PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable

Siqi Bao,[*] **Huang He,**[*] **Fan Wang, Hua Wu** and **Haifeng Wang**

Baidu Inc., China

{baosiqi, hehuang, wangfan04, wu_hua, wanghaifeng}@baidu.com

Abstract

Pre-training models have been proved effective for a wide range of natural language processing tasks. Inspired by this, we propose a novel dialogue generation pre-training framework to support various kinds of conversations, including chit-chat, knowledge grounded dialogues, and conversational question answering. In this framework, we adopt flexible attention mechanisms to fully leverage the bi-directional context and the uni-directional characteristic of language generation. We also introduce discrete latent variables to tackle the inherent one-to-many mapping problem in response generation. Two reciprocal tasks of response generation and latent act recognition are designed and carried out simultaneously within a shared network. Comprehensive experiments on three publicly available datasets verify the effectiveness and superiority of the proposed framework.

1 Introduction

Dialogue generation is a challenging task due to the limited corpus of human conversations, complex background knowledge, and diverse relationships between utterances. Recently, pre-trained large-scale language models, such as BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019), have achieved prominent success in natural language processing. Such models are usually constructed based on a massive scale of general text corpora, like English Wikipedia or BooksCorpus (Zhu et al., 2015), where distributed representations can be learned automatically from the raw text. With these representations being fine-tuned, breakthroughs have been continuously reported for various downstream tasks, especially those on natural language understanding, such as question answering, natural language inference, and so on.

This pre-training and fine-tuning paradigm also sheds light on the tasks of natural language generation, like dialogue generation. However, the previous study demonstrates that there are some deficiencies in performance while directly fine-tuning BERT on small conversation datasets (Rashkin et al., 2019; Wolf et al., 2019). Possible reasons are three-fold: 1) the underlying linguistic patterns in human conversations can be highly different from those in general text, which suggests a potentially large gap in knowledge or data distribution; 2) the training mode of uni-directional dialogue generation is also distinct from that of bi-directional natural language understating as applied in BERT; 3) unlike most of the general NLP tasks, there exists a one-to-many relationship in dialogue generation, where the dialogue context may correspond to multiple appropriate replies.

In this paper, we propose a new method to tackle the above challenges, aiming to obtain a high-quality pre-training model for dialogue generation. First of all, to reduce the gap between data distributions, large-scale Reddit and Twitter conversations are utilized to further pre-train the generation model (upon the basis of language models pre-trained with general text). Secondly, to mitigate the difference in training mode, a flexible paradigm integrating uni- and bi-directional processing is employed in this work, which is inspired by the latest unified language modeling (Dong et al., 2019). Thirdly, a discrete latent variable is introduced to model the one-to-many relationship among utterances in conversations. Each value of the latent variable corresponds to the particular conversational intent of one response, which is referred as *latent speech act*.

Distinct with those controllable dialogue generation based on explicit labels (including emotion, keywords, domain codes, and so on) (Huang et al., 2018; Keskar et al., 2019), our latent variable gets

---

[*]First two authors contributed equally to this work.

exempted from the restriction of human annotations and can be learned automatically from the corpus in an unsupervised way. In the pre-training of dialogue generation, response generation and latent act recognition are carried out simultaneously within a shared network. Based on the context and latent variable, the generation task tries to maximize the likelihood of the target response. Meanwhile, the recognition task aims to estimate the latent variable w.r.t. the given context and target response. Apparently, the accurate recognition of the latent variable is a crucial factor in boosting the quality of response generation.

We conducted experiments on three different kinds of conversation tasks: chit-chat, knowledge grounded conversation, and conversational question answering. Experimental results verify the effectiveness and superiority of our pre-trained model as compared with the other state-of-the-art methods. Our pre-trained models and source code have been released at GitHub, hoping to facilitate further research progress in dialogue generation.[1]

2 Dialogue Generation Pre-training

Given a piece of context, there exist multiple appropriate responses, leading to diverse conversation flows. It is widely recognized that the capability of modeling one-to-many relationship is crucial for the dialogue generation system (Zhao et al., 2017; Chen et al., 2019). To this end, we propose to encode discrete latent variables into transformer blocks for one-to-many relationship modeling, where two reciprocal tasks of response generation and latent act recognition are collaboratively carried out.

2.1 Model Architecture

In our model, there are three elements: dialogue context c, response r and latent variable z. The dialogue context c consists of several history utterances. (For knowledge grounded conversation, it is conventional to concatenate background knowledge into the context as well (Wolf et al., 2019).) The response r is one piece of appropriate reply towards the given context. The latent variable z is one K-way categorical variable $z \in [1, K]$, with each value corresponding to a particular latent speech act in the response.
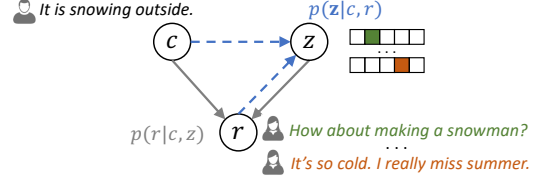
Figure 1: Graphical illustration of response generation (gray lines) and latent act recognition (dashed blue lines).

The probabilistic relationships among these elements are elaborated with the graphical model in Figure 1. Given a context c, there are multiple latent speech acts which can be taken as response intents (represented by the latent variable z). Conditioned on the context and one selected latent speech act, the response is generated as $p(r|c, z)$ (gray lines). Given a pair of context and response, the underlying latent speech act can be estimated as $p(z|c, r)$ (dashed blue lines). As such, our pre-training of dialogue generation contains the following two tasks – **response generation** and **latent act recognition**.

We propose a unified infrastructure for the joint learning of both tasks, shown as Figure 2. The backbone of our infrastructure is inspired by the transformer blocks in (Dong et al., 2019), which supports both bi-directional encoding and uni-directional decoding flexibly via specific self-attention masks. Both response generation and latent act recognition are carried out under the unified network with shared parameters. Their detailed implementations are described as follows.

Given the context c and a specific speech act z, the response generation can be estimated as

$$p(r|c, z) = \Pi_{t=1}^{T} \, p(r_t|c, z, r_{<t}) \,, \quad (1)$$

where T is the length of the target response r and $r_{<t}$ denotes previously generated words. Since the response generation is a uni-directional decoding process, each token in the response only attends to those before it, shown as dashed orange lines in Figure 2.

The latent act recognition task is included to identify the corresponding value of z for the given context and the target response in the training data. The latent act recognition shares network parameters with response generation, but has a separate self-attention mask for bi-directional encoding. As shown in Figure 2, with a special mask symbol [M] as input, it keeps collecting information from the context and target response (red lines). In this way,
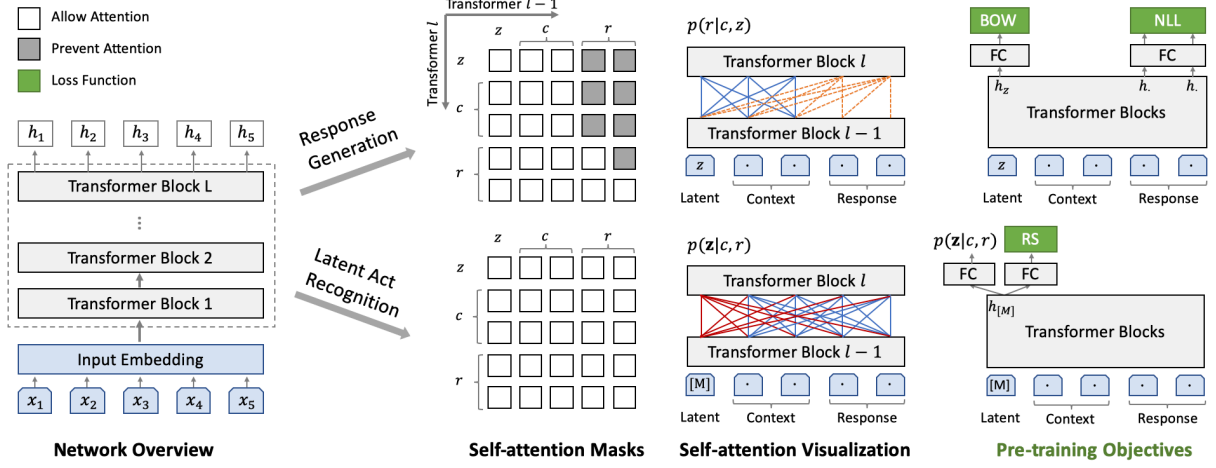
Figure 2: Architecture of dialogue generation with discrete latent variable. In self-attention visualization, red and blue lines denote bi-directional attention, and dashed orange lines denote uni-directional attention.
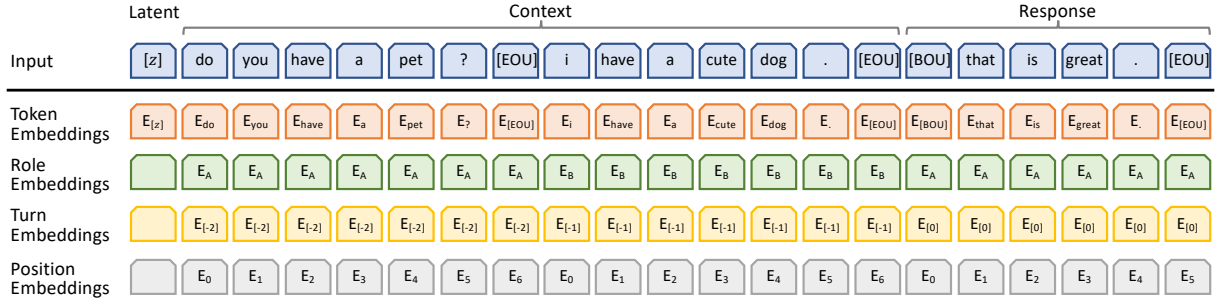

Figure 3: Input representation. The input embedding is the sum of token, role, turn and position embeddings.

the corresponding speech act for the target response can be recognized as $z \sim p(\mathbf{z}|c,r)$, where $p(\mathbf{z}|c,r)$ is the estimated posterior distribution over discrete latent values.

2.2 Input Representation

For multi-turn conversation modeling, elaborate designs have been made on the input representation in this work. For each token, its input embedding is the sum of corresponding token, role, turn and position embeddings. One visual example is shown in Figure 3 and details are described in the following.

- The input is the concatenation of latent variable, dialogue context and response. Following the pre-processing of BERT (Devlin et al., 2019), the input text is tokenized with WordPiece (Wu et al., 2016). A special end-of-utterance [EOU] token is appended to the end of each utterance for separation. Another begin-of-utterance [BOU] token is added at the beginning of the response, whose final hidden state (i.e., output of the last transformer block) is used to predict next token during generation.
- Given that z is one K-way categorical variable,

its token embedding $\mathbf{E}_{[z]}$ is mapped from the latent embedding space $\mathbf{E}_{\mathbf{z}} \in \mathbb{R}^{K \times D}$. For the rest tokens in the vocabulary, they are initialized using BERT's WordPiece embeddings.

- Role embeddings are employed to differentiate the characters evolved in the conversation. The role embedding $\mathbf{E}_A$ is added for the response, as well as dialogue utterances generated by the same character in the context. And role embedding $\mathbf{E}_B$ is used for the other character. (For knowledge grounded conversation, $\mathbf{E}_C$ is used as the role embedding of background knowledge.)
- In the interactive conversation, there are multi-turn utterances and we employ relative order in the assignment of turn embeddings. The turn embedding for the response is set to $\mathbf{E}_{[0]}$, and the turn embedding of its last utterance is $\mathbf{E}_{[-1]}$, and etc. Our utilization of relative turn embeddings instead of absolute ones enables the model to assign turn embedding $\mathbf{E}_{[0]}$ to the response consistently and makes response generation exempt from the disturbance of its round number within the dialogue.
- Position embeddings are added according to the

token position in each utterance. Note that for the special token of latent variable, its corresponding role, turn and position embeddings are all set to empty.

2.3 Pre-training Objectives

We employ three loss functions in dialogue generation pre-training: negative log-likelihood (NLL) loss, bag-of-words (BOW) loss and response selection (RS) loss. Brief illustration is shown in the last column of Figure 2 and detailed descriptions will be provided in this section.

2.3.1 Response Generation

In our model, the response is generated conditioned on the latent variable and the context. The widely adopted NLL loss is embraced in the pre-training:

$$
\begin{aligned}
\mathcal{L}_{NLL} &= -\mathbb{E}_{z \sim p(\mathbf{z}|c,r)} \ \log p(r|c,z) \\
&= -\mathbb{E}_{z \sim p(\mathbf{z}|c,r)} \sum_{t=1}^{T} \ \log p(r_t|c,z,r_{<t}) \ ,
\end{aligned}
\tag{2}
$$

where z is the latent speech act of this training pair (c, r), sampled from the probability distribution $p(\mathbf{z}|c,r)$. The posterior distribution over latent values is estimated through the task of latent act recognition:

$$
p(\mathbf{z}|c,r) = \text{softmax}(W_1 h_{[M]} + b_1) \in \mathbb{R}^K \ , \tag{3}
$$

where $h_{[M]} \in \mathbb{R}^D$ is the final hidden state of the special mask, $W_1 \in \mathbb{R}^{K \times D}$ and $b_1 \in \mathbb{R}^K$ denote the weight matrices of one fully-connected layer.

Besides the classical NLL loss, the bag-of-words loss (Zhao et al., 2017) is also employed to facilitate the training process of latent discrete variables:

$$
\begin{aligned}
\mathcal{L}_{BOW} &= -\mathbb{E}_{z \sim p(\mathbf{z}|c,r)} \sum_{t=1}^{T} \ \log p(r_t|c,z) \\
&= -\mathbb{E}_{z \sim p(\mathbf{z}|c,r)} \sum_{t=1}^{T} \ \log \frac{e^{f_{r_t}}}{\sum_{v \in V} e^{f_v}} \ ,
\end{aligned}
\tag{4}
$$

where V refers to the whole vocabulary. The function f tries to predict the words within the target response in a non-autoregressive way:

$$
f = \text{softmax}(W_2 h_z + b_2) \in \mathbb{R}^{|V|} \ , \tag{5}
$$

where h_z is the final hidden state of the latent variable and $|V|$ is the vocabulary size. f_{r_t} denotes the estimated probability of word r_t. As compared with NLL loss, the BOW loss discards the order of words and forces the latent variable to capture the global information of the target response.

2.3.2 Response Selection

Response selection helps distinguish whether the response is relevant with the dialogue context and consistent with the background knowledge. Meanwhile, its score can be regarded as an indicator of coherence during inference, helping to select the most coherent one from multiple candidate responses.

Particularly, the training of response selection is carried out together with the bi-directional encoding of latent act recognition. The positive training samples come from the dialogue context and corresponding target response (c, r), with label $l_r = 1$. And the negative samples are created by randomly selecting responses from the corpus (c, r^-), with label $l_{r^-} = 0$. The binary cross-entropy loss of response selection is defined as follows:

$$
\mathcal{L}_{RS} = -\log p(l_r = 1|c,r) - \log p(l_{r^-} = 0|c,r^-) \tag{6}
$$

The above probability is estimated through one fully-connected layer, with the final hidden state of the special mask fed as input:

$$
p(l_r = 1|c,r) = \text{sigmoid}(W_3 h_{[M]} + b_3) \tag{7}
$$

To sum up, the total objective of our pre-training model is to minimize the integrated loss:

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

2.4 Pre-training Procedure

Our pre-training model contains 12 transformer blocks, with network parameters initialized using BERT$_{BASE}$. Large-scale conversation datasets – Twitter (Cho et al., 2014) and Reddit (Zhou et al., 2018; Galley et al., 2019) are employed for pre-training, which results in 8.3 million training samples in total. For each training sample of context and target response (c, r), it needs to pass through the network twice to accomplish the tasks of latent act recognition and response generation. And the pre-training steps are summarized as follows:
1) Latent Act Recognition
   – Given a pair of context and target response, estimate the posterior distribution $p(\mathbf{z}|c,r)$
   – Randomly select r^- and calculate $\mathcal{L}_{RS}$
2) Response Generation
   – With the sampled latent value $z \sim p(\mathbf{z}|c,r)$, calculate $\mathcal{L}_{NLL}$ and $\mathcal{L}_{BOW}$
3) Optimization
   – Sum up to obtain $\mathcal{L}$, and update network parameters with back-propagation

The hyper-parameters used in pre-training are listed as follows. The maximum sequence length of context and response is set to 256 and 50, respectively. The number of transformer blocks in our model L is 12 and the hidden embedding dimension D is 768. The batch size is set to 64 and K is set to 20 for the discrete latent variable. Adam optimizer (Kingma and Ba, 2015) is employed for optimization with a learning rate of 5e-5. The pre-training of dialogue generation was carried out on 8 Nvidia Telsa V100 32G GPU cards for 3.5M steps, taking about two weeks to reach convergence.

2.5 Fine-tuning and Inference

Our pre-trained model is flexible enough to support various kinds of dialogues, including chit-chat, knowledge grounded conversation, conversational question answering, etc. The fine-tuning on small conversation datasets can be carried out by following the training objectives defined in Equation (8). As the fine-tuning process reaches convergence, the response towards the given context can be obtained through the following inference procedure:

1) Candidate Response Generation
   – Conditioned on each latent value $z \in [1, K]$, generate corresponding candidate response r.
2) Response Selection
   – Calculate the probability for each response $p(l_r = 1|c, r)$ and select the one with highest coherence value as the final response.

It is worth noting that the above fine-tuning and inference procedures are set up for the dialogue generation without any specific objectives. If there exists a specific objective within the conversation, such as letting both participants know more about each other (Bao et al., 2019), the fine-tuning can proceed to maximize the pre-defined rewards with reinforcement learning (RL). Under such circumstances, our latent discrete variable can be naturally treated as action within RL, and thus the response selection can be straightforwardly solved by selecting the action that results in the maximum reward.

3 Experiments

3.1 Settings

3.1.1 Datasets

To evaluate the performance of our proposed method, comprehensive experiments have been carried out on three publicly available datasets.

- Persona-Chat (Zhang et al., 2018) is a knowledge grounded conversation dataset. It provides both manually annotated conversations and corresponding persona profiles (background knowledge), where two participants chat naturally and try to get to know each other.
- Daily Dialog (Li et al., 2017) is a chit-chat dataset, which contains high-quality human conversations about daily life.
- DSTC7-AVSD (Alamri et al., 2019), short for Audio Visual Scene-aware Dialog of the DSTC7 challenge, is a conversational question answering dataset. In DSTC7-AVSD, the system need to generate an answer given dialogue context and background knowledge. There are two available options of knowledge utilization: 1) using single-modal information of text only, including video's caption and summary; 2) relying on multi-modal information, including text, audio and visual features. The single-modal option is adopted by our method in the experiments.

The descriptions and statistics of these datasets are summarized in Table 1.

3.1.2 Compared Methods

The following models have been compared in the experiments.

Baseline. Sequence to sequence with attention (Seq2Seq) (Vinyals and Le, 2015) is employed as the baseline for the experiments on Persona-Chat and Daily Dialog. DSTC7-AVSD has provided a baseline system, which is built upon hierarchical recurrent encoders with multi-modal features.

State of the art. Persona-Chat was also utilized in the ConvAI2 challenge (Dinan et al., 2019a), where the team of Lost in Conversation (LIC) (Golovanov et al., 2019) obtains the best performance. LIC is also one transformer based generation method and fine-tuned upon the pre-trained model of GPT (Radford et al., 2018). For the dataset of Daily Dialog, its best results are reported by the recently developed method – iVAE$_{MI}$ (Fang et al., 2019), which generates diverse responses with sample-based latent representation. In DSTC7-AVSD, the team of CMU (Sanabria et al., 2019) obtains the best performance across all the evaluation metrics.

Our method. To better analyze the effects of our latent discrete variable, we also compare to the version without latent variable (Our w/o Latent).[2]

---

[2]It shares the same training settings as our method with latent variables: network parameters are first initialized with BERT$_{BASE}$, and the pre-training is further carried out on Reddit and Twitter. The only difference lies in the incorporation of latent variable.

Dataset	Type	Knowledge	# Train	# Valid	# Test
Persona-Chat	Chit-chat with persona	Persona profiles	8,939 dialogues 131,438 turns	1,000 dialogues 15,602 turns	968 dialogues 15,024 turns
Daily Dialog	Chit-chat	N/A	11,118 dialogues 87,170 turns	1,000 dialogues 8,069 turns	1,000 dialogues 7,740 turns
DSTC7-AVSD	Conversational QA	Video caption & summary	7,659 dialogues 153,180 turns	1,787 dialogues 35,740 turns	1,710 dialogues 13,490 turns

Table 1: Summary of datasets used in the experiments.

Dataset	Model	Automatic Evaluation			Human Evaluation			
		BLEU-1/2	Distinct-1/2	Knowledge R/P/F1	Fluency	Coherence	Informativeness	Overall
Persona-Chat	Seq2Seq	0.448 / 0.353	0.004 / 0.016	0.004 / 0.016 / 0.006	1.82	0.37	0.85	0.34
	LIC	0.405 / 0.320	0.019 / 0.113	0.042 / 0.154 / 0.064	1.95	1.34	1.09	1.29
	Our w/o Latent	**0.458 / 0.357**	0.012 / 0.064	0.085 / 0.263 / 0.125	1.98	1.36	1.04	1.30
	Our Method	0.406 / 0.315	**0.021 / 0.121**	**0.142 / 0.461 / 0.211**	**1.99**	**1.51**	**1.70**	**1.50**
Daily Dialog	Seq2Seq	0.336 / 0.268	0.030 / 0.128	-	1.85	0.37	0.44	0.33
	iVAE$_{MI}$	0.309 / 0.249	0.029 / 0.250	-	1.53	0.34	0.59	0.30
	Our w/o Latent	**0.405 / 0.322**	0.046 / 0.246	-	1.91	**1.58**	1.03	1.44
	Our Method	0.397 / 0.311	**0.053 / 0.291**	-	**1.97**	1.57	**1.23**	**1.48**

Table 2: Experimental results on Persona-Chat and Daily Dialog with automatic and human evaluations, with highest value written in bold.

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
	Our Method	**0.784**	**0.637**	**0.525**	**0.435**	**0.286**	**0.596**	**1.209**
	Our Method Upper Bound	0.925	0.843	0.767	0.689	0.361	0.731	1.716

Table 3: Experimental results on DSTC7-AVSD with automatic evaluation, with highest value written in bold.

Context	breaking news on the presidential race that could change your vote
Generated Responses	i m not voting for hillary i m voting for trump
	i can t believe this is happening
	it would be interesting to see the results of this election
	trump is a role model that the news media mafia hides
	i will be voting for hillary
Context	i plan on walking into the appalachian mountains and never coming back .
Generated Responses	i've been to the appalachian mountains a few times . it's a beautiful place .
	i've never been to the appalachian mountains , but i've heard it's nice .
	i wouldn't want to live there .
	don't worry , you'll come back .
	that sounds like a good plan .

Table 4: Examples of response generation with our pre-trained model.

Index	Initialization		Training Configurations			# Fine-tuning Dialogues		
	Model	Attention	Context Attention	Data	Latent	1k	5k	9k
1.1	BERT	Bi-direction	Bi-direction	-	-	58.091	33.143	26.727
1.2	GPT-2	Uni-direction	Uni-direction	-	-	31.251	25.630	24.638
1.3	GPT-2	Uni-direction	Bi-direction	-	-	25.193	18.225	16.538
2.1	GPT-2	Uni-direction	Bi-direction	Twitter & Reddit	-	16.141	13.981	13.332
2.2	BERT	Bi-direction	Bi-direction	Twitter & Reddit	-	15.836	13.799	13.105
3.1	BERT	Bi-direction	Bi-direction	Twitter & Reddit	✓	**15.080**	**12.936**	**12.285**

Table 5: Perplexity of different pre-trained models on Persona-Chat, with best value written in bold.

3.1.3 Evaluation Metrics

Both automatic and human evaluations are employed to assess the performance of compared methods. In automatic evaluation, the following metrics are included:

- BLEU (Chen and Cherry, 2014) measures the n-gram overlap between generated response and the target response.
- Distinct-1/2 (Li et al., 2016) measures the generation diversity, which is defined as the number of distinct uni- or bi-grams divided by the total amount of generated words.
- Knowledge Recall/Precision/F1 (Dinan et al., 2019b) measures the degree of informativeness w.r.t. background knowledge.
- In DSTC7-AVSD, the MSCOCO platform (Chen et al., 2015) is employed for evaluation. It compares the generated response with six ground truth responses, using metrics of BLEU, ME-TEOR, ROUGH-L and CIDEr.

In human evaluation, we randomly select 100 dialogue contexts and generate responses with compared methods. Three crowd-sourcing workers are asked to score the response quality on a scale of [0, 1, 2] from four aspects – fluency, coherence, informativeness and overall. The higher score, the better. Details about the criteria are given as follows.

- Fluency measures whether the generated sentence is smooth and grammatically correct.
- Coherence evaluates whether the generated response is relevant with the dialogue context and consistent with the expressed information or background knowledge.
- Informativeness assesses whether the response is informative or not.
- Overall represents the general evaluation, where 0 indicates a bad response, 1 refers to a normal response and 2 stands for a good response.

After collecting the assessments from annotators, the response's final score is determined via majority voting. The average Fleiss's kappa (Fleiss and Cohen, 1973) on Persona-Chat and Daily Dialog is 0.515 and 0.480 respectively, indicating annotators have reached moderate agreement.

3.2 Experimental Results

The experimental results on Persona-Chat and Daily Dialog with automatic and human evaluations are summarized in Table 2. As suggested in the empirical study (Liu et al., 2016), the correlation between automatic metrics and human judgments is weak in open-domain dialogue generation. In the automatic evaluation, experimental results demonstrate that no method can consistently outperform the others.

During human evaluations, our method achieves better performance consistently across all the metrics on Persona-Chat and Daily Dialog. The scores of fluency almost approach the upper bound, revealing that our generated responses are very fluent. The informativeness assessments indicate that the information in our generated responses is significantly richer, as compared with the baseline methods. Our responses are coherent with the context and favored most by crowd-sourcing workers. The ablation study with our method and our w/o latent also suggests that through the incorporation of discrete latent variables, remarkable improvements can be achieved for dialogue generation. In addition, it can be observed that the generation quality of transformed-based approaches (LIC and our method) is significantly better than that of RNN-based methods (Seq2Seq and iVAE$_{MI}$).[3]

The experimental results on DSTC7-AVSD with automatic evaluation are provided in Table 3. In the

---

[3]It is a normal phenomenon that the performance of our w/o latent is close to that of LIC. Both of them initialize network parameters with pre-trained language models and continue training with large-scale conversation data as Reddit.

experiments, our response selection is strengthened with an extra ranking step, which learns to rank the candidates according to automatic scores and selects the top one as the final answer. The results in Table 3 demonstrate that our method has brought a new breakthrough for DSTC7-AVSD. Additionally, the upper bound of our method is also reported, under the ideal scenario that the optimal candidate answer can be selected.[4] The incredible results validate the great potential of our approach.

3.3 Discussions

3.3.1 Case Analysis

To further dissect the quality of our pre-trained model, several examples of generated responses are provided in Table 4. For each piece of context, our model can produce multiple responses by assigning distinct values to the latent variable and five candidate responses are selected for display in the table. It shows that our pre-trained model is able to generate diverse and appropriate responses. More examples on the conversational datasets are provided in the Appendix.

3.3.2 Comparison of Pre-trained Models

To further analyze the effectiveness of our pre-trained model, more ablation studies have been conducted on Persona-Chat. Distinct pre-trained models are included for comparison. To be fair, their transformer layers are all set to 12. There are three different sizes of training dialogues: 1k, 5k and 9k (all training data). The training configurations and experimental results measured with perplexity are summarized in Table 5. There are three groups of pre-trained models: group 1 applies direct fine-tuning of BERT or GPT-2 (Radford et al., 2019) on Persona-Chat; group 2 employs Twitter and Reddit for further training upon the basis of pre-trained language models; group 3 carries out the training process with latent variable.[5] (Model 2.2 is our w/o latent one and model 3.1 is our method.)

These results demonstrate that our method outperforms the other pre-trained models consistently

with lower perplexity across different training sets. Several interesting conclusions can be also drawn from these results. Firstly, the comparison between model 1.2 and model 1.3 encourages the adoption of flexible attention mechanism to fully leverage the bi-directional context information.[6] Secondly, the superiority of group 2 over group 1 mainly comes from the employment of Twitter and Reddit, which are closer to human conversations than general text. Thirdly, the comparison between model 2.2 and model 3.1 reflects that the incorporation of discrete latent variable is able to boost the quality of response generation, whose effects have also been verified in Table 2.

4 Related Work

Related work involves pre-trained language models and one-to-many modeling in dialogue generation. **Pre-trained Language Models.** Pre-trained language models, which are trained on massive general text, have brought many breakthroughs on various NLP tasks. These models can be roughly divided into two categories according to their attention mechanisms. GPT (Radford et al., 2018) and GPT-2 (Radford et al., 2019) are representative uni-directional language models, where one token is only allowed to attend its previous tokens and the objective is to maximize left-to-right generation likelihood. BERT (Devlin et al., 2019) and XL-Net (Yang et al., 2019) are bi-directional language models, where bi-directional context attention is enabled for token prediction. The latest unified language model UniLM (Dong et al., 2019) is able to support both uni- and bi-directional attention with flexible self-attention mask designs. Recently, some attempts (Golovanov et al., 2019; Wolf et al., 2019; Zhang et al., 2019) have been made to adapt generative language models GPT or GPT-2 for dialogue generation. Whereas the special issues of conversations, such as impacts from background knowledge and problems of one-to-many relationship, are not fully considered and tackled in these adaptations.

One-to-many Modeling. Given one piece of context, there exists multiple appropriate responses, which is know as the one-to-many mapping problem. To model this one-to-many relationship, CVAE (Zhao et al., 2017) employs Gaussian distri-

---

[4]Given a dialogue context and background knowledge, our model is able to generate K diverse responses. Each of them will be evaluated using MSCOCO and the one obtaining the best score will be treated as the optimal candidate answer.

[5]Overall, group 1 involves two-stage training: pre-training of language model with general text and fine-tuning on small conversation datasets. Whereas, group 2 and group 3 involve three-stage training: pre-training of language model with general text, further pre-training of dialogue generation with Twitter and Reddit, and fine-tuning on small conversation datasets.

[6]The results of model 1.1 demonstrate that there are some deficiencies in performance to apply direct fine-tuning of BERT on small conversation datasets, as discussed in the introduction.

bution to capture the discourse-level variations of responses. To alleviate the issue of posterior collapse in VAE, some extension approaches are further developed, including conditional Wasserstein auto-encoder of DialogWAE (Gu et al., 2019) and implicit feature learning of iVAE$_{MI}$ (Fang et al., 2019). SpaceFusion(Gao et al., 2019) aims to jointly optimize diversity and relevance in the latent space, which are roughly matched by the distance and direction from the predicted response vector. Besides the continuous representation in VAE, discrete categorical variables are also utilized for interpretable generation (Zhao et al., 2018). Additionally, multiple mapping modules as latent mechanisms are introduced for diverse generation (Chen et al., 2019), where accurate optimization is carried out via posterior mapping selection. However, due to the small scale of annotated conversation data and limited capacity of generation network, it remains challenging for these methods to balance the diversity and fluency during response generation.

5 Conclusion

A novel pre-training model for dialogue generation is introduced in this paper, incorporated with latent discrete variables for one-to-many relationship modeling. To pre-train our model, two reciprocal tasks of response generation and latent recognition are carried out simultaneously on large-scale conversation datasets. Our pre-trained model is flexible enough to handle various down-stream tasks of dialogue generation. Extensive and intensive experiments have been carried out on three different kinds of publicly available datasets. And the results demonstrate that our model obtains significant improvements over the other state-of-the-art methods.

Our work can be potentially improved with more fine-grained latent variables. In the future, we will also explore to boost the latent selection policy with reinforcement learning and extend our pre-training to support dialogue generation in other languages.

Acknowledgments

References

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Siqi Bao, Huang He, Fan Wang, Rongzhong Lian, and Hua Wu. 2019. Know more about each other: Evolving dialogue strategy via compound assessment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5382–5391.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 362–367.

Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 2019. Generating multiple diverse responses with multi-mapping and posterior mapping selection. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4918–4924.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019a. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. *International Conference on Learning Representations*.

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. *arXiv preprint arXiv:1905.03197*.

Le Fang, Chunyuan Li, Jianfeng Gao, Wen Dong, and Changyou Chen. 2019. Implicit deep latent variable models for text generation. *arXiv preprint arXiv:1908.11527*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. In *Educational and Psychological Measurement*, pages 613–619.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7. In *AAAI Dialog System Technology Challenge Workshop*.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly optimizing diversity and relevance in neural response generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1229–1238.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058.

Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim. 2019. Dialogwae: Multimodal response generation with conditional wasserstein autoencoder. *International Conference on Learning Representations*.

Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–54.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 986–995.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Technical report, OpenAI*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Technical report, OpenAI*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Ramon Sanabria, Shruti Palaskar, and Florian Metze. 2019. Cmu sinbad's submission for the dstc7 avsd challenge. In *AAAI Dialog System Technology Challenge Workshop*.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1098–1107.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 654–664.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4623–4629.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27.

A Additional Case Analysis

In Table 6, it provides the cases of our method and compared approaches on Persona-Chat, where two participants chat with each other according to their personas. As shown in the example, participant P2 needs to produce a response towards the given dialogue context, conditioned on his/her persona profile. The baseline Seq2Seq tends to generate common replies with low informativeness and poor coherence. LIC and our w/o latent are able to produce some coherent responses, whereas deficient in informativeness. In comparison, the response by our method is not only coherent with the context, but also expressive of the background personas. Besides, we also observe the phenomenon of diverse knowledge usage in our response generation, which suggests that the latent variable helps control the knowledge selection and utilization in an implicit way.

Table 7 provides the generated responses on Daily Dialog, where two participants chat about daily life. This example shows that Seq2Seq is able to generate fluent utterances, while lacking coherence with the context. As for iVAE$_{MI}$, it suffers

from the difficulty to balance diversity and fluency. By contrast, our method is able to generate more coherent and high-quality responses.

Table 8 provides the generated responses on DSTC7-AVSD, where two participants discuss the objects and events in a video. Participant P1 is responsible to raise questions, who only has access to the first, middle and last frames of the video. Participant P2 has watched the whole video and needs to answer the partner's questions. The generation system is developed to mimic P2 and answer the questions based on the background knowledge. The baseline approach relies on the multi-modal information, including text, audio and visual features, to produce the answer. Due to the limited network capacity, it lacks fidelity to the background knowledge and makes the generated response less accurate. (As the team of CMU has not released their codes or models, their samples are omitted here.) Our method utilizes the video caption and summary as background knowledge. It can be observed that our method generates a more appropriate answer as compared with the baseline approach.

Persona Profiles [P2]	Context
i love trying barbeque restaurants . i am an administrative assistant . i am looking forward to retiring in five years . i like ot play racquetball .	[P2] hi , do you like bbq ? [P1] hello yes i love bbq [P2] i love restaurants with bbq , they are a lot of fun . [P1] my poor toes are so webbed [P2] really ? afterwards , i go and play racquetball to loose some calories . [P1] cool i like to paint [P2] what do you do as a profession ? i work as administrative assistant [P1] i dig in people trash for things to build [P2] sounds unique but that is more of a hobby instead of a profession [P1] true . i paint for a living
Seq2Seq	that sounds like a lot of work . do you have a job ?
LIC	yeah it is a lot of fun
Our w/o Latent	do you have any hobbies ?
Our Method	that is cool , i am looking forward to retiring in 5 years

Table 6: Case analysis of response generation on Persona-Chat.

Context	[P2] have you seen the new cosmo magazine that i was reading this morning ? [P1] i think i may have seen it on the kitchen table . [P2] you're right , thanks . there's an interesting interview with a top director about his latest film that i want to finish . [P1] who's the director ?
Seq2Seq	yes , i did . i found it very disappointing .
iVAE$_{MI}$	the cake and is very popular .
Our w/o Latent	i'm not sure of his name , but i know it's going to be a thriller .
Our Method	i'm not sure , but i think he might be the director of gone with the wind .

Table 7: Case analysis of response generation on Daily Dialog.

Video Caption	a man closes his window , then he sneezes twice before taking a drink . then he opens up a bag and digs through it looking for something before walking out of the room .
Video Summary	a man closes the window , goes to the table and goes through the items in a bag , takes a drink from the green cup and leaves the room .
Context	[P1] what is the guy doing at the window ? [P2] the guy is closing the window [P1] what does he do after that ?
Baseline	he picks up a book from the table
Our Method	he goes to the table and takes a drink from a green cup

Table 8: Case analysis of response generation on DSTC7-AVSD.