to identify speaker's intent from a given utterance, while slot filling is to extract from the utterance the correct argument value for the slots of the intent [2]. Despite being the 17th most spoken language in the world [3] (about 100M speakers), data resources for Vietnamese SLU are limited. There is only one Vietnamese dataset relevant to intent detection, which is a dialog act corpus containing ISO-24617-2 based annotations over communication acts [4], however, this corpus is not publicly available for the research community. To the best of our knowledge, there is no public Vietnamese dataset available specifically for either intent detection or slot filling.

In this paper, we present the first public dataset for Vietnamese intent detection and slot filling. We create this dataset through three manual phases. The first phase manually translates each English utterance from the well-known intent detection and slot filling dataset ATIS [5] into Vietnamese. Note that this is not a direct translation as performed for the multilingual ATIS datasets [6, 7] where American-specific entities in English are kept intact during translation. In fact, we require modifications to ensure that our Vietnamese utterances are natural, fitting in real-world scenarios in Vietnam and of high-quality, e.g. replacing American-popular slot values, such as locations, airline names and the like, with their counterparts in Vietnam. In the second manual phase, we project intent and slot annotations from each ATIS English utterance to its Vietnamese-translated version. In the last phase, we manually fix inconsistencies among projected annotations in our Vietnamese dataset.

Recent research on intent detection and slot filling have shown that jointly learning these two tasks helps improve performance results [2, 8, 9]. In addition, previous works [10, 11, 12, 13, 14] present attention mechanisms to incorporate intent context information via an utterance representation into slot filling to boost the performances. We argue that instead of using the utterance representation, we can incorporate more explicit intent context information via a "soft" intent label embedding that is computed based on intent prediction probabilities. Thus, we present a new joint model for intent detection and slot filling, that extends JointBERT+CRF [1] with an intent-slot attention layer to explicitly convey the intent context information via the "soft" intent label embedding into slot filling. Our contributions are summarized as follows:

- We introduce the first public intent detection and slot filling dataset—named **PhoATIS**—for Vietnamese.
- We propose a joint model for intent detection and slot filling. Experimental results on our dataset show that: (i) our proposed model does significantly better than JointBERT+CRF; (ii) our attention mechanism is more effective than the previous ones [10, 11, 12, 13, 14]; and (iii) automatic Vietnamese word segmentation and pre-trained monolingual language model are less effective for these Vietnamese intent detection and slot filling tasks than for other Vietnamese NLP tasks [15, 16, 17].
- We publicly release our dataset and our model implementation for research or educational purpose. We hope that our dataset and model can serve as a starting point for future Vietnamese SLU research and applications.

2. Related Work

In addition to ATIS, SNIPS [18] is also commonly used for intent detection and slot filling. However, recent performance scores reported on SNIPS are almost perfect [9, 13, 19], thus resulting in a less challenging dataset. Given the popularity of ATIS, there are efforts to translate it into other languages. In particular, Upadhyay et al. [6] and Xu et al. [7] extend ATIS to eight more languages across four different language families. See [9] for a summary of other intent detection and slot filling datasets.

Early research on intent detection and slot filling tackle these two tasks independently, where intent detection and slot filling are formulated as utterance classification and sequence labeling problems, respectively [20, 21, 22, 23]. Recent studies have shown that jointly learning intent detection and slot filling produces significant performance improvements over independent models [8, 9]. Two joint training strategies are investigated in the literature [2]. The first strategy is through parameter and hidden state sharing, employing a shared BiLSTM/BERT encoder and two separate decoders for intent detection and slot filling that are structured on top of the encoder [1, 24, 25, 26, 27]. The second strategy extends the first one to model the relationship between slots and intent labels. In particular, several research works [10, 11, 12] present attention mechanisms to compute the correlation between a global intent context representation and each slot vector outputted by the encoder; while other research works [13, 14] first learn an utterance representation (i.e.

*equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-618

Table 1: *Statistics of our Vietnamese dataset PhoATIS with 28 intent labels and 82 slot types.*

Statistic	Train	Valid.	Test	All
# Utterances	4478	500	893	5871
# Slots	14859	1713	2842	19414

equivalent to the global intent context representation) through self-attention and concatenate this representation with each of the encoder's vector outputs, before feeding the concatenated vectors into a slot filling decoder. See [9] for an overview of other methods for intent detection and slot filling.

3. Our PhoATIS Dataset

Our dataset construction process includes three manual phases. The first phase is to create a raw natural Vietnamese utterance set that is translated based on the ATIS dataset [5]. The second phase is to project intent and slot annotations from ATIS to its Vietnamese-translated version. The last one is to fix inconsistencies among projected annotations.

First phase of translation: We manually translate all 5871 English utterances from ATIS into Vietnamese, including 4478, 500 and 893 utterances for training, validation and test, respectively [1, 13]. The translation work is done by one NLP researcher and two research engineers with good English proficiency (IELTS 7.0+). We randomly split those English utterances into two non-overlapping and equal subsets. Every utterance from a subset is first translated by one engineer and then cross-checked and corrected by the second engineer; after that, the NLP researcher verifies each translated utterance and makes further revisions if needed. Here, there is a discussion session to finalize the best-translated version for each complicated case.

NOTE that unlike the multilingual ATIS datasets [6, 7] where the original slot values of American-specific entities in English are kept intact when translating into other languages; during our translation phase, we require adaptive modifications to make the translated utterances reflect real-world scenarios in the context of airline booking in Vietnam. In particular, there are 9336 slot values of American locations (e.g. airports, cities, and the like) and other American-popular entities (e.g. ticket codes, airlines, and the like); and the translation process replaces 8837/9336 slot values with their counterparts in Vietnam and all over the world. We also require to preserve spoken modalities (e.g. disfluency, word repetition and collocation) as much as possible to obtain a translated dataset that is correct, natural and similar to the real-world scenarios in Vietnam.

Second phase of annotation projection: We manually project the intent and slot annotations from the original ATIS dataset in English onto our Vietnamese-translated version. In this phase, each utterance in our Vietnamese dataset would have the same intent and slot label types as its corresponding English utterance. This annotation projection process is performed independently by the two research engineers. We again divide the dataset into 2 non-overlapping and equal subsets in which each subset is then annotation-projected by one engineer. This is a non-trivial annotation task because slot values and word orders are different between English utterances and their Vietnamese counterparts. After that, cross-checking is performed to ensure that there are no projection mistakes.

Third phase of fixing inconsistencies: Previous works also point out that there are errors in ATIS reference labels [28, 29].

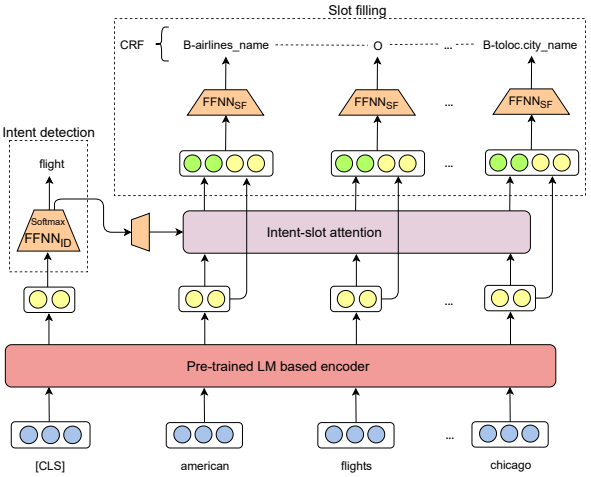

Figure 1: *Illustration of our proposed model JointIDSF.*

For example, all occurrences of the word token "noon" are labeled with the slot label type "time" in the training set, however, they are annotated with the type "period_of_day" in the validation set despite having similar contexts (e.g. "noon" in the training utterance "which northwest flights stop in denver before **noon**" and "noon" in the validation utterance "please list only the flights from cleveland to dallas that leave before **noon**" have labels "time" and "period_of_day", respectively). Regarding intent detection, there are also inconsistencies, e.g. the utterance "what is the lowest fare from denver to atlanta" annotated with the intent label "airfare" while the utterance "show me the lowest price from dallas to baltimore" annotated with the intent "flight".

For our Vietnamese dataset, we also find similar inconsistent labels in both intent detection and slot filling when projecting the annotations in the previous second phase. We thus host another discussion session to refine annotation guidelines for handling annotation inconsistencies in our dataset. Then, based on the refined guidelines, we revisit each annotated Vietnamese utterance to make further corrections if needed,[1] producing a final Vietnamese dataset of 5871 gold annotated utterances with 28 intent labels and 82 slot types. Statistics of our dataset are shown in Table 1.

4. Our Model

Figure 1 illustrates the architecture of our joint model—named **JointIDSF**—that consists of four layers including: an encoding layer (i.e. encoder), an intermediate intent-slot attention layer and two decoding layers of intent detection and slot filling.

Encoding layer: Given an utterance consisting of n tokens $w_1, w_2, \ldots, w_n$, we insert a special classification token of "[CLS]" at the front of the utterance, resulting in an input utterance of $n+1$ tokens $w_0, w_1, w_2, \ldots, w_n$ for the encoding layer (here, w_0 is "[CLS]"). The encoding layer employs a pre-trained Transformer-based language model (LM), e.g. BERT [30], to produce contextualized latent feature embeddings $\mathbf{c}_i \in \mathbb{R}^{d_e}$ each representing the i^{th} token w_i:

$$\mathbf{c}_i = \text{PretrainedLM}\big(w_{0:n}, i\big) \quad (1)$$

Here, d_e is the embedding size of the encoder's contextualized embedding outputs.

[1] Compared to the Vietnamese dataset outputted from the second phase, there are 146 changes in intent labels and 91 changes in slot annotations, across 198 utterances.

Intent detection layer: Following a common strategy when fine-tuning pre-trained LMs for a sequence classification task [30], the intent detection layer is a linear prediction layer that is appended on top of the contextualized embedding $\mathbf{c}_0$ of the classification token "[CLS]". In particular, the intent detection layer feed $\mathbf{c}_0$ into a single-layer feed-forward network (FFNN_{ID}) followed by a softmax predictor for intent prediction:

$$\mathbf{p} = \mathsf{softmax}\big(\text{FFNN}_{\text{ID}}\big(\mathbf{c}_0\big)\big) \tag{2}$$

where the output layer size of FFNN_{ID} is k being the total number of intent labels. Here, we formulate the prediction task as a multi-class classification problem. Based on the probability vector $\mathbf{p} \in \mathbb{R}^k$, a cross-entropy objective loss $\mathcal{L}_{\text{ID}}$ is calculated for intent classification during training.

Intent-slot attention layer: We introduce an attention mechanism to align the importance of the intent information with each of the original utterance's tokens. In particular, the intent-slot attention layer takes the outputs from the encoding layer and the intent detection layer to produce intent-specific vectors that are then used as part of the input for the slot filling layer. Formally, this attention layer first creates a "soft" intent label embedding $\mathbf{w} \in \mathbb{R}^{d_e}$ by multiplying a label weight matrix $\mathbf{W} \in \mathbb{R}^{d_e \times k}$ with the probability vector $\mathbf{p} \in \mathbb{R}^k$. Then it uses the intent label embedding $\mathbf{w}$ and the contextualized embeddings $\mathbf{c}_i$ to generate the intent-specific vectors $\mathbf{s}_i$ ($i \in \{1, 2, ..., n\}$) as follows:

$$\mathbf{w} = \mathbf{W}\mathbf{p} \tag{3}$$

$$\alpha_i = \frac{\mathsf{exp}(\mathbf{w}^\top \mathbf{c}_i)}{\sum_{j=1}^{n} \mathsf{exp}(\mathbf{w}^\top \mathbf{c}_j)} \tag{4}$$

$$\mathbf{s}_i = \alpha_i \mathbf{w} \tag{5}$$

Slot filling layer: The slot filling layer formulates the slot filling task as a BIO-based sequence labeling problem. First, it creates a sequence of vectors $\mathbf{v}_{1:n}$ in which each $\mathbf{v}_i$ is resulted in by concatenating the intent-specific vector $\mathbf{s}_i$ and the corresponding contextualized embedding $\mathbf{c}_i$:

$$\mathbf{v}_i = \mathbf{s}_i \circ \mathbf{c}_i \tag{6}$$

It then passes each vector $\mathbf{v}_i$ into another FFNN (FFNN_{SF}):

$$\mathbf{h}_i = \text{FFNN}_{\text{SF}}\big(\mathbf{v}_i\big) \tag{7}$$

where the output layer size of FFNN_{SF} is the number of BIO-based slot types. Lastly, the slot filling layer feeds the output vectors $\mathbf{h}_i$ into a linear-chain CRF predictor [31] for slot type prediction. A cross-entropy loss $\mathcal{L}_{\text{SF}}$ is computed for slot filling during training while the Viterbi algorithm is used for inference.

Joint training: The final training objective loss $\mathcal{L}$ of our joint model JointIDSF is the weighted sum of the intent detection loss $\mathcal{L}_{\text{ID}}$ and the slot filling loss $\mathcal{L}_{\text{SF}}$:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{ID}} + (1 - \lambda)\mathcal{L}_{\text{SF}} \tag{8}$$

where the hyper-parameter λ is a mixture weight: $0 < \lambda < 1$.

Discussion: Our JointIDSF can be viewed as an extension of the recent state-of-the-art JointBERT+CRF model [1], where we introduce the intent-slot attention layer to explicitly incorporate intent context information into slot filling. In particular, without the intent-slot attention layer, Equation 6 would become $\mathbf{v}_i = \mathbf{c}_i$, and our model thus reduces to JointBERT+CRF.

Furthermore, our intent-slot attention layer is different from previous attention mechanisms [10, 11, 12] in two important aspects: (i) we propose to use the intent label representation instead of the "[CLS]"-based utterance context representation (i.e. using $\mathbf{w} = \mathbf{W}\mathbf{p}$ in Equation 3 instead of $\mathbf{w} = \mathbf{c}_0$); and (ii) our intent-slot attention layer's outputs are the scalar multiplication between the attention weights and the intent label representation instead of the slot vector representations (i.e. using $\mathbf{s}_i = \alpha_i \mathbf{w}$ in Equation 5 instead of $\mathbf{s}_i = \alpha_i \mathbf{c}_i$). In addition, using $\mathbf{v}_i = \mathbf{c}_0 \circ \mathbf{c}_i$ (instead of $\mathbf{v}_i = \mathbf{s}_i \circ \mathbf{c}_i$ in Equation 6) is equivalent to the approach used in [13, 14]. Our ablation study results in Section 5.2 show the effectiveness of our attention mechanism.

5. Experiments

5.1. Experimental setup

Note that the utterances in our PhoATIS dataset are annotated at the syllable level (as when written, the white space is both used to mark word boundaries as well as to separate syllables that constitute words). To obtain a word-level variant of the dataset, we perform automatic Vietnamese word segmentation by employing RDRSegmenter [32] from VnCoreNLP [33]. For example, a 4-syllable written text "sân bay Nội Bài" (Noi Bai airport) is word-segmented into 2-word text "$\text{sân\_bay}_{\text{airport}}$ $\text{Nội\_Bài}_{\text{Noi\_Bai}}$". Here, the outputs of automatic Vietnamese word segmentation do not affect the span boundaries of slot annotations.

We conduct experiments on our dataset to study: (i) a quantitative comparison between our model JointIDSF and the baseline JointBERT+CRF, (ii) the influence of Vietnamese word segmentation (here, input utterances can be formed in either syllable or word level), and (iii) the usefulness of pre-trained language model-based encoders. Here, we employ XLM-R [34] and PhoBERT [15]—two recent state-of-the-art pre-trained language models that support Vietnamese—as the encoders. XLM-R, a multilingual variant of RoBERTa [35], is pre-trained on a 2.5TB multilingual dataset that contains 137GB of syllable-level Vietnamese texts. PhoBERT, a monolingual variant of RoBERTa for Vietnamese, is pre-trained on 20GB of word-level texts.

For both JointIDSF and JointBERT+CRF, we employ the AdamW optimizer [36] and set the batch size to 32. We also perform grid search on the validation set to select their optimal hyper-parameters with the Adam initial learning rate in {1e-5, 2e-5, 3e-5, 4e-5, 5e-5} and the mixture weight λ in {0.05, 0.1, 0.15, ..., 0.9, 0.95}. We train for 50 epochs, and calculate the average score of the intent accuracy for intent detection and the F_1-score (in %) for slot filling after each training epoch on the validation set. We select the model checkpoint that obtains the highest average score over the validation set to apply to the test set. Note that our JointIDSF implementation initializes its encoder by a trained JointBERT+CRF's encoder. All our reported results are the average over 5 runs with 5 different random seeds.

5.2. Main results

Table 2 reports the test set results of our JointIDSF and the baseline JointBERT+CRF, employing standard evaluation metrics of the intent accuracy for intent detection, the F_1-score for slot filling and the overall sentence accuracy [2, 9]. The results are categorized into two comparable settings of using the syllable-level dataset and its automatically word-segmented variant for training and evaluation, associated with the encoders XLM-R and PhoBERT, respectively. In each setting, we find that JointIDSF significantly outperforms JointBERT+CRF. Here, the highest improvements are accounted for the sentence accuracy

Table 2: *Results on the test set. "Intent Acc." and "Sent. Acc." denote intent detection accuracy and sentence accuracy, respectively. Each score improvement over JointBERT+CRF with the same encoder is statistically significant with p-value* < 0.05 *(except 97.56 vs 97.42 w.r.t. intent accuracy).*

	Model	Encoder	Intent Acc.	Slot F1	Sent. Acc.
Syll.	JointBERT+CRF	XLM-R	97.42	94.62	85.39
	Our JointIDSF	XLM-R	**97.56**	**94.95**	**86.17**
Word	JointBERT+CRF	PhoBERT	97.40	94.75	85.55
	Our JointIDSF	PhoBERT	**97.62**	**94.98**	**86.25**

Table 3: *Ablation study results on the validation set. Recall that JointBERT+CRF is a simplified variant of JointIDSF when using* $\mathbf{v}_i = \mathbf{c}_i$ *in Equation 6. Each score difference between our full model JointIDSF and its ablated one is significant with p-value* < 0.05 *(except 98.45 vs. 98.35 w.r.t. intent accuracy).*

Model	Intent Acc.	Slot F1	Sent. Acc.
Our JointIDSF$_{\text{PhoBERT encoder}}$	**98.45**	**97.03**	**89.55**
(i) $\mathbf{w} = \mathbf{c}_0$ in Eq. 3	98.05	96.62	88.30
(ii) $\mathbf{s}_i = \alpha_i \mathbf{c}_i$ in Eq. 5	98.10	96.67	88.55
(iii) $\mathbf{v}_i = \mathbf{c}_0 \circ \mathbf{c}_i$ in Eq. 6	98.35	96.78	88.85
JointBERT+CRF$_{\text{PhoBERT encoder}}$	98.20	96.54	88.15

(i.e., 85.39% $\rightarrow$ 86.17% and 85.55% $\rightarrow$ 86.25%), thus showing that our intent-slot attention layer helps better capture correlations between intent labels and slots in the same utterances. We also find that the performances of word-level models are higher, but not significantly, than their syllable-level counterparts. Thus, automatic Vietnamese word segmentation and the pre-trained monolingual language model PhoBERT are less effective for these Vietnamese intent detection and slot filling tasks than for other Vietnamese NLP tasks [15, 16, 17]. It is probably because the utterances in our dataset are domain-specific and medium-length ones with an average length of 15 word tokens.

We perform an ablation study to understand the model influences on the validation set using the word-level setup (here, using the syllable-level setup results in similar findings). In particular, as discussed in Section 4, we investigate the following factors: (i) using the "[CLS]"-based utterance context representation $\mathbf{c}_0$ instead of the intent label representation (i.e. using $\mathbf{w} = \mathbf{c}_0$ in Equation 3), (ii) using the scalar multiplication between the attention weights and the slot vector representations $\mathbf{c}_i$ instead of the intent label representation (i.e. using $\mathbf{s}_i = \alpha_i \mathbf{c}_i$ in Equation 5), and (iii) concatenating the utterance context representation $\mathbf{c}_0$ instead of the attention layer's output vectors $\mathbf{s}_i$ to all the slot vector representations (i.e. using $\mathbf{v}_i = \mathbf{c}_0 \circ \mathbf{c}_i$ in Equation 6). Table 3 shows that all these factors degrade the performance of our full model, clearly showing the more effectiveness of our attention mechanism compared to the previous ones [10, 11, 12, 13, 14].

5.3. Error analysis

We also provide an illustration example to compare prediction outputs of JointIDSF and JointBERT+CRF w.r.t. the validation utterance "chuyến bay nào rời sân_bay vân_đồn đến côn_đảo và hạ_cánh lúc 10 giờ tối" (what flights leave Van Don airport for Con Dao and arrive at 10 pm). Both JointIDSF and JointBERT+CRF predict the intent "flight" for this utterance correctly. In addition, our JointIDSF correctly recognizes "sân_bay$_{\text{airport}}$ vân_đồn$_{\text{Van_Don}}$" as an airport of departure location, that is tagged with the slot type "fromloc.airport_name". However, JointBERT+CRF is failed, tagging "sân_bay$_{\text{airport}}$" with the slot type "fromloc.city_name"—the name of a departure city, and not recognizing "vân_đồn$_{\text{Van_Don}}$" as a part of any slot. Another example is "cho tôi danh_sách các chuyến bay vào ngày 27 tháng 12 từ đài_bắc đến singapore và giá vé tương_ứng" (give me a list of flights on 27 December from Taipei to Singapore and their corresponding airfare). Both JointIDSF and JointBERT+CRF predict slots correctly. However, while our JointIDSF produces a correct intent label "airfare#flight", JointBERT+CRF produces an incorrect intent label of "flight".

Table 4: *Counts for error types on the validation set of our JointIDSF$_{\text{PhoBERT encoder}}$ (average over the 5 different runs).*

Definition	#errors
Wrong Intent (**WI**): Predicted intent label is not the gold-annotated one.	8
Missing Slot (**MS**): A gold slot's span is not entirely or partly recognized.	5
Spurious Slot (**SS**): A predicted slot matches a gold O label.	10
Wrong Boundary (**WB**): A predicted slot's span is partly overlapped with a gold slot's span, while the predicted slot's label type is the gold slot's.	14
Wrong Label (**WL**): The predicted slot has exact span boundary while having incorrect slot label.	29

To understand the source of errors, we perform an error analysis using the best performing model JointIDSF$_{\text{PhoBERT encoder}}$ on the word-level validation set. We categorize all error cases into different categories of WI, MS, SS, WB and WL, as listed in Table 4. There are 8 errors counted for the WI category, and most of them are induced by the multi-intent labels since the model is likely to predict the most clearly manifested or first appeared intent. For example, the model predicts an intent label of "airfare" instead of the gold one "airfare#flight_time" for the utterance "cho tôi biết chi_phí và thời_gian của các chuyến bay từ phú_quốc đến cam_ranh" (show me the cost and time for flights from Phu Quoc to Cam Ranh). There are 5 and 10 errors counted for the error categories MS and SS, respectively. For these two types of errors, the model is often ambiguous about the slot types that rarely appear in the training set such as "connect", "airport_code" and the like. The WB error category has 14 error cases that are related to multi-word spanned slots. The remaining 29 error cases are accounted for the WL error category. These cases often are induced by ambiguities between which the "departure" part is and which the "arrival" part is in an utterance since many utterances do not have an explicit context. For example, given the utterance "tôi cần đến phú_quốc vào tối thứ_tư từ đà_lạt" (I need to go to Phu Quoc on Wednesday's night from Da Lat), it is relatively ambiguous to determine whether the phrase "tối thứ_tư" (Wednesday's night) refers to as an arrival time or a departure time without a clearer context.

6. Conclusion

In this paper, we have presented the first public dataset for Vietnamese intent detection and slot filling. In addition, we also have proposed an effective model, namely JointIDSF, for jointly learning intent detection and slot filling. In particular, JointIDSF extends the recent state-of-the-art JointBERT+CRF [1] by introducing the intent-slot attention layer to incorporate intent context information into slot filling explicitly. We empirically conduct experiments and perform a detailed error analysis on our dataset, and show that: JointIDSF significantly outperforms JointBERT+CRF and our attention mechanism is more effective than the previous ones [10, 11, 12, 13, 14]. We hope that the public release of our dataset and JointIDSF implementation can serve as the starting point for further research and applications in Vietnamese spoken and natural language understanding.

7. References

[1] Q. Chen, Z. Zhuo, and W. Wang, "BERT for Joint Intent Classification and Slot Filling," *arXiv preprint*, vol. arXiv:1902.10909, 2019.

[2] S. Louvan and B. Magnini, "Recent Neural Methods on Slot Filling and Intent Classification for Task-Oriented Dialogue Systems: A Survey," in *Proceedings of COLING*, 2020, pp. 480–496.

[3] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the World, 22nd edition*. United States: SIL International, 2019.

[4] T. L. Ngo, P. K. Linh, and T. Hideaki, "A Vietnamese Dialog Act Corpus Based on ISO 24617-2 standard," in *Proceedings of LREC*, 2018.

[5] P. J. Price, "Evaluation of spoken language systems: the ATIS domain," in *Proceedings of HLT*, 1990, pp. 91–95.

[6] S. Upadhyay, M. Faruqui, G. Tür, H.-T. Dilek, and L. Heck, "(Almost) Zero-Shot Cross-Lingual Spoken Language Understanding," in *Proceedings of ICASSP*, 2018, pp. 6034–6038.

[7] W. Xu, B. Haider, and S. Mansour, "End-to-End Slot Alignment and Recognition for Cross-Lingual NLU," in *Proceedings of EMNLP*, 2020, pp. 5052–5063.

[8] C. Zhang, Y. Li, N. Du, W. Fan, and P. Yu, "Joint Slot Filling and Intent Detection via Capsule Neural Networks," in *Proceedings of ACL*, 2019, pp. 5259–5267.

[9] H. Weld, X. Huang, S. Long, J. Poon, and S. Han, "A survey of joint intent detection and slot-filling models in natural language understanding," *arXiv preprint*, vol. arXiv:2101.08091, 2021.

[10] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, "Slot-Gated Modeling for Joint Slot Filling and Intent Prediction," in *Proceedings of NAACL*, 2018, pp. 753–757.

[11] C. Li, L. Li, and J. Qi, "A Self-Attentive Model with Gate Mechanism for Spoken Language Understanding," in *Proceedings of EMNLP*, 2018, pp. 3824–3833.

[12] H. E, P. Niu, Z. Chen, and M. Song, "A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling," in *Proceedings of ACL*, 2019, pp. 5467–5471.

[13] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu, "A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Language Understanding," in *Proceedings of EMNLP-IJCNLP*, 2019, pp. 2078–2087.

[14] Z. Zhang, Z. Zhang, H. Chen, and Z. Zhang, "A Joint Learning Framework With BERT for Spoken Language Understanding," *IEEE Access*, vol. 7, pp. 168 849–168 858, 2019.

[15] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of ACL: EMNLP 2020*, 2020, pp. 1037–1042.

[16] A. T. Nguyen, M. H. Dao, and D. Q. Nguyen, "A Pilot Study of Text-to-SQL Semantic Parsing for Vietnamese," in *Findings of EMNLP 2020*, 2020, pp. 4079–4085.

[17] T. H. Truong, M. H. Dao, and D. Q. Nguyen, "COVID-19 named entity recognition for Vietnamese," in *Proceedings of NAACL-HLT*, 2021, p. to appear.

[18] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril *et al.*, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint*, vol. arXiv:1805.10190, 2018.

[19] J. Wang, K. Wei, M. Radfar, W. Zhang, and C. Chung, "Encoding Syntactic Knowledge in Transformer Encoder for Intent Detection and Slot Filling," in *Proceedings of AAAI*, 2021, p. to appear.

[20] G. Tür, L. Deng, D. Hakkani-Tür, and X. He, "Towards deeper understanding: Deep convex networks for semantic utterance classification," in *Proceedings of ICASSP*, 2012, pp. 5045–5048.

[21] S. Ravuri and A. Stolcke, "Recurrent neural network and LSTM models for lexical utterance classification," in *Proceedings of INTERSPEECH*, 2015, pp. 135–139.

[22] G. Mesnil, X. He, L. Deng, and Y. Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding," in *Proceedings of INTERSPEECH*, 2013, pp. 3771–3775.

[23] N. T. Vu, P. Gupta, H. Adel, and H. Schütze, "Bi-directional recurrent neural network with ranking loss for spoken language understanding," in *Proceedings of ICASSP*, 2016, pp. 6060–6064.

[24] P. Xu and R. Sarikaya, "Convolutional neural network based triangular CRF for joint intent detection and slot filling," in *Proceedings of ASRU*, 2013, pp. 78–83.

[25] B. Liu and I. Lane, "Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling," in *Proceedings of INTERSPEECH*, 2016, pp. 685–689.

[26] X. Zhang and H. Wang, "A Joint Model of Intent Determination and Slot Filling for Spoken Language Understanding," in *Proceedings of IJCAI*, 2016, pp. 2993–2999.

[27] D. Hakkani-Tür, G. Tür, A. Celikyilmaz, Y.-N. Chen, J. Gao, L. Deng, and Y.-Y. Wang, "Multi-Domain Joint Semantic Frame Parsing using Bi-directional RNN-LSTM," in *Proceedings of INTERSPEECH*, 2016, pp. 715–719.

[28] F. Béchet and C. Raymond, "Is ATIS too shallow to go deeper for benchmarking Spoken Language Understanding models?" in *Proceedings of INTERSPEECH*, 2018, pp. 3449–3453.

[29] J. Niu and G. Penn, "Rationally Reappraising ATIS-based Dialogue Systems," in *Proceedings of ACL*, 2019, pp. 5503–5507.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL*, 2019, pp. 4171–4186.

[31] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of ICML*, 2001, pp. 282–289.

[32] D. Q. Nguyen, D. Q. Nguyen, T. Vu, M. Dras, and M. Johnson, "A Fast and Accurate Vietnamese Word Segmenter," in *Proceedings of LREC*, 2018, pp. 2582–2587.

[33] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "VnCoreNLP: A Vietnamese Natural Language Processing Toolkit," in *Proceedings of NAACL: Demonstrations*, 2018, pp. 56–60.

[34] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of ACL*, 2020, pp. 8440–8451.

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint*, vol. arXiv:1907.11692, 2019.

[36] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proceedings of ICLR*, 2018.