Exemplar Guided Latent Pre-trained Dialogue Generation

Miaojin Li[1,2], Peng Fu[1] *, Zheng Lin[1] *, Weiping Wang[1], and Wenyu Zang[3]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
[3] China Electronics Corporation
{limiaojin, fupeng, linzheng, wangweiping}@iie.ac.cn
wenyuzang@sina.com

Abstract. Pre-trained models with latent variables have been proved to be an effective method in the diverse dialogue generation. However, the latent variables in current models are finite and uninformative, making the generated responses lack diversity and informativeness. In order to address this problem, we propose an *exemplar guided latent pre-trained dialogue generation* model to sample the latent variables from a continuous sentence embedding space, which can be controlled by the exemplar sentences. The proposed model contains two parts: exemplar seeking and response generation. First, the exemplar seeking builds a sentence graph based on the given dataset and seeks an enlightened exemplar from the graph. Next, the response generation constructs informative latent variables based on the exemplar and generates diverse responses with latent variables. Experiments show that the model can effectively improve the propriety and diversity of responses and achieve state-of-the-art performance.

Keywords: Multiple Responses Generation · Pre-trained Generation · Dialogue Generation.

1 Introduction

Text generation, which is a challenging task due to the limited dataset and complex background knowledge, is one of the most popular branches of machine learning. By pre-trained on large-scale text corpora and finetuning on downstream tasks, self-supervised pre-trained models such as GPT[1], UNILM[2], and ERNIE[3, 4] achieve prominent improvement in text generation. Dialogue generation is one of the text generation tasks, so the pre-trained models also gain remarkable success in dialogue generation[5, 6]. However, these researches treat the dialogue dataset as the general text and ignore the unique linguistic pattern in conversations.

Different from general text generation, dialogue generation requires the model to deal with the one-to-many relationship. That is, different informative replies

* Corresponding Author

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

can answer the same query in human conversation. To better address this relationship, some studies[7–12] focus on improving pre-trained conversation generation with the help of external knowledge. For example, incorporating additional texts [8], multi-modal information[13], and personal characters[7, 11, 12]. They have shown external knowledge can effectively improve response generation. However, the defect of these methods is obvious since they all heavily rely on high-quality knowledge beyond the dialogue text. In addition, the external knowledge is unavailable in a real conversation scenario. To model the one-to-many relationship without external knowledge, inducing latent variables into a pre-trained model is a valid way. With discrete latent variables corresponding to latent speech acts, PLATO[14] and PLATO-2[15] can generate diverse responses and outperform other state-of-the-art methods. However, the discrete latent variables in their work are finite and contain insufficient information, so that the diversity and informativeness of generated responses are limited.

In order to allow latent variables to be more diverse and more informative, we propose the exemplar guided latent pre-trained dialogue generation model. The model treats latent variables as the continuous embeddings of sentences instead of discrete ones. Besides, it has been proved that the usage of exemplar can bring more beneficial and diverse information [16–19] for dialogue generation, so we utilize exemplars to guide the model to construct more informative latent variables. The exemplars contain explicit referable exemplary information from the given dataset. Different from the previous work, which directly searches exemplars with similar semantics to the current conversation, we first constructed a sentence graph and then find the relevant exemplars in the graph. By altering different exemplars, the model can generate more diverse and informative responses.

To achieve our goal, the proposed model is designed with two main parts: exemplar seeking and response generation. In the exemplar seeking, we construct a sentence graph according to the given dataset offline, and then find the proper exemplar in the sentence graph for the current dialogue. Next, in the response generation, we construct the response variable based on the dialogue context and the given exemplar and use a transformer structure to generate a response based on the response variable.

Experimental results show that this approach can achieve state-of-the-art performance. Our contributions can be summarized as follows:

- To generate diverse and informative responses, we propose an exemplar guided latent pre-trained dialogue generation model, in which latent variables are continuous sentence embeddings.
- To obtain referable exemplars, we build a sentence graph and search for an enlightened exemplar in the graph. By altering exemplars and sampling variables, we can get various latent variables to generate diverse and informative responses.
- We conduct our model on three dialogue datasets and achieve state-of-the-art performance.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

2 Related Work

Pre-trained dialogue generation models. DialoGPT[6] exhibits a compelling performance in generating responses. But it pays not enough attention to the one-to-many phenomenon in the dialogue. And the one-to-many relationship can be effectively built with extra knowledge. The researcher[7] uses conditional labels to specify the type of target response. The study[8] provides related information from wiki articles to the pre-trained language model. The work[9] learns different features of arXiv-style and Holmes-style from two separate datasets to fine-tune the pre-trained model to generate responses towards the target style. TOD-BERT[10] incorporates two special tokens for the user and the system to model the corresponding dialogue behavior. The researches [11] and [12] both take the usage of personality attributes to improve the diversity of pre-trained models. However, those methods train models based on the characteristic datasets with specific and additional labels, which is not easy to collect. Without additional information besides dialogue text, PLATO[14, 15] uses discrete latent variables, which are dialogue acts, to improve the diversity of generation. However, the diversity and informativeness of dialogue acts are limited, so their model's improvement is limited.

Hybrid neural conversation models that combine the benefits of both retrieval and generation methods can promote dialogue generation effectiveness. Studies [20] and [21] apply retrieved sentences to assist the generation network in producing more informative responses. Still, retrieved responses may better than generated ones, so studies [22, 23] rerank all responses achieved by these two methods to find the best response. To further improve the quality of generated responses, Researchers [24] and [25] induce the reinforcement learning process. Researcher[25] employs the top N retrieved exemplars to estimate the reward for the generator, and researcher[24] proposes a generative adversarial framework based on reinforcement learning. However, they ignore the phenomenon that irrelevant information in retrieved sentences may mislead the generation. Therefore, researchers[26–28] focus on refining the retrieved sentence into a useful skeleton and utilize this skeleton instead of the whole sentence to guide the generation. Besides, the work[29] applies a two-stage exemplar retrieval model to retrieves exemplar responses in terms of both the text-similarity and the topic proximity. In recent, the transformer structure becomes a popular method in text generation because of its effectiveness. Studies[16–19] conduct this structure to improve the quality of dialogue generation. However, those works ignore the shift phonemes of content in human dialogue[30]. To better capture the direction of this shifting, we form a sentence graph and search for a guiding exemplar in it to direct the generation.

3 Methodology

3.1 Framework

To generate diverse and informative responses, we propose an exemplar guided latent pre-trained dialogue generation model that composed of two main parts: exemplar seeking and response generation.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

The exemplar seeking is designed to get the exemplar sentence. This module uses the dataset to construct the sentence graph at first. In this graph, nodes and edges represent keywords and sentences respectively. Then the module starts from the keyword nodes of the query, wandering along the edge in the graph, to find the end keyword nodes of the response. Finally, it selects the decent sentence to be the exemplar sentence in the edges between start and end.

To generate responses according to the dialogue context and intrusive exemplary, we invent the responses generation module. It first encodes the dialogue context to form a Gaussian distribution and samples a context variable from it. At the same time, this module encodes the exemplar to form another Gaussian distribution and samples an exemplar variable from it. Next, this module calculates the response variable based on the context variable and exemplar variable. Finally, this module generates a response based on the response variable.

3.2 Exemplar Seeking

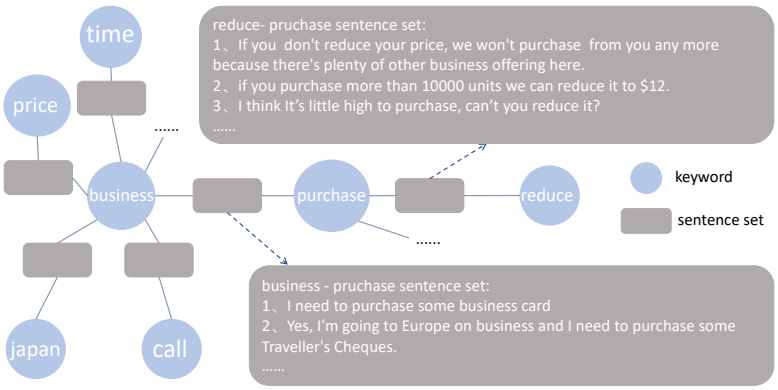

Fig. 1. An Example of Sentence Graph. The blue nodes in the graph represent keywords, while the gray edge between the two nodes represents the set of sentences with these two nodes as keywords.

This module tries to get an enlightened exemplar to guide the response generation stage. Instead of directly retrieving a sentence just by literally-similarity methods in endeavors[16–19], we hope to find an exemplar that can reflect the content shift[30, 31] in conversation, and use this exemplar to guide the generation process. For this purpose, we design a graph-based method to capture the semantic transfer phenomenon, which consists of two parts: graph construction and exemplar selection. The former first constructs a sentence graph based on the given dataset, and the latter selects a suitable exemplar from the sentence graph for current dialogue.

Graph Construction This section purposes to form a sentence graph, from which we can quickly locate a suitable exemplar for the current conversation. To

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

this end, we treat the sentences as the edges and the keywords in the sentence as the nodes. What's more, we define words with the Top-k TF-IDF values in each sentence as the keywords of this sentence. After that, we extract tuples like (keywords1, sentence set, keywords2), where keywords1 and keywords2 are two keywords that exist in the same sentence, and the sentence set is the set of sentences that contains keywords1 and keywords2. As shown in the Fig.1, we next construct the graph from those tuples.

Exemplar Selection After preparing the whole graph, exemplar selection searches for an exact exemplar in the graph. We set keywords in the query as start nodes, and collect keywords from n hops. If the keywords in the response are reachable in n hops, we random choose a sentence from the edge of the last hop as the exemplar sentence. For example, as shown in the Fig.1, the keyword "price" is the keyword in the query, and the "reduce" is the keyword in the response, so we chose the sentence "If you don't reduce your price, we won't purchase from you any more because there's plenty of other business offering here." as the exemplar.

3.3 Responses Generation

After an exemplar is selected, responses generation module attempts to generate the final response based on the dialogue context and the given exemplar. To achieve the goal, we design this module with three parts: an input construction, a latent construction, and a multi-task decoder. At first, The input construction is assigned to construct the input of the latent construction. Second, the latent construction encodes the input into latent variables. Finally, the multi-task decoder completes three generation tasks with these variables.

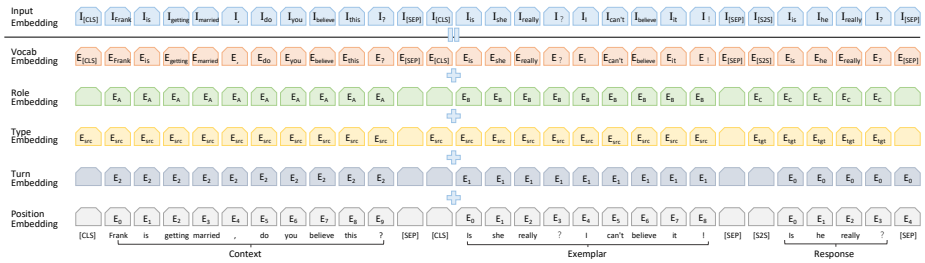

Fig. 2. Input Construction. The input embedding of each token is the embedding sum of the role, turn, position, segment, and vocab embeddings.

Input Construction Input construction aims to construct a comprehensive input of generation stage. To enrich input knowledge, we summarize four kinds of embeddings as the final input embeddings for each token, including the vocab, role, type, turn, and position embeddings. Vocab embedding are intialized with UNILM[2]. Role embedding is managed to distinguish the replier and the interlocutor in a conversation. Type embedding mainly separates dialogue utterances

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

and knowledge information since persona and DSTC dataset contains external knowledge about current dialogue, such as summary and personal profile. Turn embedding is numbered from reply to the beginning of the conversation, the reply is numbered 0, the last statement is numbered 1, and so on. Position embeddings are computed according to the token position in each utterance. As shown in the Fig.2, we concatenate the embeddings of special tags, context, exemplar, and response as the final input series.

Latent Construction The latent construction module works to produce the latent variables based on the input. Since a response is related to the dialogue context and guided by the examplar, the latent response variable can be measured according to the context variable and the examplar variable:

$$Z_{res} = Z_{con} + (Z_{exe} - Z_{con}) * L \tag{1}$$

Where Z_{con}, Z_{exe}, and Z_{res} are latent variables of the dialogue context, the exemplar, and the response separately. L represents the distance of the Z_{res} in the direction from Z_{con} to Z_{exe}. As shown in Fig.3, we use Multi-Layer Transformer[32] as the backbone network and apply a special attention mask for the latent construction. This structure encodes contextual information from both directions and can encode better contextual representations of text than its unidirectional counterpart.

We further induce Gaussian distribution to approximate the ideal distribution of latent variables and use the reparameterization method[33] to get samples of the latent variable. The latent variable of context can be constructed as:

$$Z_{con} = \mu_{con} + \sigma_{con} * \varepsilon \tag{2}$$

where $\varepsilon \sim N(0, 1)$ is a random sampling error, the μ_{con} and the σ_{con} are mean and standard deviation of the distribution, which derived by a linear layer:

$$\begin{bmatrix} \mu_{con} \\ \log\left(\sigma^2_{con}\right) \end{bmatrix} = W_{con} O_{con} + b_{con} \tag{3}$$

where W_{con} and b_{con} are training parameters, O_{con} is the output of last transformer block, at the position of [CLS] before the context as shown in Fig.3. Also, we conduct the same method to get the Z_{exe}.

Multi-task Decoder Multi-task decoder strives to do multiple generation tasks with latent variables. In this section, we first reconstruct the input of the transformer by replacing embeddings of [CLS] in context with the latent variable Z_{con}, replacing embeddings of [CLS] in exemplar with the latent variable Z_{exe} and replacing embeddings of [S2S] in response with latent variable Z_{res}. Then, based on this input, we use a unidirectional network to accomplish multiple generation tasks. The multiple tasks method has been confirmed to produce an outstanding influence in the field of text generation, so we also apply multi-task to improve our model, including Masked LM, Unidirectional LM, and bag-of-words prediction. The three tasks are all generation tasks, and the bidirectional structure of

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

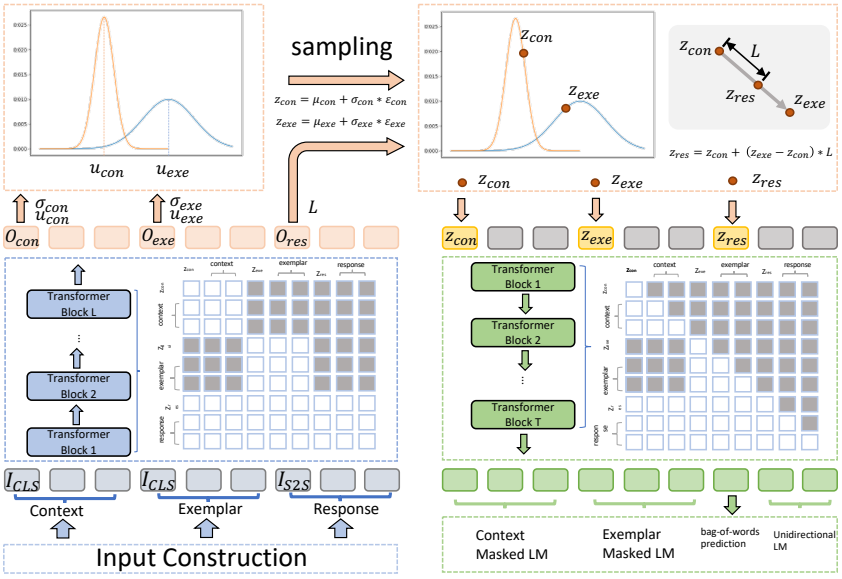

Fig. 3. The architecture of Responses Generation with latent construction and multiple tasks decode. The self-attentions of latent construction and multi-task decoder are shown at the right of transformer blocks. Grey dots represent preventing attention and blank ones represent allowing attention.

Bert will bring future knowledge to the current period. Therefore, as shown in the Fig.3, we use the unidirectional attention mask so that the generation only receive the data ahead of the current time sequence.

For Masked LM, the model could trivially predict the target masked word in a multi-layered network and the prediction objective of masked words in context can be formed as:

$$\mathcal{L}_{MASKC} = -E_{z\sim p(\mathbf{z_{con}}|c)} \log p(w|c) = -E_{z\sim p(\mathbf{z_{con}}|c)} \sum_{i\in I} \log p(w_i|z, w_{<i}) \quad (4)$$

where w_i is the masked word at i-th position and the I is the set of positions of all masked words in context. At the same time, $\mathcal{L}_{MASKE}$ is the prediction goal of masked words in exemplar and is obtained similarly to $\mathcal{L}_{MASKC}$.

For Unidirectional LM, we use the Unidirectional BERT[2] without token mask operation to generate all tokens in response. That is, when generating the t th token in the response r , former tokens are always given to the model as the input. Then generation probability of response is:

$$\mathcal{L}_{NLL} = -E_{z\sim p(\mathbf{z_{res}}|c,e,r)} \log p(r|c, e, z) = -E_{z\sim p(\mathbf{Z_{res}}|c,e,r)} \sum_{t=1}^{T} \log p(r_t|z, c, e, r_{<t}) \quad (5)$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

where z_{res} is estimated by Latent Construction module given a tuple of (c, e, r), and T is the length of response sequence.

For bag-of-words prediction, we use the bag-of-words loss[34] to facilitate the training process of responses' latent variables and tackle the vanishing latent variable problem:

$$\mathcal{L}_{BOW} = -E_{z\sim p(\mathbf{z_{res}}|c,e,r)} \sum_{t=1}^{T} \log p(r_t|c,e,r) = -E_{z\sim p(\mathbf{z}|c,r)} \sum_{t=1}^{T} \log \frac{e^{f_{r_t}}}{\sum_{v\epsilon V} e^{f_v}} \tag{6}$$

where V is vocabulary size and f if a network layer to generate the words in the target response in a non-autoregressive way:

$$f = softmax(W_{bow} h_{res} + b_{bow}) \tag{7}$$

where h_{res} is the output of z_{res} decoded by the T-th transfomer block in multi-task decoder as shown in Fig.3. W_{bow} and b_{bow} are the training parameters.

Overall, the total objective of our model is to jointly minimize the integrated loss:

$$\mathcal{L} = \mathcal{L}_{MASKC} + \mathcal{L}_{MASKE} + \mathcal{L}_{NLL} + \mathcal{L}_{BOW} \tag{8}$$

3.4 Post-training and Fine-tuning

We employ the pre-trained parameters of the UNILM[2] to initialize our network. Though UNILM has been proven to be an effective language model, it can not directly adapt to the task of dialogue generation. Since UNILM is trained on a large non-conversational corpus, and there is a huge natural gap between the dialogic corpus and other corpora. In order to make up for the reduced effect caused by different corpus, we carry out post-training in dialogue corpus. For each dataset, we first post-train the UNILM structure with masked LM task on the dialogue corpus without exemplar sentences. After that, we implement the proposed model to fine-tune corpus with exemplar with all three tasks.

4 Experiments

4.1 Datasets

In order to evaluate the performance of our proposed method, we conducted comprehensive experiments on three different datasets: Persona chat, Daily dialog, and Dstc7-avsd. The daily dialog contains only the dialogue text, but Persona chat and Dstc7-avsd involve knowledge beyond dialogue. To avoid changing the structure of our model, we concatenate knowledge text with dialogue context and treat this combination text as the dialogue context in training.

– Persona chat [35] is a knowledge-based dialogue dataset consisting of 164,356 utterances between crowdworkers who were randomly paired and asked to act the part of a given provided persona (randomly assigned, and created by another set of crowdworkers).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

– Daily dialog [36] is a high-quality multi-turn dialogue dataset containing conversations about our daily life, in which human communicate with others for two main reasons: exchanging information and enhancing social bonding.
– Dstc7-avsd [37] is an abbreviation of dstc7 challenged audio-visual scene aware dialogue. It is a conversational QA dataset. Given dialogue context and background knowledge, the system tries to generate answers in this challenge. In our experiment, we utilize the unimodal information of text, which includes the title and abstract of video.

4.2 Baseline and Evaluation

The following models have been compared in the experiments.

– **Seq2Seq**: Sequence to sequence with attention is employed as the baseline for the experiments.
– **LIC**: LIC obtains the well-known performance[38] in the ConvAI2 challenge [39], in which Persona-Chat dataset is utilized.
– **iVAE MI**: iVAE MI[40] generates diverse responses with sample-based latent representation and achieves state-of-the-art performance on the dataset of Daily Dialog.
– **CMU**: CMU[41] gains excellent achievement in all the evaluation metrics of DSTC7-AVSD.
– **PLATO**: PLATO[14] outperforms other methods in all datasets for now.

We use three metrics to evaluate our proposed model, including automatic evaluation BLEU, DISTINCT, and MSCOCO platform.

– **BLEU**: We adopt BLEU-(1-4) to measure the overlap of candidates and references at the character level.
– **DISTINCT**: This metric is proposed by [42] to evaluate the diversity of the responses. In the generated responses, the number of distinct unigrams and bigrams are divided by the total number of generated unigrams and bigrams in the test set.
– **MSCOCO**: We employ the MSCOCO platform[43] to evaluate the performance in DSTC7-AVSD, including metrics of BLEU, METEOR, ROUGH-L and CIDEr.

The evaluation methods and results of the baselines are from the project of PLATO[14].

4.3 Results and Analysis

Comparison of BLEU As shown in Table 2 and Table 3, the large-scale pre-trained model achieves better performance than the seq2seq on three datasets, which shows the effectiveness of the large-scale pre-trained model in dialogue generation tasks. Our method achieves the best results on Persona-Chat, Daily Dialog datasets. Because our approach introduces a suitable exemplar, providing

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

Table 1. The automatic evaluation results of DSTC7-AVSD

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH-L	CIDEr
DSTC7-AVSD	Baseline	0.626	0.485	0.383	0.309	0.215	0.487	0.746
	CMU	0.718	0.584	0.478	0.394	0.267	0.563	1.094
	PLATO	**0.784**	0.637	0.525	0.435	0.286	0.596	1.209
	Our Method	0.736	**0.644**	**0.607**	**0.522**	**0.297**	**0.610**	**1.219**

better guidance for the generation process, and improving the accuracy. Our model's score is lower than PLATO in the BLEU-1 with DSTC7-AVSD dataset but higher than PLATO in BLEU-2/3/4, which indicates that our model has more advantages in generating more extended responses. It is because much information, which our latent variables provide, may confuse the model when generating short text but be beneficial when generating long text.

Table 2. The automatic evaluation results of Persona-Chat

Dataset	Model	BLEU-1	BLEU-2	Distinct-1	Distinct-2
Persona-Chat	Seq2Seq	0.448	0.353	0.004	0.016
	PLATO w/o Latent	0.405	0.320	0.019	0.113
	LIC	0.458	0.357	0.012	0.064
	PLATO	0.406	0.315	**0.021**	0.121
	Our method	**0.459**	**0.362**	0.020	**0.123**

Table 3. The automatic evaluation results of Daily Dialog

Dataset	Model	BLEU-1	BLEU-2	Distinct-1	Distinct-2
Daily-Dialog	Seq2Seq	0.336	0.268	0.030	0.128
	iVAE MI	0.309	0.249	0.029	0.250
	PLATO w/o Latent	0.405	0.322	0.046	0.246
	PLATO	0.397	0.311	0.053	0.291
	Our method	**0.417**	**0.334**	**0.056**	**0.295**

Comparison of Distinct Our model gets the highest score in distinct-1/2 on the daily dataset, but it is slightly lower than that of PLATO on the distinct-1 on the persona dataset. The most likely reason is that the data in these two datasets are considerably different. The data of daily dialogue only contains dialogue text. The information in conversation without the exemplar is limited, so that information is insufficient to generate various responses. In this case, our method of introducing more information is effective. However, the corpus of the persona dataset contains enough additional information about the dialogue. In this case, the exemplar cannot bring much more useful knowledge about the current conversation, so our method's improvement is limited. However, in most conversation scenes, external knowledge does not exist, so that our model can be more practical than PLATO.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

4.4 Case Study

Table 4. Some responses generated by our method.

Query	Exemplar	Response
The taxi drivers are on strike again .	I'm going to work .	I ' m going to work late .
	OK , let's go and ask.	Let ' s go and try to find one .
	We can have a company car take you there .	You have to take a bus .
	What ? Are you serious ?	Are you serious ?
	Why don't we go for a walk ?	You have to be quick .
I really need to start eating healthier .	Anyway health is the most important thing .	I really need to lose weight .
	Oh I see , okay .	I think so , too .
	In fact your body will only store fat if you miss meals .	I really need to get some exercise .
	Do you know of another good restaurant ?	Do you have any plans ?
	That's a good choice .	That ' s a good idea .

We show some generation examples in the Table 4. According to the same query, we search for different exemplars in the graph to guide the generation process. Combined with different exemplars, the model generates different informative responses. Some of the generated responses have obvious overlap with the exemplar, such as " I'm going to work late." And some others have a weak connection with the exemplar, such as "You have to be quick." This phenomenon demonstrates that both exemplars can guide the generation effectively. Also, most exemplars contain knowledge beyond the query so that exemplars can provide more extra information for model generation.

4.5 The Influence of Exemplar Quality

Table 5. The Influence of Exemplar Quality

Exemplar Position	BLEU-1	BLEU-2	Distinct-1	Distinct-2
Ground Truth	0.470	0.361	0.058	0.310
Exemplar in Graph (Our method)	0.417	0.334	0.055	0.295
Random	0.351	0.326	0.048	0.241
NULL(Our method w/o Latent)	0.356	0.329	0.039	0.234

In order to verify the effectiveness of the exemplar seeking module, we use different quality sentences as the exemplar in our method. We consider the quality of ground truth is the best, and use random sentences and blank sentences to replace exemplars in our model. As shown in Table 5, Random sentences get the worst performance in BLEU because they do not contain accurate information and bring misleading noise to the generation. Blank sentences get the worst

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

score in Distinct owning to they convey the least information to the generation process. Our method gains the closest performance to the ground truth, which shows that the exemplar seeking module is valid.

5 Conclusion

In this paper, we propose the exemplar guided latent pre-trained dialogue generation model to generate diverse and informative responses. In the proposed model, we treat the latent variable as the continuous sentence embedding and induce an enlightened exemplar to guide the generation. Results of experiments prove the proposed model improves the diversity and informativeness of responses.

Acknowledgment

This work was supported by National Natural Science Foundation of China (No. 61906187, No. 61976207, No. 61902394).

References

1. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.
2. L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H. Hon, "Unified language model pre-training for natural language understanding and generation," in *NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*.
3. Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "ERNIE 2.0: A continual pre-training framework for language understanding," in *AAAI,New York, NY, USA, February 7-12, 2020*, 2020.
4. D. Xiao, H. Zhang, Y. Li, Y. Sun, H. Tian, H. Wu, and H. Wang, "ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation," in *Proceedings of IJCAI 2020*, 2020.
5. H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proceedings of ACL 2019, Florence, Italy, July 28- August 2, 2019,*.
6. Y. Zhang, S. Sun, M. Galley, Y. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proceedings of ACL 2020, Online, July 5-10, 2020*.
7. Y. Zeng and J. Nie, "Generalized conditioned dialogue generation based on pre-trained language model," *CoRR*, vol. abs/2010.11140, 2020.
8. X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan, "Knowledge-grounded dialogue generation with pre-trained language models," in *Proceedings of EMNLP 2020, Online, November 16-20, 2020*.
9. Z. Yang, W. Wu, C. Xu, X. Liang, J. Bai, L. Wang, W. Wang, and Z. Li, "Styledgpt: Stylized response generation with pre-trained language models," in *Proceedings of Findings, EMNLP 2020, Online Event, 16-20 November 2020*.
10. C.-S. Wu, S. C. Hoi, R. Socher, and C. Xiong, "TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue," in *Proceedings of EMNLP*, Online, 2020.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

11. Y. Zheng, R. Zhang, M. Huang, and X. Mao, "A pre-training based personalized dialogue generation model with persona-sparse data," in *Proceedings of AAAI 2020, New York, NY, USA, February 7-12, 2020*.
12. Y. Cao, W. Bi, M. Fang, and D. Tao, "Pretrained language models for dialogue generation with multiple input sources," in *Proceedings of EMNLP 2020, Online Event, 16-20 November 2020*.
13. H. Le and S. C. H. Hoi, "Video-grounded dialogues with pretrained generation language models," in *Proceedings of ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds., 2020.
14. S. Bao, H. He, F. Wang, H. Wu, and H. Wang, "PLATO: pre-trained dialogue generation model with discrete latent variable," in *Proceedings of ACL 2020, Online, July 5-10, 2020*.
15. S. Bao, H. He, F. Wang, H. Wu, H. Wang, W. Wu, Z. Guo, Z. Liu, and X. Xu, "PLATO-2: towards building an open-domain chatbot via curriculum learning," *CoRR*, vol. abs/2006.16779, 2020.
16. L. Zhang, Y. Yang, J. Zhou, C. Chen, and L. He, "Retrieval-polished response generation for chatbot," *IEEE Access*, vol. 8, 2020.
17. I. Shalyminov, A. Sordoni, A. Atkinson, and H. Schulz, "Hybrid generative-retrieval transformers for dialogue domain adaptation," *CoRR*, vol. abs/2003.01680, 2020.
18. P. Gupta, J. P. Bigham, Y. Tsvetkov, and A. Pavel, "Controlling dialogue generation with semantic exemplars," *CoRR*, vol. abs/2008.09075, 2020.
19. T. Ma, H. Yang, Q. Tian, Y. Tian, and N. Al-Nabhan, "A hybrid chinese conversation model based on retrieval and generation," *Future Gener. Comput. Syst.*, vol. 114, 2021.
20. J. Weston, E. Dinan, and A. H. Miller, "Retrieve and refine: Improved sequence generation models for dialogue," in *Proceedings of SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018*.
21. G. Pandey, D. Contractor, V. Kumar, and S. Joshi, "Exemplar encoder-decoder for neural conversation generation," in *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*.
22. Y. Song, C. Li, J. Nie, M. Zhang, D. Zhao, and R. Yan, "An ensemble of retrieval-based and generation-based human-computer conversation systems," in *Proceedings of IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*.
23. L. Yang, J. Hu, M. Qiu, C. Qu, J. Gao, W. B. Croft, X. Liu, Y. Shen, and J. Liu, "A hybrid retrieval-generation neural conversation model," in *Proceedings of CIKM 2019, Beijing, China, November 3-7, 2019*.
24. J. Zhang, C. Tao, Z. Xu, Q. Xie, W. Chen, and R. Yan, "Ensemblegan: Adversarial learning for retrieval-generation ensemble model on short-text conversation," in *Proceedings of SIGIR 2019, Paris, France, July 21-25, 2019*.
25. Q. Zhu, L. Cui, W. Zhang, F. Wei, and T. Liu, "Retrieval-enhanced adversarial training for neural response generation," in *Proceedings of ACL 2019, Florence, Italy, July 28- August 2, 2019,*.
26. Y. Wu, F. Wei, S. Huang, Y. Wang, Z. Li, and M. Zhou, "Response generation by context-aware prototype editing," in *Proceedings of AAAI 2019,, Honolulu, Hawaii, USA, January 27 - February 1, 2019*.
27. D. Cai, Y. Wang, W. Bi, Z. Tu, X. Liu, W. Lam, and S. Shi, "Skeleton-to-response: Dialogue generation guided by retrieval memory," in *Proceedings of NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10

28. D. Cai, Y. Wang, W. Bi, Z. Tu, X. Liu, and S. Shi, "Retrieval-guided dialogue response generation via a matching-to-generation framework," in *Proceedings of EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*.
29. H. Cai, H. Chen, Y. Song, X. Zhao, and D. Yin, "Exemplar guided neural dialogue generation," in *Proceedings of IJCAI 2020*.
30. S. Wu, Y. Li, D. Zhang, Y. Zhou, and Z. Wu, "Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact," in *Proceedings of IJCAI 2020*.
31. H. Zhang, Z. Liu, C. Xiong, and Z. Liu, "Grounded conversation generation as guided traverses in commonsense knowledge graphs," in *Proceedings of ACL 2020, Online, July 5-10, 2020*.
32. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*.
33. D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: http://arxiv.org/abs/1312.6114
34. T. Zhao, R. Zhao, and M. Eskénazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proceedings of ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*.
35. S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" in *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*.
36. Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*.
37. H. AlAmri, V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson, S. Lee, and D. Parikh, "Audio visual scene-aware dialog," in *CVPR,2019, Long Beach, CA, USA, June 16-20, 2019*.
38. S. Golovanov, R. Kurbanov, S. I. Nikolenko, K. Truskovskyi, A. Tselousov, and T. Wolf, "Large-scale transfer learning for natural language generation," in *Proceedings of ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*.
39. E. Dinan, V. Logacheva, V. Malykh, A. H. Miller, K. Shuster, J. Urbanek, D. Kiela, A. Szlam, I. Serban, R. Lowe, S. Prabhumoye, A. W. Black, A. I. Rudnicky, J. Williams, J. Pineau, M. S. Burtsev, and J. Weston, "The second conversational intelligence challenge (convai2)," *CoRR*, vol. abs/1902.00098, 2019.
40. L. Fang, C. Li, J. Gao, W. Dong, and C. Chen, "Implicit deep latent variable models for text generation," in *Proceedings of EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*.
41. S. P. Ramon Sanabria and F. Metze, "Cmu sinbads submission for the dstc7 avsd challenge," in *AAAI Dialog System Technology Challenge Workshop*, 2019.
42. J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of NAACL HLT 2016, San Diego California, USA, June 12-17, 2016*.
43. X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *CoRR*, vol. abs/1504.00325, 2015.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_10