Modeling Long Context for Task-Oriented Dialogue State Generation

Jun Quan and **Deyi Xiong***
School of Computer Science and Technology, Soochow University, Suzhou, China
terryqj0107@gmail.com, dyxiong@suda.edu.cn

Abstract

Based on the recently proposed transferable dialogue state generator (TRADE) (Wu et al., 2019) that predicts dialogue states from utterance-concatenated dialogue context, we propose a multi-task learning model with a simple yet effective utterance tagging technique and a bidirectional language model as an auxiliary task for task-oriented dialogue state generation. By enabling the model to learn a better representation of the long dialogue context, our approaches attempt to solve the problem that the performance of the baseline significantly drops when the input dialogue context sequence is long. In our experiments, our proposed model achieves a 7.03% relative improvement over the baseline, establishing a new state-of-the-art joint goal accuracy of 52.04% on the MultiWOZ 2.0 dataset.

1 Introduction

Dialogue state tracking (DST, also known as belief tracking) predicts user's goals in task-oriented dialogue system, where dialogue states are normally represented in the form of a set of slot-value pairs. A variety of approaches to dialogue state tracking are devoted to dealing with two different settings: DST over a predefined domain ontology and DST with slot-value candidates from an open vocabulary. Most of the previous work is based on the first setting, assuming that all possible slot-value candidates are provided in a domain ontology in advance. The task of the dialogue state tracking with this setting is therefore largely simplified to score all predefined slot-value pairs and select the value with the highest score for each slot as the final prediction. Although predefined ontology-based approaches are successfully used on datasets with small ontologies, such as DSTC2 (Henderson et al., 2014) and WOZ2.0 (Wen et al., 2017), they are quite limited in both scalability to scenarios with infinite slot values and prediction of unseen slot values.

In order to address these issues of DST over predefined ontologies, recent efforts have been made to predict slot-value pairs in open vocabularies. Among them, TRADE (Wu et al., 2019) proposes to encode the entire dialogue context and to predict the value for each slot using a copy-augmented decoder, achieving state-of-the-art results on the MultiWOZ 2.0 dataset (Budzianowski et al., 2018). As TRADE simply concatenates all the system and user utterances in previous turns into a single sequence as the dialogue context for slot-value prediction, it is difficult for the model to identify whether an utterance in the dialogue context is from system or user when the concatenated sequence becomes long. We observe that the longest dialogue context after concatenation on the MultiWOZ 2.0 dataset contains 880 tokens. Our experiments also demonstrate that the longer the dialogue context sequence is, the worse TRADE performs.

To deal with this problem, we propose two approaches to modeling long context for better dialogue state tracking. The first method is tagging. While constructing the dialogue context sequence, we insert a tag of *[sys]* symbol in front of each system utterance, and a tag of *[usr]* symbol in front of each user utterance. The purpose of adding such symbolic tags in the concatenated dialogue context sequence is to explicitly enhance the capability of the model in distinguishing system and user utterances. In the second method, we propose to integrate a bi-directional language modeling module into the upstream of the model as an auxiliary task to gain better understanding and representation of the dialogue context. The bi-directional language modeling task is to predict the next word by using forward hidden states and the previous word by using backward hidden states based on the

*Corresponding author

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7119–7124

dialogue context sequence without any annotation. With these two approaches, we perform dialogue state tracking in a multi-task learning architecture.

In summary, the contributions of our work are as follows:

- We propose a simple tagging method to explicitly separate system from user utterances in the concatenated dialogue context.
- We propose a language modeling task as an auxiliary task to better model long context for DST.
- We conduct experiments on the MultiWOZ 2.0 dataset. Both methods achieve significant improvements over the baselines in all evaluation metrics. The joint of the two methods establish a new state-of-the-art results on the MultiWOZ 2.0. In addition, we provide a detailed analysis on the improvements achieved by our methods.

2 Related Work

Predefined ontology-based DST assumes that all slot-value pairs are provided in an ontology. Mrkšić et al. (2017) propose a neural belief tracker (NBT) to leverage semantic information from word embeddings by using distributional representation learning for DST. An extension to the NBT is then proposed by Mrkšić and Vulić (2018), which learns to update belief states automatically. Zhong et al. (2018) use slot-specific local modules to learn slot features and propose a global-locally self-attentive dialogue state tracker (GLAD). Nouri and Hosseini-Asl (2018) propose GCE model based on GLAD by using only one recurrent networks with global conditioning. Ramadan et al. (2018) introduce an approach that fully utilizes semantic similarity between dialogue utterances and the ontology terms. Ren et al. (2018) propose StateNet which generates a fixed-length representation of the dialogue context and compares the distances between this representation and the value vectors in the candidate set for making prediction. These predefined ontology-based DST approaches suffer from their weak scalability to large ontologies and cannot deal with previously unseen slot values.

In open vocabulary-based DST, Xu and Hu (2018) propose a model that learns to predict unknown values by using the index-based pointer network for different slots. Wu et al. (2019) apply an encoder-decoder architecture to generate dialogue states with the copy mechanism. However, their method simply concatenates the whole dialogue context as input and does not perform well when the dialogue context is long. We study this problem and propose methods to help the DST model better model long context. Inspired by Zhou et al. (2019) who use an additional language model in question generation, we attempt to incorporate language modeling into dialogue state tracking as an auxiliary task.

3 Our Methods

In this section, we describe our proposed methods. First, section 3.1 briefly introduces the recent TRADE model (Wu et al., 2019) as background knowledge, followed by our methods: utterance tagging in section 3.2 and multi-task learning with language modeling in section 3.3.

3.1 Transferable Dialogue State Generator

TRADE is an encoder-decoder model that encodes concatenated previous system and user utterances as dialogue context and generates slot value word by word for each slot exploring the copy mechanism (Wu et al., 2019). The architecture of TRADE is shown in Figure 1 without the language model module. In the encoder of TRADE, system and user utterances in previous dialogue turns are simply concatenated without any labeling. In our experiments, we find that the performance of the TRADE model significantly drops when the length of the dialogue context is long. On the MultiWOZ 2.0 dataset, the maximum length of a dialogue context is up to 880 tokens. About 27% of instances on the test set have dialogue context sequences longer than 200 tokens. The joint accuracy of the TRADE on these cases drops to lower than 22%. This suggests that TRADE suffers from long context.

3.2 Utterance Tagging

To deal with this problem, we first propose a simple method to label system and user utterances by inserting a tag of *[sys]* just at the beginning of each system utterance and a tag of *[usr]* in front of each user utterance when they are concatenated into the dialogue context. We conjecture that mixing system and user utterances in one single sequence may confuse the encoder. It may also mislead the decoder to attend to inappropriate parts and the copy network to copy from wrong utterances. The

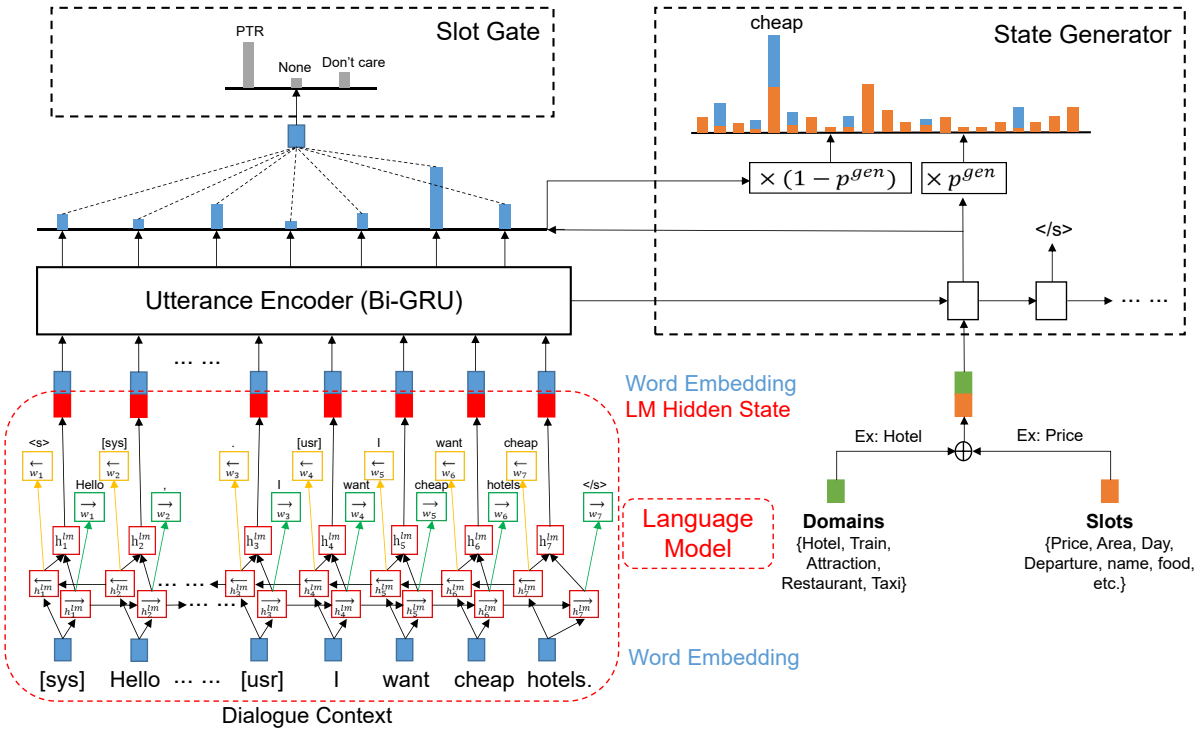

Figure 1: Multi-task Learning Framework with Language Modeling Task for Dialogue State Tracking

explicit indicators from the two tags are to help TRADE differ system from user utterances.

3.3 Multi-task Learning with Language Modeling

We further propose to incorporate a bi-directional language modeling module into the dialogue state tracking model in a multi-task learning framework for DST, which is shown in Figure 1.

The bi-directional language modeling module is to predict the next word and the previous word in the concatenated sequence with the forward and the backward GRU network respectively. We first feed the concatenated dialogue context into the embedding layer. We initialize each word embedding in the dialogue context by concatenating Glove embedding (Pennington et al., 2014) and character embedding (Hashimoto et al., 2017). This word embedding sequence is then fed into a bi-directional GRU network to get the hidden representations $\overrightarrow{h_t^{lm}}$ and $\overleftarrow{h_t^{lm}}$ in two directions, which are used to predict the next and the previous word through a softmax layer as follows:

$$P^{lm}(w_{t+1}|w_{<t+1}) = softmax(W_f \overrightarrow{h_t^{lm}}) \quad (1)$$

$$P^{lm}(w_{t-1}|w_{>t-1}) = softmax(W_b \overleftarrow{h_t^{lm}}) \quad (2)$$

The loss function is defined as the sum of the negative log-likelihood of the next and previous words in the sequence. The language modeling loss L^{lm} is therefore calculated as follows (T is the length of the concatenated dialogue context sequence):

$$L^{lm} = -\sum_{t=1}^{T-1} log(P^{lm}(w_{t+1}|w_{<t+1})) - \sum_{t=2}^{T} log(P^{lm}(w_{t-1}|w_{>t-1})) \quad (3)$$

The sum of the forward and backward hidden states in the language model module is used as the hidden representation h_t^{lm} for word w_t in the dialogue context: $h_t^{lm} = \overrightarrow{h_t^{lm}} + \overleftarrow{h_t^{lm}}$. We further sum it with the word embedding of w_t and feed the sum into the utterance encoder. Following Wu et al. (2019), we include the slot gate and state generator modules in our model and calculate the dialogue state tracking loss L^{dst}.

The training objective for the multi-task learning framework is to minimize the total loss L^{total} which is the sum of DST and language modeling loss:

$$L^{total} = L^{dst} + \alpha L^{lm} \quad (4)$$

where α is a hyper-parameter which is used to balance the two tasks.

Model	Joint Accuracy	Slot Accuracy
Baselines		
GLAD (Zhong et al., 2018)	35.57	95.44
TRADE (Wu et al., 2019)	48.62	96.92
COMER (Ren et al., 2019)	48.79	-
NADST (Le et al., 2020)	50.52	-
SOM-DST (Kim et al., 2019)	51.38	-
DSTQA (Zhou and Small, 2019)	51.44	97.24
Ours		
Ours	**52.04**	**97.26**
-LM	50.15	97.10
-Tagging	51.36	97.23

Table 1: Experimental results on the MultiWOZ 2.0 dataset.

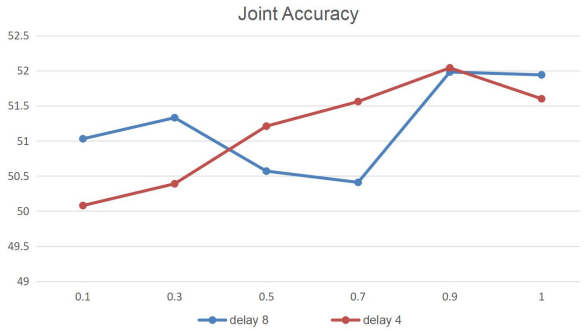

Figure 2: The impact of hyper-parameter α and delay update step on DST joint accuracy.

4 Experiments

In this section, we evaluated our proposed methods on the public dataset.

4.1 Datasets & Settings

We conducted experiments on the MultiWOZ 2.0 (Budzianowski et al., 2018) which is the largest multi-domain task-oriented dialogue dataset, consisting of over 10,000 dialogues from seven domains. Each dialogue is composed of 13.68 turns on average. Following Wu et al. (2019), we used five domains excluding *hospital* and *police* domains which account for a small portion and do not appear on the test set.

In our multi-task learning model, both the sizes of hidden states and word embeddings were set to 400. We set the batch size to 8 and applied the delay update mechanism with different step sizes to train the model.

4.2 Results

Joint accuracy and slot accuracy are the two metrics we used to evaluate the performance on dialogue state tracking. Table 1 shows the results of our methods and other baselines on the test set of the MultiWOZ 2.0 dataset. Our full model (tagging + language modeling) significantly outperforms several previous state-of-the-art models, including TRADE, and achieves new state-of-the-art results, 52.04% of joint accuracy and 97.26% of slot accuracy on the MultiWOZ 2.0. The tagging alone (-LM) can improve the joint accuracy on the MultiWOZ 2.0 by 1.53% while the auxiliary language modeling (-Tagging) by 2.74%.

Figure 2 shows the impact of α and the number of delay update steps on DST. Consequently, our model performs best when we set α to 0.9 and the number of delay update steps to 4.

Length	Total	Correct Turns		Joint Accuracy(%)	
		TRADE	Ours	TRADE	Ours
0 - 99	2,940	2,115	2,190 (+75)	71.94	74.49 (+2.55)
100-199	2,466	1,028	1,129 (+101)	41.69	45.78 (+4.09)
200-299	1,494	356	445 (+89)	23.83	29.79 (+5.96)
$\geqslant$ 300	468	57	70 (+13)	12.18	14.96 (+2.78)

Table 2: Results and statistics on different lengths of dialogue context on the test set.

Model	Total	Correct	Not exactly correct		
			Over pred.	Partial pred.	False pred.
TRADE	7,368	3,556	791	1,480	1,541
Ours	7,368	**3,834 (+278)**	877 (+86)	**1,201 (-279)**	**1,456 (-85)**

Table 3: Statistics and analysis on different types of prediction errors. The red indicates positive effects, while the blue indicates negative effect.

4.3 Analysis

We further provide a deep analysis on our results on the MultiWOZ 2.0 according to the length of concatenated dialogue context, which are shown in Table 2. We can clearly observe that the performance of the baseline model drops sharply with the increase of the dialogue context length. We can also find that our model performs better than the baseline in all cases, suggesting that the proposed methods are able to improve modeling long dialogue context for DST.

Table 3 shows the statistics of different kinds of prediction errors on the test set of the MultiWOZ 2.0. We define three types of dialogue state prediction errors. Over prediction is that the predicted states not only fully cover the golden states, but also include some redundant slot values. Partial prediction is an error that the predicted states are just part of the golden states with some slot values missing. False prediction denotes that false slot values are predicted for some slots. As shown in Table 3, our model significantly reduces the number of partial and false prediction errors, with the help of better representation of dialogue context.

5 Conclusion

In this paper, we have presented the utterance tagging and auxiliary bi-directional language modeling in a multi-task learning framework to model long dialogue context for open vocabulary-based DST. Experiments on the MultiWOZ 2.0 dataset show that our model significantly outperforms the baselines and achieves new state-of-the-art results.

Acknowledgments

The present research was supported by the National Natural Science Foundation of China (Grant No. 61861130364) and the Royal Society (London) (NAF\R1\180122). We would like to thank the anonymous reviewers for their insightful comments.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Kazuma Hashimoto, Yoshimasa Tsuruoka, Richard Socher, et al. 2017. A joint many-task model: Growing a neural network for multiple nlp tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2019. Efficient dialogue state tracking by selectively overwriting memory. *arXiv preprint arXiv:1911.03906*.

Hung Le, Richard Socher, and Steven C.H. Hoi. 2020. Non-autoregressive dialog state tracking. In *International Conference on Learning Representations*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788.

Nikola Mrkšić and Ivan Vulić. 2018. Fully statistical neural belief tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 108–113.

Elnaz Nouri and Ehsan Hosseini-Asl. 2018. Toward scalable neural dialogue state tracking model. In *Advances in neural information processing systems (NeurIPS), 2nd Conversational AI workshop*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Osman Ramadan, Paweł Budzianowski, and Milica Gasic. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 432–437.

Liliang Ren, Jianmo Ni, and Julian McAuley. 2019. Scalable and accurate dialogue state tracking via hierarchical sequence generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1876–1885.

Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards universal dialogue state tracking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2786.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.

Puyang Xu and Qi Hu. 2018. An end-to-end approach for handling unknown slot values in dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1448–1457.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1458–1467.

Li Zhou and Kevin Small. 2019. Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering. *arXiv preprint arXiv:1911.06192*.

Wenjie Zhou, Minghua Zhang, and Yunfang Wu. 2019. Multi-task learning with language modeling for question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3385–3390, Hong Kong, China. Association for Computational Linguistics.