Prompt-based Dialogue State Tracking Method Jointly Modeled with Natural Language Understanding

Anonymous ACL submission

Abstract

Cross-domain dialogue state tracking has become a hot topic in recent years, it profoundly influences the generalizability of task-oriented dialogue systems. In this paper, we propose a prompt-based dialogue state tracking method jointly modeled with natural language understanding (PLDT) to address the problem of multi-domain adaptation in the state tracking task and optimize the existing models. We introduce the joint modeling method to reduce the cumulative errors between DST and NLU in pipeline dialogue system. Based on this, in analyzing current dialogue state tracking methods, we combine T5 with Ptr-Net in a proper way to solve both the redundancy and inaccuracy shortcomings in generative methods and the out-of-vocabulary (OOV) problem in pointer network methods, respectively. We also design a continuous prompt learning approach that uses a few discrete samples (labeled by a keyword extraction algorithm in an automatic way) to train the model in an unsupervised way and generate a suitable prompt. Our model outperforms other existing approaches on MultiWOZ2.0 and CrossWOZ in both slot and joint accuracy and has better performance in zero-shot tasks than other cross-domain models.

1 Introduction

Dialogue state tracking (DST), as a critical component in the pipeline approach of task-oriented dialogue systems, profoundly affects the agent's performance. It takes input from natural language understanding (NLU) and outputs the current turn's intents and slot-value pairs. Table 1 provides a typical example of DST. Due to the task similarity of DST and NLU, recent research usually models them jointly (Zhang et al., 2020; Chen et al., 2017), where the joint model receives the user's utterances directly and effectively solves the error accumulation. In addition, unseen slots tracking task which is belong to a zero-shot domain adaptation problem (Peng et al., 2018) has become a popular issue with the development of cross-domain task-oriented dialogue systems (Huang et al., 2020).

In terms of cross-domain research, there are three classical approaches. The first approach makes the model independent of the ontology/belief states and predicts the value by calculating semantic similarity between the dialogue context and ontology terms. Therefore, the model can address the cross-domain problem by training on different domain data (Ramadan et al., 2018; Lee et al., 2019). Obviously, a new domain requires training from scratch, which can lead to lower generalizability. The second approach extracts the dialogue states directly from user utterances, using copy mechanisms (See et al., 2017; Xu and Hu, 2018; Gao et al., 2019), This method can capture information well from the context, but it fails when slot values do not appear in the dialogue. The last approach regards DST as a generation task that can extend the ontology to the entire vocabulary (Le et al., 2020). However, it generates semantically similar values repeatedly and makes the dialogue state redundant.

In this work, we proposed the Prompt-based Dialogue State Tracking jointly modeled with Natural Language Understanding (PLDT) method to tackle these challenge. We combined the generative model with the extractive model, which not only solves the excessive dependence of terms extraction on user utterance, but also avoids the problem of repeated generation. Then we designed a prompt learning to fine-tuning the pretraining model for the zero-shot domain adaptation scenario. The main contributions are as follows:

1. We combine the Seq2Seq structure with the Ptr-Net, solve the OOV(out-of-vocabulary) problem and make slot values more accurate.

2. We designe a continuous prompt learning method that used a keyword extraction algorithm to generate few discrete training data, and train

Turn	Actor	Input	Dialogue			
			Name	Ticket	Play-time	Score
1	User	Hello, I'm looking for a scenic spot with a rating of 4.5 or above. Do you have any good places to recommend?	none	none	none	More than 4.5 points
2	Agent	There are so many good places. You can go to the Forbidden City, Badaling Great Wall, the Summer Palace and so on.	Badaling Great Wall	none	none	More than 4.5 points
2	User	I want to go to Badaling Great Wall. Where is the address? How long can I play?	Badaling Great Wall	none	3-4 hours	More than 4.5 points
3	Agent	Take a right at Exit 58, Beijing-Tibet Expressway, Yanqing District, Beijing; You can play for 3-4 hours.	Badaling Great Wall	none	3-4 hours	More than 4.5 points
3	User	Thanks! No more questions, bye!	Badaling Great Wall	none	3-4 hours	More than 4.5 points
4	Agent	You're welcome! Wish you a happy life! Bye!	Badaling Great Wall	none	3-4 hours	More than 4.5 points

Table 1: Example of dialogue state tracking.

the generative model in an unsupervised manner, thereby improving the model's generalization and extensibility.

3. The experimental result shows that our method outperforms existing cross-domain DST models. We also analyzed the influence of each component on the model's performance to prove the validity of our method.

2 Related Work

2.1 Prompt Learning

GPT3 (Brown et al., 2020) puts forward a new paradigm of pretraining model based on Prompt learning, that is, add prompt to the input of the pretraining model to make the target of the downstream task more close to the target of the pretraining task, so as to improve the model's performance on the downstream task. In recent years, with the launch of various large models, prompt training has gradually become more and more prominent (Han et al., 2021). Unlike earlier hand-designed prompts, Shin et al. (2020) generated prompts by the model automatically, but this discrete prompt approach lacks flexibility. The works like Li and Liang (2021) and Lester et al. (2021) called the continuous prompt, they parameterize the prompt as a token to enhance the expressive ability of the prompt. In this work, we choose the continuous prompt learning to fine-tune our model.

2.2 Dialogue State Tracking (DST)

There has been a lot of research on cross-domain DST task in recent years. Zhong et al. (2018) uses semantic similarity matching to predict the dialogue state and Lee et al. (2019) regards the field slot pair as the question, the slot value pair as the answer, and finally uses the classifier to select the dialogue state with the highest probability. They are all limited to the ontology. Heck et al. (2020); Xu and Hu (2018); Wu et al. (2019) introduced a pointer network to avoid experts manually designing the ontology, and Wu et al. (2019) combined the pointer network with RNN, fixed an issue where slot terms could not be found directly in dialog statements. Ren et al. (2019); Lin et al. (2021); Kim et al. (2019), and Zeng and Nie (2020) chose a generative way. Lin et al. (2021) uses the T5 pre-training model as the encoder and decoder to directly generate the dialog state between the system's and the user's utterance, Zeng and Nie (2020) uses BERT as encoder and decoder at the same time, and uses the attention-mask matrix to control BERT for state prediction and slot value generation, which implements the flat modeling of encoder and decoder, improves the efficiency of the model, and solves the problem that the model using hierarchical decoder cannot be jointly optimized. Those generation models can get rid of the ontology limitation but the generated slot-value is often not precise enough.

3 Proposed Method

3.1 Model Structure

In order to have a better performance on cross-domain DST tasks, we propose the prompt based dialogue state tracking jointly modeled with natural language understanding (PLDT) method. Figure 1 shows the overall framework of PLDT. We input user's and agent's utterances history and use Posi-

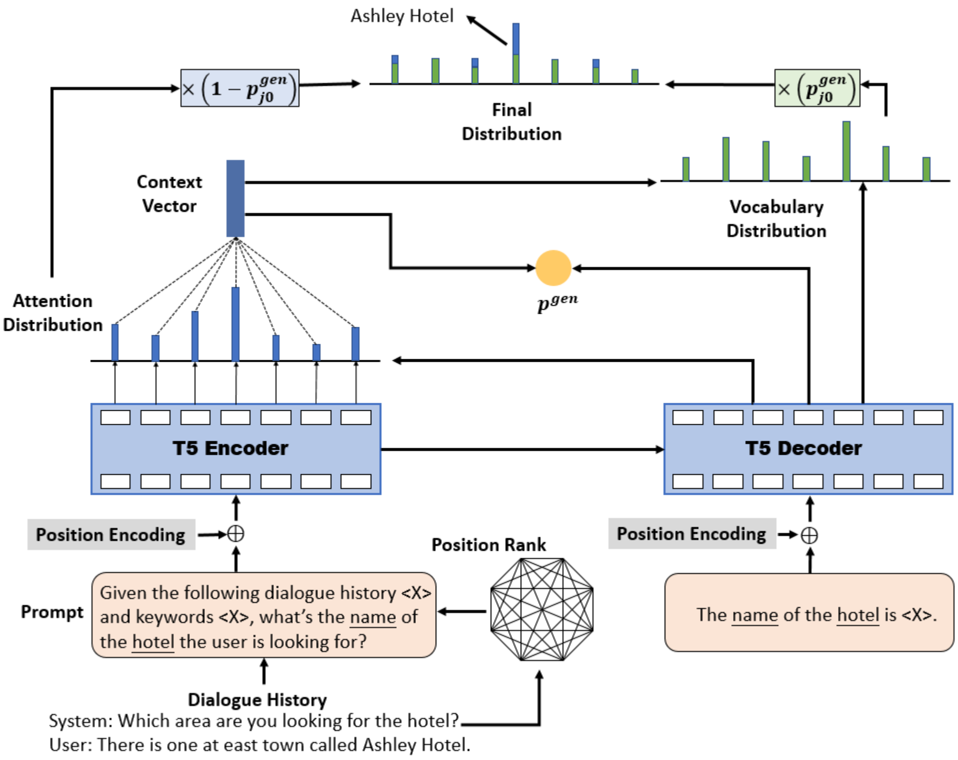

Figure 1: Architecture of PLDT framework

tion Rank (Florescu and Caragea, 2017) algorithm to extract the keywords from utterances, then concatenate utterances, keywords, and prompt texts (or vectors) as the input of the T5 (Raffel et al., 2020) pre-training model encoder. By integrating the hidden layer of the T5 encoder and the decoder with an attention, the attention weight distribution and context vector for the input sequence are obtained. Then, on the one hand, we use the context vector and decoder result to calculate the probability distribution on the vocabulary; on the other hand, we concatenate the context vector, the input of the decoder and the hidden layer of the decoder to obtain the generated pointer P_{gen}. Finally, we use P_{gen} to weight the attention distribution of the input sequence and the probability distribution on the vocabulary to get the final text probability distribution.

3.2 Continuous Prompt Learning

We designed a continuous prompt generating method to deal with prompt generation in multi-domain data sets. As Figure 2 shows, we use the keyword extraction algorithm on dialogue history and then initialize the discrete prompt text into a vector representation, then input the prompt word vector into the pre-training model. Through the automatic learning of the pre-training model, a continuous prompt word vector is obtained.

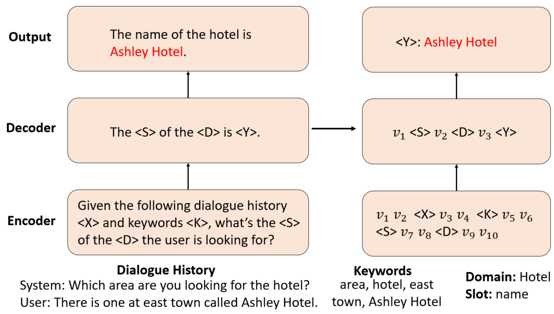

Figure 2: A continuous prompt learning examples

We represent each prompt text as $t = \{t_1, t_2, , t_n\}$. And then encode the prompt text through the Tokenizer of the T5 to obtain vectors $V = \{V_1, V_2, \ldots, V_n\}$ on R^d. Meanwhile, we encode the multi-round dialogue X and the keywords K by T5 Tokenizer, and concatenate prompt vectors to get the synthetic input sequence V'.

$$V' = [V, X, K] \tag{1}$$

Then we fix the parameters of the T5 model, and input V' into the T5 for fine-tuning. The set prediction distribution is defined as:

$$p(y|x) = \sum_{\mathbf{t} \in T} PLM(y|\mathbf{t}.x) \tag{2}$$

Domain	Slot	Train set	Dev set	Test set
Attraction	area, name, type	2717	401	395
Hotel	area, day, internet, name, parking, people, price, stars, stay, type	3381	416	394
Restaurant	area, day, food, name, people, price, time	3813	438	437
Taxi	arrive by, departure, destination, leave at	1654	207	195
Train	arrive by, day, departure, destination, leave at, people	3103	484	494

Table 2: Slot statistics on the MultiWOZ dataset.

Domain	Slot	Train set	Dev set	Test set
Attraction	name, rating, fee, duration, address, phone, nearby attract, nearby rest, nearby hotels	4154	421	413
Hotel	name, rating, price, type, services, phone, address, nearby attract, nearby rest	4156	410	409
Restaurant	name, rating, cost, dishes, address, phone, open, nearby attract, nearby rest, nearby hotels	4200	429	427
Taxi	from, to, car type, plate number	688	78	73
Metro	from, to, car type, plate number	669	62	82

Table 3: Slot statistics on the CrossWOZ dataset.

Where $p(y|x)$ obeys distribution on prompt vector T. By maximizing the set prediction distribution, the loss is:

$$L_{prompt} = \sum_{(x,y)\in\epsilon} -log \sum_{\mathbf{t}\in T} p(y|\mathbf{t}.x) \quad (3)$$

Where ϵ is state tracking task on a training set.

3.3 T5 Model

After Prompt learning convergence, the Prompt word vector has been fixed and the input sequence V' will input into T5 for further fine-tune the pointer generation network.

Then, we use attention mechanism to calculate the attention weight distribution of each word α^t:

$$e_i^t = v^T tanh(W_h h_i + W_s s_t + b_{attn}) \quad (4)$$

$$\alpha_t = softmax(e_t) \quad (5)$$

Where h_i is the encoder hidden layer in V', s_t is the decoder hidden layer in time t. The attention context vector c_t is calculated from:

$$c_t = \sum_i \alpha_i^t h_i \quad (6)$$

c_t contains the contextual semantic information of all the words in the input sequence, which will be used in the pointer generation network to assist the decoder's output at time t.

3.4 Pointer Generation Network

We concatenate the state s_t and the c_t, and put them into two fully connected layers and Softmax activation function, to obtain the probability distribution pvocab on the word list:

$$P_{vocab} = softmax(W'(W[s_t, c_t] + b) + b') \quad (7)$$

Then, in order to combine the attention weight distribution with the prompt distribution, we calculate the generated pointer P_{gen} for controlling the generated word:

$$P_{gen} = \sigma(w_c^T c_t + w_s^T s_t + w_x^T x_t + b_{ptr}) \quad (8)$$

Where x_t is the input of decoder hidden layer. The range of P_{gen} is $[0, 1]$, The larger the P_{gen} is, the more candidates from generated words. The smaller the P_{gen} is, the more candidates from the original input sequence. The final generation probability distribution of the word ω is:

$$P(\omega) = P_{gen} P_{vocab}(\omega) + (1 - P_{gen}) \sum_{i:\omega_i=\omega} \alpha_i^t \quad (9)$$

For the entire output sequence Y the loss is:

$$\mathcal{L} = \frac{1}{Y} \sum_{t=0}^{Y} -logP(\omega_t^*) \quad (10)$$

Model	MultiWOZ		CrossWOZ	
	Slot	Joint	Slot	Joint
GLAD*	95.44	35.57	86.02	25.84
SUMBT**	95.72	42.40	86.92	30.49
SpanPtr+	93.85	30.28	85.98	25.96
COMERE**	96.37	48.79	88.43	41.50
TRADE+*	96.92	48.62	90.71	36.08
SOM-DST**	97.54	51.72	91.03	36.55
Transformer-DST***	97.69	54.64	93.45	39.36
SST'	96.85	51.17	\	\
OPAL'	97.24	54.10	\	\
PLDT(Transformer)/ Transformer	96.02/85.16	51.54/30.53	93.70/78.44	36.06/20.75
PLDT(UNILM) / UNLLM	97.14/96.02	55.23/49.52	94.85/89.89	39.77/35.59
PLDT(BART) / BART	97.80/96.50	57.24/50.65	**95.51**/90.64	41.86/36.02
PLDT(T5) / T5	**98.11**/97.17	**57.83**/52.70	95.46/91.74	**42.14**/38.20

Table 4: cross-domain DST model performance comparison on MultiWOZ and CrossWOZ. * represents using LSTM as encoder, ** represents using BERT as encoder, *** represents using BERT as encoder and decoder, + represents use pointer network, ' represents End-to-End model. The lower part shows the experimental results of PLDT combined with the seq2seq model, and the right part of the slash represents the results of DST task using only the seq2seq model.

4 Experimental

4.1 Dataset

To verify the validity of our proposed model, we used two different open source data sets: MultiWOZ2.0(Ramadan et al., 2018) and CrossWOZ(Zhu et al., 2020). They are suitable for English and Chinese dialogue state tracking domain tasks respectively.

(1) **MultiWOZ2.0**

MultiWOZ2.0 [1] is a multi-domain English dialogue data set that contains real conversations between visitors and staff of the Visitor Center in multiple domains. There are 3406 single-domain conversations and 7032 multi-domain conversations, and 8438 multi-round conversations with an average of 8.93 single-domain conversations. The average number of rounds of multi-field dialogues was 15.39.

Because the data of Hospital and Police are very small, we only experiment on the other five, the statistical information is shown in Table 2:

(2) **CrossWOZ**

CrossWOZ [2] is a multi-domain Chinese conversation data set, which contains multi-rounds of task-based conversation data in five fields: restaurants, scenic spots, hotels, taxis and subways. There are altogether 6012 conversations with an average number of 16.9. Statistical analysis was performed on the CrossWOZ dataset are shown in Table 3.

4.2 Baseline

We compare our model with other state-of-the-art methods on MultiWOZ2.0 and CrossWOZ. Ontology based method: GLAD(Zhong et al., 2018), SUMBT(Lee et al., 2019); pointer network method: SpanPtr(Xu and Hu, 2018), TRADE(Wu et al., 2019); generation method: COMER(Ren et al., 2019), T5DST(Lin et al., 2021), SOM-DST(Kim et al., 2019), and Transformer-DST(Zeng and Nie, 2020), the end-to-end model SST(Chen et al., 2020) and OPAL(Chen et al., 2022). We also used other seq2seq models to compare with the T5 model: BiLSTM, Transformer(Vaswani et al., 2017), UNILM(Dong et al., 2019), BART(Lewis et al., 2019).

4.3 Evaluation Measures

We use the follow metrics to evaluate the model's performance. Slot Accuracy: percentage of domain-slot-value are correctly predicted.

$$P_{slot} = \frac{N^{+}_{slot}}{N_{slot}} \quad (11)$$

[1] `https://github.com/budzianowski/multiwoz`
[2] `https://github.com/thu-coai/CrossWOZ.`

Joint Accuracy: percentage of the turns in current dialogue whose slots are all correctly predicted.

$$P_{joint} = \frac{N_{turn}^{+}}{N_{turn}} \tag{12}$$

4.4 Training Setting

We use Large version(Xue et al., 2020) of the T5 model, witch hidden layer size is 1024, the total parameters of the model is 780M, and choose the Adam optimizer. The initial learning rate is 0.0001, batch size is 16, and the default epoch is 50. In the training process, we adopt the early stop strategy to evaluate the performance of the model on the validation set every other round. When the performance on the validation set did not improve for three consecutive epochs, the training was stopped.

4.5 Experimental Results

Table 4 shows the cross-domain DST task result. Comparing the SUMBT, COMER and GLAD models, we can see that BERT based encoder model improved significantly in each performance than the LSTM based encoder model, especially in joint accuracy. By comparing the SpanPtr, TRADE and COMER, we can see that the SpanPtr that only uses pointer network to extract slot values from dialogue utterances has poor performance. While the TRADE combining Seq2Seq with pointer network has achieved a good result, and its slot accuracy is even better than that of COMER which use BERT as an encoder. By comparing SOM-DST, COMER and Transformer-DST, when all encoders use BERT, the decoder that also uses BERT performances better than that use RNN structure. It shows again that the pre-training model can bring stronger semantic modeling ability. Meanwhile, we compare with two end-to-end SOAT methods and only get the experimental results of MultiWOZ from the paper for comparison due to the lack of source code. Our model achieved optimal results on all indexes of both data sets,indicating that the combination of Prompt learning and T5 pre-training model with pointer generation network can further improve the context comprehension and semantic modeling ability of the model. Additionally, We replace T5 with other seq2seq models, and the results show that the use of PLDT method has a great improvement on the DST task of the seq2seq model.

Table 5 shows the zero-shot prediction performance of our method in the four fields is better than other models, which reflects strong generalization ability and domain scalability. TRADE and COMER use randomly initialized RNN decoders and behave generally in this task. T5DST uses T5 as an encoder and decoder has strong language understanding, but the result is slightly less than pointer generation networks. As shown in Figure 3, we also find that it is more difficult to identify slots in specific field, while it is relatively less difficult to identify slots overlapping in different fields.

4.6 Ablation Study

In order to verify the components of our model, we conducted ablation experiments.

Effect of prompt

To prove the improvement of the prompt, we compared the differences between no prompt, discrete prompt, and continuous prompt learning. As shown in Table 6, using prompt is better than not using, continuous prompt learning is better than discrete prompt, because the discrete prompt is designed manually which makes it difficult to ensure the quality of each prompt. The continuous Prompt learning method can automatically learn the locally optimal prompt to make the model easily understand the conversation, thus improving the model's performance. On the other hand, it also shows that a monotonous prompt for DST is not enough, and diverse prompts are needed to improve the accuracy of the model.

Model	MultiWOZ		CrossWOZ	
	Slot	Joint	Slot	Joint
without Prompt	97.62	56.32	94.13	41.08
Discrete Prompt	97.88	56.90	95.06	41.65
Continuous Prompt Learning	**98.11**	**57.83**	**95.46**	**42.14**

Table 6: Prompt ablation result on MultiWOZ and CrossWOZ.

Effect of keyword enhancement

In continuous prompt learning we use the keyword extraction to enhance the performance of prompt generation. We compared position rank with other keyword extraction algorithms: YAKE(Campos et al., 2018), TF-IDF(Sammut and Webb, 2011), and TextRank(Mihalcea and Tarau, 2004). Table 7 illustrates the enhancement effect,

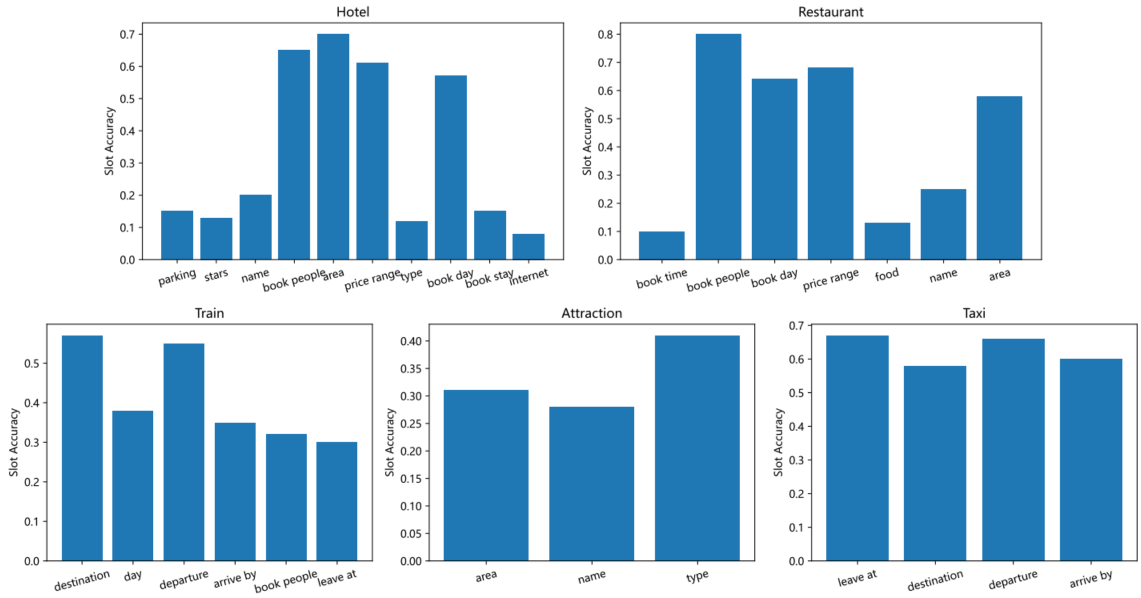

Figure 3: Slot accuracy statistics on five domains of the MultiWOZ dataset.

Model	**joint/MultiWOZ**					
	Attraction	Hotel	Restaurant	Taxi	Train	Average
TRADE	19.87	13.70	11.52	60.58	22.37	25.76
SUMBT	22.60	19.80	16.50	59.50	22.50	28.18
T5DST	32.66	18.73	20.55	**64.62**	31.27	33.56
PLDT	**35.91**	**22.36**	**23.44**	62.57	**35.12**	**35.88**

Table 5: Zero-shot performance in five domains on MultiWOZ.

and Table 9 directly shows that other algorithms have problems such as keyword repetition and inaccuracy.

Model	**MultiWOZ**		**CrossWOZ**	
	Slot	Joint	Slot	Joint
without keyword	97.89	57.25	95.22	41.71
YAKE	97.72	57.14	95.15	41.78
TF-IDF	97.52	56.96	95.30	41.58
TextRank	97.95	57.44	95.20	41.93
PositionRank	**98.11**	**57.83**	**95.46**	**42.14**

Table 7: Different keyword extraction algorithm comparation.

Effect of pointer network

Finally, we in investigate the effectiveness of pointer network in Table 11, the results show that the method of using pointer generation network is better than removing pointer generation network. That is to say, only rely on decoder to generate answers, some words will be generated repeatedly.

Model	**MultiWOZ**		**CrossWOZ**	
	Slot	Joint	Slot	Joint
without pointer network	97.43	53.10	92.78	38.77
PLDT	**98.11**	**57.83**	**95.46**	**42.14**

Table 8: Pointer generates network ablation results.

5 Conclusion

In this paper, we propose a prompt-based dialogue state tracking method jointly modeled with natural language understanding (PLDT). The method combines the advantages of generative models and pointer networks, and uses T5 as the seq2seq model for the pointer generation network. We then design a prompt learning method that uses unsupervised training to generate a continuous prompt. Furthermore, we introduce the position rank algorithm to avoid manual prompt design and reduce labeling costs. We verify the outstanding performance and generalization of our model on benchmark datasets MultiWOZ2.0 and CrossWOZ by comparing it with existing state-of-the-art DST methods and analyze the validity of each component at the end.

Limitations

In this section, we'll discuss the limitation of our PLDT model. First of all, a generative structure could inevitably result in a large number of network parameters, which would undoubtedly increase the training cost of the model, although we used prompt to fine-tune the LLM, but it still took a lot of time. Furthermore, our experiment is only trained on English and Chinese datasets, so there is no in-depth discussion of whether the model has generalization on other different languages and what's the factors that affect DST tasks in different languages.

Acknowledgements

This work is funded by **********(anonymous)*********

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. A text feature based automatic keyword extraction method for single documents. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pages 684–691. Springer.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7521–7528.

Zhi Chen, Yuncong Liu, Lu Chen, Su Zhu, Mengyue Wu, and Kai Yu. 2022. Opal: Ontology-aware pretrained language model for end-to-end task-oriented dialogue. *arXiv preprint arXiv:2209.04595*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.

Corina Florescu and Cornelia Caragea. 2017. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialog state tracking: A neural reading comprehension approach. *arXiv preprint arXiv:1908.01946*.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašić. 2020. Trippy: A triple copy strategy for value independent neural dialog state tracking. *arXiv preprint arXiv:2005.02877*.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

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 CH Hoi. 2020. Non-autoregressive dialog state tracking. *arXiv preprint arXiv:2002.08024*.

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. Sumbt: Slot-utterance matching for universal and scalable belief tracking. *arXiv preprint arXiv:1907.07421*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Zhaojiang Lin, Bing Liu, Seungwhan Moon, Paul Crook, Zhenpeng Zhou, Zhiguang Wang, Zhou Yu, Andrea Madotto, Eunjoon Cho, and Rajen Subba. 2021. Leveraging slot descriptions for zero-shot cross-domain dialogue state tracking. *arXiv preprint arXiv:2105.04222*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Kuan-Chuan Peng, Ziyan Wu, and Jan Ernst. 2018. Zero-shot deep domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 764–781.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Osman Ramadan, Paweł Budzianowski, and Milica Gašić. 2018. Large-scale multi-domain belief tracking with knowledge sharing. *arXiv preprint arXiv:1807.06517*.

Liliang Ren, Jianmo Ni, and Julian McAuley. 2019. Scalable and accurate dialogue state tracking via hierarchical sequence generation. *arXiv preprint arXiv:1909.00754*.

Claude Sammut and Geoffrey I Webb. 2011. *Encyclopedia of machine learning*. Springer Science & Business Media.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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. *arXiv preprint arXiv:1905.08743*.

Puyang Xu and Qi Hu. 2018. An end-to-end approach for handling unknown slot values in dialogue state tracking. *arXiv preprint arXiv:1805.01555*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yan Zeng and Jian-Yun Nie. 2020. Jointly optimizing state operation prediction and value generation for dialogue state tracking. *arXiv preprint arXiv:2010.14061*.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

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.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.