Multi-domain Task Oriented Dialogue System Using Word Constraints

Chang-Uk Shin and Jeong-Won Cha

Artificial Intelligence Research Lab., Changwon National University, Republic of Korea
{papower1, jcha}@changwon.ac.kr

Abstract

We introduce a new modeling method for multi-domain dialogue modeling. The proposed method infers domain and dialogue acts from utterance for a multi-domain dialogue model. With the inferred domain and dialogue act information, the model constrains the system utterance dictionary to enable the generation of utterances corresponding to the domain. In the experiments set up for the MultiWOZ dataset provided by the DSTC8 Track1 Multi-domain task completion track, the proposed method achieves 0.032 higher BLEU-1 score than the baseline method. We found that the modeling method and the results in this paper show that the domain estimation and dialogue act estimation in dialog modeling improve the performance. Also, when we implement the proposed dictionary constraint method, we could improve the domain cohesion performance and the utterance generation performance by limiting system utterance words.

Introduction

A task-oriented dialog system refers to a system that communicates with a user on a specific subject in a specific domain and helps to achieve the user's purpose. Therefore, the high-performance natural language understanding and dialog status tracking module should be able to identify the user's intention and track the achievement of the goal. Previously, research on single-domain task-oriented dialogue modeling has been conducted. The Multi-Domain Task Completion track, the track1 of DSTC8, was held to build a task-oriented dialogue system for seven domains including hotels.

Many tasks in the field of natural language processing use artificial neural networks to model languages and achieve high performance. In the case of dialog modeling, there has been a case where high performance is achieved by modeling by E2E method using an artificial neural network.

Expansion from a single domain to multiple domains leads to an increase in modeling difficulty. First, the categories and vocabulary of entity names to be processed increase. These entity vocabularies are difficult to model because they appear in small amounts in the dataset. Second, the types and distribution of dialog states in each domain are different. And the modeling complexity becomes larger when domain transfer is freely performed. Due to this increase in modeling difficulty, the performance of the final model is reduced. Therefore, there is a need for a study on how to achieve high performance in the multi-domain.

We aim to improve performance through domain and dialogue act estimation in multi-domain task-oriented dialog systems. The domain inference module and dialogue act inference module estimate the domain and dialogue act of the current user utterance and provide the features to the system utterance generation module, respectively. These features directly intervene in system utterance generation, placing constraints on system utterance generation words. This aims to increase domain cohesion of system utterance and reduce modeling complexity to improve modeling performance.

Related Research

Before the neural network, dialogue modeling was performed by connecting two or more statistics-based modules. This module usually includes a natural language understanding module, a dialogue state tracking module, and a utterance generation module. Each module was learned individually. The natural language understanding module performs language analysis on the input utterance. Afterward, the dialogue status tracking module receives the analysis result to determine the user's intention. Finally, the utterance generation module generates the system utterance based on the result of the dialogue state tracking module. Since these statistics-based dialogue modeling techniques are written separately, the whole module cannot perform inference organically.

As the neural network model was applied to the dialogue modeling study, the end-to-end modeling that performed the whole dialogue modeling as a structure began to be studied.

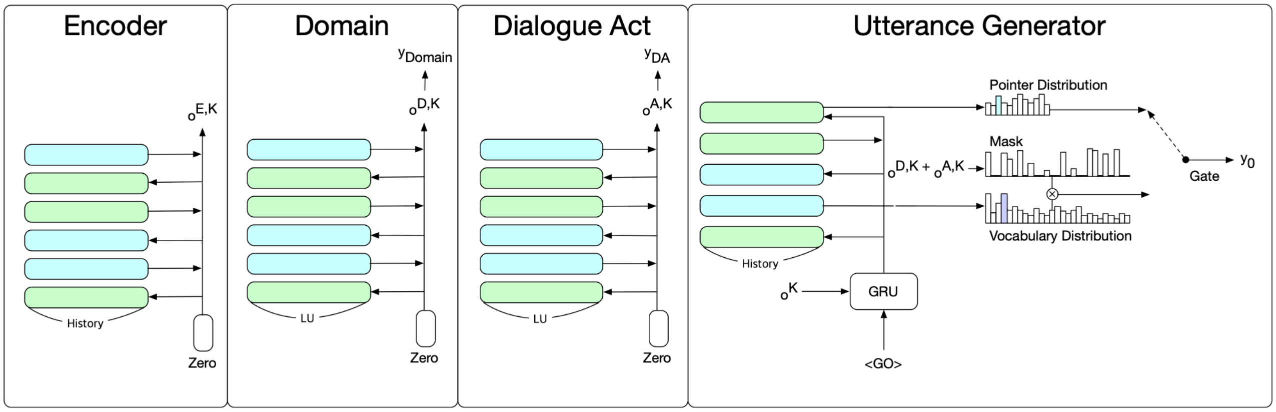

Figure 1 Proposed architecture

Initially, dialogue modeling was performed using the sequence-to-sequence structure(Sutskever et al., 2014, Vinyals et al., 2015).

The sequence-to-sequence structure models the dialogue modeling problem as a source to target transduction problem. It encodes dialogue history or user's last utterance via encoder network using LSTM or GRU. And it generates system utterance based on the distributed representation(Wen et al., 2017). The sequence-to-sequence-based model learns using only dialogue history. Thus, it has the advantage of not requiring the researcher's handcrafted features, compared to the previous statistics-based method.

Since then, improvements have been made to apply the attention mechanism to sequence-to-sequence structures (Bahdanau et al., 2015) or to improve performance by combining external knowledges with rules(Wen et al., 2017, Williams et al., 2017).

On the other hand, there have been cases of dialogue modeling using the structure of End-to-End Memory Networks(Sukhbaatar et al., 2015) (Sakai et al., 2017, Bordes et al., 2017, Madotto et al., 2018, Chen et al., 2019). The RNN-based model updates the state of a sequence when it takes input at each timestep. However, this state can be unstable for long sequences. On the contrary, they pointed out the limitation of RNN and proposed a structure that can compute the entire input simultaneously a certain number of times.

Among them, (Madotto et al., 2018, Chen et al., 2019) combine the Pointer Network(Vinyals et al., 2018) with the memory network and improve the dialog modeling performance by utilizing external knowledge.

A good dataset is essential for learning modeling dialogues. To create and evaluate a high-performance dialogue model, many researchers have created and published high-quality multidomain dialog datasets.

(Eric et al., 2017) published a three-domain task-oriented dialog dataset: calendar scheduling, weather information retrieval, and point-of-interest navigation. (Budzianowski et al., 2018) proposed a multi-domain dialog dataset named MultiWOZ, which DSTC8 Track1 uses. It consists of seven domains and has a total of 8,438 dialogs, which can be useful for studying multi-domain dialog modeling compared to existing multi-domain task-oriented dialog datasets.

(Zhao et al., 2019) performed reinforcement-learning-based dialogue modeling on the MultiWOZ dataset and shared the results. The author points out the limitations of previous word-level reinforcement learning and proposes a new reinforcement learning technique called Latent Action Reinforcement Learning(LaRL).

Proposed Method

The core of our proposed method is the restriction of system utterance with domain and dialogue act information. We constrain the system utterance generation words to alleviate the difficulty of modeling the system and achieve high performance. We assumed that the previous user utterance and the domain information of the dialog play an important role in determining the utterance of the system. We apply constraints by masking the system utterance generated words. This method borrows the DropMax proposed in (Lee et al., 2019), and DropMax masks an arbitrary category each time for classification problems, and this has been reported to produce an ensemble classifier.

The data set provided by DSTC8 has domain information per conversation, but no dialogue act. Thus, to implement the proposed method, we attached dialogue acts and domain information in utterance units.

We used Mem2Seq(Madotto et al., 2018) as the base model for the experiment. Mem2Seq is a proposed structure for task-oriented dialogue modeling. It follows the encoder-decoder framework, which takes as input all the dialog history performed so far and generates system utterances based on the results of the encoder. Mem2Seq processes the hidden representation of the utterance generator by searching the entire dialog history to generate each word in the system utterance. This method can achieve the same effect as the attention mechanism of the existing sequence to sequence architecture. Also, it uses two generation methods with priority for utterance generation. The two generation methods are pointer generator and vocab generator. If the pointer generator points to one of the words in the input dialog history, the system uses it as the next token; otherwise, the system infers the next token from the entire vocabulary.

Mem2Seq consists of one dialog history encoder and one decoder to perform multi-hop inference. We added two modules to this infrastructure. The two modules added are the dialogue act estimation module and the domain estimation module for user utterance. Each module of the proposed method is as follows.

Dialog History Encoder: We construct the memory with a total of $k+1$ learnable embedding matrices $C = \{ C^1, \ldots, C^{k+1} \}$. We perform the following operation with q^0 set to zero vector for every input.

$$p_i{}^k = \text{Softmax}\big((q_i{}^k)^T C_i{}^k\big), \quad (1)$$

$$o^k = \textstyle\sum_i p_i^k C_i^{k+1} \quad (2)$$

$$q^{k+1} = q^k + o^k \quad (3)$$

In Equations 1, 2, and 3, q is a query that pays attention to memory, p is the result of attention operation, $p \in R^l$, and l is the number of words in the dialog history. We first pay attention to the embedding matrix C^k of each hop (Equation 1) and then use the result p to read the next layer of memory (Equation 2). Finally, we update the query q with the result o^k. And we repeat this operation by the number of hops we specify. In other words, we pay attention to the entire dialog history by a fixed number of hops.

Domain Inference Module: We use this to do the same thing as the dialog encoding module. The dialog history encoder uses the entire dialog history, but the domain estimator uses only the last utterance.

Dialogue Act Inference Module: The module that predicts the dialogue act uses the same method as the domain estimator. Therefore, only the last utterance is used as input.

Decoder: We use the result vector $\mathrm{o}^{E,K}$ of the dialog encoder that encodes the entire dialog history as the initial state vector of the utterance generation module. We use $\mathrm{o}^{\mathrm{D,K}}$ and $\mathrm{o}^{\mathrm{S,K}}$ as masks that restrict the distribution of words through equations 4 and 5.

$$m_s = \sigma(o^{D,K} U_1 + o^{S,K} U_2) \quad (4)$$

$$m = \text{Ber}(m_s) * m_s \quad (5)$$

In Equation 4 and 5, $\mathrm{U}_1, \mathrm{U}_2 \in \mathbb{R}^{\mathrm{h,v}}$, h is the hidden dimension of the structure, and v is the number of words in the utterance generation module. Ber is a sample of the Bernoulli distribution. In Equation 4, two estimation results are summed and passed through a sigmoid function to obtain a sigmoid mask. As shown in Equation 5, we generate the final mask by multiplying the sample of the Bernoulli distribution by the sigmoid mask.

The above mask may also mask every step correct answer token. Therefore, when training, the correct answer token is set to 1 in m, which is an operation result, so that the correct answer is not masked. Since the mask cannot be adjusted using the correct answer during the evaluation, we change the equation of the mask so that only the sigmoid mask works. As mentioned earlier, we estimate the dialogue act and domain of the user's last utterance and constrain the generated words of the utterance generation module. Through this process, the system can infer words based on domains and dialogue acts and generate utterances in consideration of domain characteristics.

Experiments

Experimental Settings

We used the MultiWOZ dataset provided by the DSTC8 organizer for the experiment. We attached a dialogue act and domain to each user's utterance in the dataset to supervise the proposed method DA and domain. Statistics of attached dialogue acts and domains are shown in table 1.

Dialogue Act	Count	Ratio (%)
none	3,072	4.30
request	49,375	69.06
request_book	695	0.97
ask_property	9,640	13.49
accept	535	0.75
greeting	8,170	11.43

Domain	Count	Ratio (%)
Restaurant	21190	29.64
Hotel	19113	26.74
Train	15090	21.11
Attraction	11558	16.17
Taxi	2643	3.70
Police	744	1.04
General	195	0.27
Hospital	954	1.33

Table 1 Statistics of dialogue acts and domains in MultiWOZ dataset

Experiments	BLEU1	BLEU2	BLEU3	BLEU4	Corpus BLEU
Baseline	0.235 (± 0.020)	0.084 (± 0.014)	0.049 (± 0.011)	0.030 (± 0.008)	8.64 (± 1.52)
+Domain Estimator	0.244 (± 0.008)	0.090 (± 0.008)	0.053 (± 0.006)	0.032 (± 0.005)	9.28 (± 1.10)
+Dialogue Act Estimator	0.249 (± 0.016)	0.100 (± 0.009)	0.059 (± 0.005)	0.036 (± 0.003)	10.03 (± 0.80)
+Domain Estimator + Dialogue Act Estimator (Proposed Method)	**0.267** (± 0.017)	**0.107** (± 0.012)	**0.064** (± 0.008)	**0.039** (± 0.006)	**11.01** (± 1.21)

Table 2 Multi-domain dialogue modeling evaluation on MultiWOZ dataset

We compared the performance of the proposed method with the Mem2Seq baseline to verify the effectiveness of the proposed method. In our experiments, we used the optimizer as RAdam(Liu et al., 2019) and fixed the learning rate to 0.001. The batch size was set to the highest value that could be set during the experiment. We experimented by fixing the number of hops and the hidden dimensions to the values that achieved the highest performance in the Mem2Seq baseline experiment. In the proposed method, two newly added modules have the same hop number and hidden dimension as Mem2Seq's encoder.

Experimental Results & Analysis

Table 2 shows the performance of the baseline and the proposed method. The proposed method could improve the performance of the baseline model in all indicators.

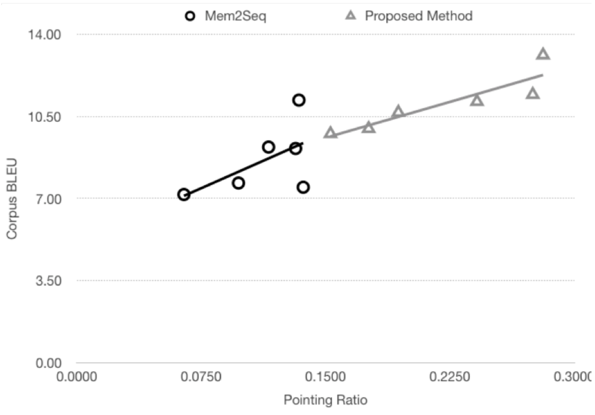

Figure 2 Performance trend according to the ratio of the pointer distribution

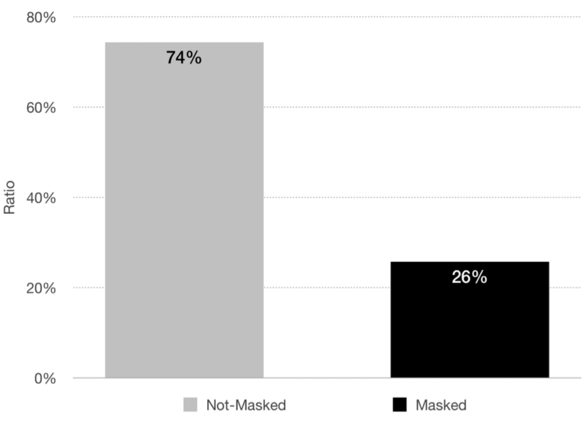

Figure 3 Total ratio of masked word

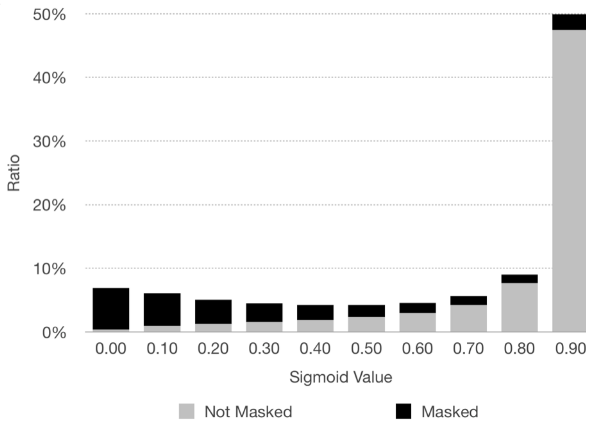

Figure 4 Mask distribution over value of the sigmoid mask

Figure 2 illustrates a figure to compare the performance of the pointer generator and the vocabulary generator. As shown in Figure 2, if many words were selected in the pointer distribution, the performance was better than when selected in the vocabulary distribution. We could analyze that the pointer distribution had a positive effect on dialogue generation. Using these traits, we can improve performance when we generate dialogue using additional sources other than conversation history.

Figure 3 and 4 show the rate at which the proposed method constrains words. Figure 3 shows the ratio of masked words and not-masked words for the entire sample. The proposed method masked approximately 26% of unnecessary words. In Figure 4, the x-axis is the value of the sigmoid mask, and the y-axis is the proportion of the samples with that value. In the figure, there are many large sigmoid mask values. The larger this value, the more words that are not masked by the Bernoulli distribution. As a result, there are a lot of gray areas. We found that this figure affected performance. Therefore, more research is needed to increase the number to increase learning speed and performance.

Experiments	Baseline	Proposed Method
Restaurant	0.6905	**0.7264** (+0.0359)
Attraction	0.6429	**0.7579** (+0.1150)
Taxi	0.5600	**0.8632** (+0.3032)
Train	0.9921	**0.9932** (+0.0011)
Hotel	0.3077	**0.3909** (+0.0832)
Police	0.0000	0.0000
Hospital	0.0000	0.0000

Table 3 Entity cohesion evaluation on MultiWOZ dataset

The last analysis we performed is the domain cohesion analysis. We wanted to improve system utterance using domain and dialogue act information. Thus, if the approach worked, the system utterance of the proposed method would have high domain cohesion. Table 3 shows the analysis result. The analysis was performed using the frequency of the appearance of entities in system utterances.

$$c_d = \frac{\sum_j f(w_{dj})}{\sum_i f(w_{di})}, \text{where } w_i, w_j \in U_d, w_i \in E_{all}, w_j \in E_d \quad (6)$$

The closer this number is to 1, the more entities in the domain are uttered, resulting in higher domain cohesion. In the evaluation, the proposed method was found to improve the performance in five domains including the restaurant. In Table 3, the cohesion value of police and hospital is 0, because only five conversations showed the police domain, and the hospital did not occur even once.

Conclusion

We use domain inference information and dialogue act inference information to constrain words that system utterance can generate to perform multidomain dialogs. We evaluated how the operation helped to generate domain utterances. The proposed method can improve the not only utterance generation performance but also the domain cohesion of system utterance.

We used domain and dialogue act information for word constraints. Therefore, domain and dialogue act information about utterance is needed. After that, we will research how to perform such masking without labeling work.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT). (No. 2018-0-00247, Development of Data Self-proliferation and Validation Technology for Deep Learning Using GANs)

References

Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bordes, A., Boureau, Y.L. and Weston, J., 2016. Learning end-to-end goal-oriented dialog. arXiv preprint arXiv:1605.07683.

Budzianowski, P., Wen, T. H., Tseng, B. H., Casanueva, I., Ultes, S., Ramadan, O., & Gašić, M., 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. arXiv preprint arXiv:1810.00278.

Chen, X., Xu, J. and Xu, B., 2019, July. A Working Memory Model for Task-oriented Dialog Response Generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2687-2693).

Eric, M. and Manning, C.D., 2017. Key-value retrieval networks for task-oriented dialogue. arXiv preprint arXiv:1705.05414.

Lee, H.B., Lee, J., Kim, S., Yang, E. and Hwang, S.J., 2018. DropMax: Adaptive variational softmax. In Advances in Neural Information Processing Systems (pp. 919-929).

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J. and Han, J., 2019. On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265.

Madotto, A., Wu, C.S. and Fung, P., 2018. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. arXiv preprint arXiv:1804.08217.

Sukhbaatar, S., Weston, J. and Fergus, R., 2015. End-to-end memory networks. In Advances in neural information processing systems (pp. 2440-2448).

Sakai, A., Shi, H., Ushio, T. and Endo, M., 2017. End-to-end memory networks with word abstraction and contextual numbering for goal-oriented tasks. Dial. Syst. Technol. Challenges, 6.

Sutskever, I., Vinyals, O. and Le, Q.V., 2014. Sequence to sequence learning with neural networks. Advances in NIPS.

Vinyals, O., Fortunato, M. and Jaitly, N., 2015. Pointer networks. In Advances in Neural Information Processing Systems (pp. 2692-2700).

Vinyals, O. and Le, Q., 2015. A neural conversational model. arXiv preprint arXiv:1506.05869.

Wen, T.H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L.M., Su, P.H., Ultes, S. and Young, S., 2016. A network-based end-to-end trainable task-oriented dialogue system. arXiv preprint arXiv:1604.04562.

Williams, J.D., Asadi, K. and Zweig, G., 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. arXiv preprint arXiv:1702.03274.

Zhao, T., Xie, K. and Eskenazi, M., 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. arXiv preprint arXiv:1902.08858.