Joint Intent Detection and Slot Filling with BERT in Burmese

Sann Su Su Yee
Natural Language Processing Lab
University of Computer Studies,
Yangon, Myanmar
sannsusuyee@ucsy.edu.mm

Khin Mar Soe
Natural Language Processing Lab
University of Computer Studies,
Yangon, Myanmar
khinmarsoe@ucsy.edu.mm

***Abstract*— Natural language understanding (NLU) in dialogue system is a core component to interpret the language where intent detection and slot filling are two main subtasks for building a spoken and natural language understanding. In high-resource language, multiple deep learning-based joint models have demonstrated excellent results on the two tasks. However, Burmese can be regarded as a low-resource language due to a lack of pre-created dataset for specific research interested. In this paper, we proposed small-scale human-labeled intent/slot training dataset (consists of 25K utterances), related with transportation and restaurant data, which extracted from MmTravel corpus. For the encoder, we used pretrained BERT (Bidirectional Encoder Representation from Transformers) model to encode the input sequence as context representations, which is deep bidirectional representations from large-scale unlabeled text by jointly conditioning on both left and right context in all layers. For the decoder, we employ an intent deteciton decoder in the initial stage and slot filling decoder to forecast the semantic concept tags for each word using the intent contextual information as second stage. The experimental results of the model achieve state-of-the-art results.**

Keywords— natural language understanding, intent detection, slot filling, MmTravel, pre-trained language model, BERT.

I. BACKGROUND

In recent years, numerous intelligent personal assistants have been developed and obtained great success. The essential component of natural language understanding is essential prerequisite to build an excellent dialogue system, where intent and slot classification are two critical tasks [1]. Intent classification is a kind of sentence classification problem that can predict the intent tag y^{intent} of an utterance, which focus on automatically determining the intent of a user from an utterance. Slot filling is a sequence labeling task that mapping an input word sequence $x = (x_1, x_2, \dots, x_T)$ with the slot label sequence $y^{slot} = (y_1^{slot}, y_2^{slot}, \dots, y_T^{slot})$. In our dataset, we using labeled data as SLOT and O (out of scope slot) labels across domains, which can improve the non-slot token reduction step in the target domain and slot name prediction step. Table 1 shows a sample user query of intent detection and slot filling "ရန်ကုန် ရထား လက်မှတ် နှစ် စောင် ပေး ပါ, Give me two Yangon railway tickets".

Deep learning models have been widely researched in NLU which is the recent trends and based on machine reading comprehension on dialog or text to interpret the language. NLU models has been divided into traditional pipeline approaches and joint modelling approaches based on whether intent and slot classification are modelled separately or jointly. Intent detection or recognition, which is regarded as an utterance classification problem to predict an intent label. This classification is based on what the user wants to achieve, which can also be denoted as dialogue acts. The intent classifier can be used first to recognize the user help to understand the user wants, needs, and purposes such as inquire for bus schedules, reservation a table, etc. This task can be modeled using LSTM by Ravuri and Stolcke in 2015, CNN by Zhang et al. in 2015, attention-based CNN by Zhao in 2016, hierarchical attention networks by Yang et al. in 2016, and others. Slot filling approaches include RNN-EM by Peng et al. in 2015, deep LSTM by Yao et al. in 2014, encoder-labeler deep LSTM by Kurata et al. in 2016 and CNN by Vu et al. in 2016, among others. Considering this strong correlation between the two tasks, the tendency is to develop a joint model using CNN-CRF by Xu and Sarikaya et al. in 2013, joint RNN-LSTM by Hakkani et al. in 2016 and attention-based BiRNN by Goo et al. in 2018.

TABLE I. AN EXAMPLE WITH INTENT AND SLOT ANNOTATION, WHICH INDICATES THE SLOT OF TO_PLACE_NAME, TRANSPORT_TYPE AND NUMBER FROM AN UTTERANCE WITH AN INTENT TRANSPORT_BUY_TICKET .

Utt.	ရန်ကုန်	ရထား	လက်မှတ်	နှစ် စောင်	ပေး	ပါ
	yə̃ kòʊ̃	yatʰá	leʔ maʔ	niʔ saʊ̃	péi	pà
	(Give me two Yangon railway tickets.)					
Slot	to_place_name	transport_type	O	number	O	O
Intent	transport_buy_ticket					

However, rely on deep learning approaches require a large amount of data for specific domain. If there are lack of human-annotated data result in poor generalization capabilities, intent detection and slot filling might be highly challenge task. It is also quite difficult for deep learning-based models to produce synthetic utterances with semantic preservation in limited data settings. Recently, language model pre-trained have achieved great success on addressing the data sparsity challenge, which are using an enormous amount of unlabeled data for training, such as ELMo [2], Generative Pre-trained Transformer (GPT) [3], and BERT [4]. Pre-trained models can be fine-tuned on NLP tasks and have achieved significant improvement over training on task-specific annotated data. The key contributions of this paper are as follows:

(1) We introduce human-annotated intent/slot Burmese dataset; (2) We proposed BERT based joint intent detection and slot filling model which design a two-state decoder process to address the poor generalization capability and illustrate that the proposed model achieves state-of-the-art result. We hope that the dataset and model can serve as a starting point for future Burmese NLU research and applications.

II. EVALUATING WITH MINIMAL DATA

A. Dataset

The MmTravel corpus [5] was designed to cover utterances for all potential topics in travel conversations. Since it is almost infeasible to collect them by transcribing actual conversations or simulated dialogs during the period, utterances in the corpus are built on the ASEAN-MT corpus. To determine the effectiveness of of our proposed model, we conduct experiments using our human-labeled intent/slot training dataset. The intent/slot Burmese dataset is extracted

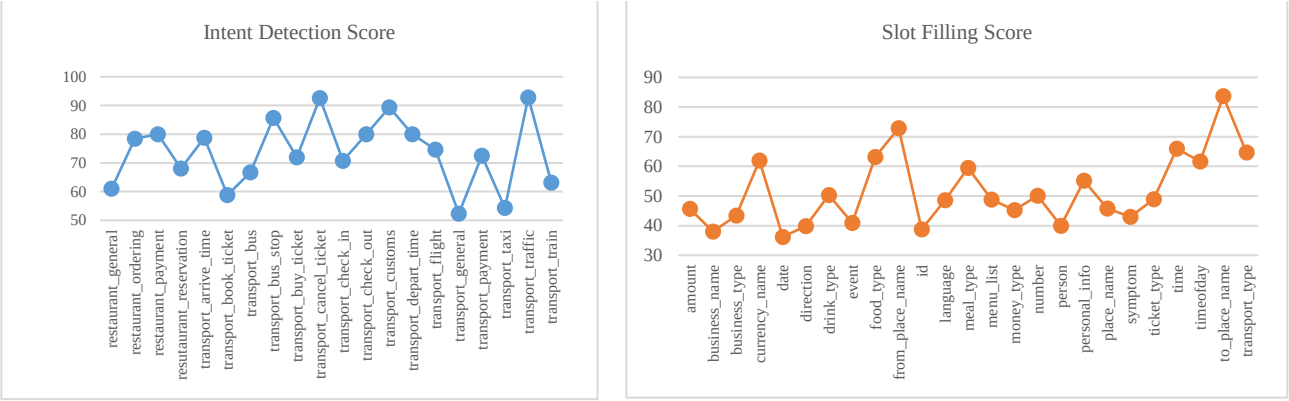

Fig. 1. Experimental results (F1 scroes %) on the test set with pre-trained BERT model on intent/slot Burmese dataset.

25K utterances from MmTravel corpus across domains transportation and restaurant data. The statistics of the datasets are summarized in Table II and there are 20 intent labels and 25 slot types in the training set.

TABLE II. STATISTICS OF SMALL-SCALE HUMAN-LABELED INTENT/SLOT DATASET WITH 20 INTENT LABELS AND 24 SLOT TYPES

Statistic	Train	Valid.	Test	All
# Utterances	19,854	2,039	3,823	25,485
# Slots	188,289	19,562	36,679	244,532

B. BERT Models

Recently, the language pre-trained model has been demonstrated to be highly effective at learning universal language representations by using large amounts of unlabeled data. In 2019, Devlin et al. [4] proposed bidirectional encoder representation from transformer (BERT) which is a contextualized word representation model based on a multilayer bidirectional transformer-encoder that uses parallel attention layers in the transformer neural network rather than sequential recurrence. BERT comes in two sizes: BERT-Base and BERT-Large, both of which have been pre-trained. BERT can be used with either unannotated data from the pre-trained model or task-specific data to fine-tune it.

Sentence Representation: Our proposed model used BERT as encoder, the input representation is a concatenation of WordPiece embeddings, the segment embedding and positional embeddings. In particular, the segment embedding has no discrimination for single sentence classification and labelling tasks. The first token has been inserted by special classification embedding ([CLS]) and a special token ([SEP]) is added as the final token. The BERT encoder computes the semantic representations of the sentence given an input token sequence $x = (x_1, \dots, x_T)$, and the output is $H = BERT\ (x_1, \dots, x_T)$.

Joint Modeling: The intent label prediction of the y^{intent} based on the utterance semantic representation H' according to: $y^{intent} = softmax\ (W^{intent} H' + b^{intent})$. For slot filling, we feed the hidden states $(h_1^s, \dots, h_T^s)$ of slot filling decoder into a softmax layer to predict the slot tags. Each tokenized input word is passed via a Word-Piece tokenizer in order to make this process compatible with WordPiece tokenization, and the hidden state corresponding to the first sub-token is then provided as the input to the softmax classifier, which can be denoted as: $y_i^{slot} = softmax\ (W^{intent} h_i^s + b^{slot})$, where h_i^s is the hidden state corresponding to the first sub-token of word x_i. The formulated of the jointly model is $p(y^S, y^I|x) = p(y^I|x) \prod_{t-1}^{T} p(y_t^s|x)$, where $p(y^S, y^I|x)$ is the conditional probability of the result given the input word sequence x. The model is end-to-end fine-tuned by minimizing the cross-entropy loss.

C. Implementation and Results

In this paper, we implemented our model by using the PyTorch framework and built on pre-trained $\text{BERT}_{\text{BASE}}$ model which has 12-layer transformer blocks, 768 hidden states, and 12 self-attention heads, totally 110M parameters. On the development set, all hyper-parameters have been fine-tuned. The longest utterance is 50 words. The additional criteria are based on BERT articles. Adam is utilized for optimization with a 2e-5 initial learning rate. We used weighted loss to enables weighted class balancing of the loss. Figure 1 shows the experimental results of the proposed models for each intent and slot label on intent/slot Burmese datasets. Joint BERT achieves the state-of-the-art result with the intent classification F1-score of 76.85, and slot filling F1 of 60.29%.

III. CONCLUSION AND FUTURE WORK

In this work, we have presented the hand-annotated Burmese dataset for joint intent detection and slot filling. In addition, we proposed BERT architecture based joint learning for intent detection and slot filling model. Experimental results show that the proposed joint BERT model does not achieve high performance as expected. However, we believe that the dataset and model might act as a foundation for further research. In our future work, we will leverage the monolingual BERT model from the UCSY NLP team to enhance the performance of joint intent and slot classification.

REFERENCES

[1] G. Tur and R. De Mori, "Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. Hoboken", NJ, USA:Wiley, 2011.

[2] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations", In NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 2227–2237.

[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning", In Technical report, OpenAI, 2018.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding,", in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., vol. 1, Jun. 2019.

[5] S. S. S. Yee, K. M. Soe, "Myanmar Dialogue Act Recognition Using Bi-LSTM RNN", 23rd Conference of the Oriental COCOSDA, (Oriental COCODSA 2020), Yangon, Myanmar, November 2020.