Towards Efficient Dialogue Pre-training with Transferable and Interpretable Latent Structure

Xueliang Zhao[1,3]*, Lemao Liu[2], Tingchen Fu[4], Shuming Shi[2], Dongyan Zhao[1,3,5]†, Rui Yan[4]†
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Tencent AI Lab [3]Center for Data Science, AAIS, Peking University
[4]Gaoling School of Artificial Intelligence, Renmin University of China
[5]Beijing Institute for General Artificial Intelligence
{xl.zhao,zhaody}@pku.edu.cn {redmondliu,shumingshi}@tencent.com
lucas.futingchen@gmail.com ruiyan@ruc.edu.cn

Abstract

With the availability of massive general-domain dialogue data, pre-trained dialogue generation appears to be super appealing to transfer knowledge from the general domain to downstream applications. In most existing work, such transferable ability is mainly obtained by fitting a large model with hundreds of millions of parameters on massive data in an *exhaustive* way, leading to inefficient running and poor interpretability. This paper proposes a novel dialogue generation model with a latent structure that is easily transferable from the general domain to downstream tasks in a *lightweight* and *transparent* way. Experiments on two benchmarks validate the effectiveness of the proposed model. Thanks to the transferable latent structure, our model is able to yield better dialogue responses than four strong baselines in terms of both automatic and human evaluations, and our model with about 22% parameters particularly delivers a 5x speedup in running time compared with the strongest baseline. Moreover, the proposed model is explainable by interpreting the discrete latent variables.

1 Introduction

Conversation between humans and machines has long been a goal of artificial intelligence (AI). Building an open-domain dialogue system with data-driven techniques has gotten a lot of attention in the AI and NLP fields in recent years, thanks to breakthroughs in deep learning (Sutskever et al., 2014; Gehring et al., 2017; Vaswani et al., 2017). In particular, with the availability of massive human dialogue data (e.g., the Reddit comments) on social media (Adiwardana et al., 2020), pre-trained dialogue generation appears to be super appealing to alleviate potential discrepancies between general domain and downstream applications (Zhang et al., 2020; Bao et al., 2020, 2021; Li et al., 2021).

The common idea behind the pre-trained dialogue generation can be highlighted as a two-step pipeline: a) it firstly trains a deep neural model on massive general-domain dialogue data, b) and then transfers the model into downstream tasks via fine-tuning or zero-shot learning. Under this pipeline, the transferability is mainly obtained by fitting a large model with millions of parameters on massive data in an *exhaustive* way. Consequently, the downsides in existing works are obvious: their running is inefficient and their outputs are difficult to explain.

This paper thereby aims to build a pre-trained dialogue model which is easily transferable from the general domain to downstream tasks in a *lightweight* and *transparent* way. To this end, we propose a novel dialogue model with a latent structure consisting of several latent variables. By using some self-supervised tasks to endow its latent variables with some prior properties during training, the latent structure makes the knowledge better transferable across different domains. Specifically, we first propose to incorporate the transformer architecture with a discrete conversation flow. Given a dialogue session, our model will sequentially infer the discrete state for each utterance which provides essential hints for future states and has an effect on the generation of the associated utterance. We further propose a method to disentangle the context-sensitive information from the conversation flow, which is achieved by two disentangled latent variables to capture the context-sensitive information (e.g., topic and persona) and the context-independent information (e.g., dialogue logic for each utterance) respectively. Through tailor-designed self-supervised tasks, the context-sensitive latent variable is able to capture the holistic information of a dialogue session while the context-independent variable is supposed to reflect the dynamic flow of dialogue in each utterance. Meanwhile, the model is optimized with variational

*This work was done while X. Zhao was an intern at Tencent AI Lab.
†Corresponding authors: Dongyan Zhao and Rui Yan.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10051 - 10063
December 7-11, 2022

inference by maximizing the evidence lower bound of the likelihood.

We conduct experiments with two multi-turn dialogue generation benchmarks, including DailyDialog (Li et al., 2017) and ConvAI2 (Dinan et al., 2020). Thanks to the transferable latent structure, our model is able to yield better dialogue responses than four strong baselines in terms of both automatic and human evaluations, and our model including about 22% - 66% parameters particularly delivers a 2x - 30x speedup in running time. Moreover, the proposed model is explainable by visualizing the discrete latent variables.

Our contributions in the paper are three-fold: (1) We present a context-free dialogue structure that captures the prior knowledge about state transition in a large-scale dialogue corpus. Furthermore, with the help of this dialogue structure, our model outperforms the state-of-the-art dialogue pre-training method with much fewer parameters. (2) We propose a disentangled structure learning framework to induce a context-free dialogue structure that enjoys better transferability and interpretability. (3) We empirically verify the effectiveness and efficiency of the proposed model on two benchmarks.

2 Related Work

The success of neural networks in machine translation promotes early research on end-to-end open-domain dialogue generation (Ritter et al., 2011; Shang et al., 2015; Vinyals and Le, 2015). Various adaptations to the vanilla encoder-decoder architecture have been built to model the structure of dialogue contexts (Serban et al., 2016, 2017; Zhang et al., 2019); improve response diversity (Li et al., 2015; Zhao et al., 2017; Tao et al., 2018); introduce external knowledge (Dinan et al., 2019; Zhao et al., 2020a,b); and control response qualities (Xu et al., 2019; Zhou et al., 2017; Zhang et al., 2018; Wang et al., 2018; See et al., 2019).

Large-scale pre-training for open-domain dialogue generation has recently become promising as a way to bridge the gap between conversation with existing systems and conversation with humans. Inspired by the successfulness of GPT-2 (Radford et al., 2019), Zhang et al. (2020) propose to train the transformer models on a very large dialogue dataset to generate informative text. Bao et al. (2020) further use discrete latent variables to address the one-to-many mapping problem in open-domain dialogue. Despite prior successes, the dialogue context is simply concatenated as a long sequence, which may fail to capture the discourse-level coherence among utterances. To this end, Gu et al. (2021) and Li et al. (2021) introduce more self-supervision objectives to capture the discourse-level coherence and the dynamic information flow respectively.

The concept of dialogue structure has proven useful in modeling the complicated relationships between utterances. In the field of task-oriented dialogue, Shi et al. (2019) propose a discrete variational recurrent neural network (DVRNN) to learn the dialogue structure through unsupervised learning; Qiu et al. (2020) further propose to enhance prior work with a structured attention mechanism; and Sun et al. (2021) propose a conversational graph to represent deterministic dialogue structure, where nodes and edges represent the utterance and context information, respectively. In the field of open-domain dialogue, Xu et al. (2021) construct a large dialogue structure graph with around 1.6 million vertices to cover a wide range of topics. This work introduces a disentangled structure learning framework, which can induce a transferable substructure and an interpretable dialogue substructure, to incorporate the structural bias in dialogue pre-training. Thanks to the tailor-designed self-supervised tasks, our latent structure is more general than the dialogue structure in existing work.

3 Approach

3.1 Overview

Let $X = (u_1, u_2, \cdots, u_n)$ denote a dialogue session, with $u_t = (w_{t,1}, w_{t,2}, \cdots, w_{t,m})$ denoting the t-th utterance and $w_{t,i}$ the i-th token in it. The number of utterances in a session and the number of tokens in each utterance are represented by n and m, respectively. The conversational context for u_t is $u_{<t} = (u_1, u_2, \cdots, u_{t-1})$. Our ultimate goal is to develop a generation model $p(u_t|u_{<t})$ that can predict the next utterance based on the context of the conversation.

Figure 1 illustrates the overview of our graphical model, which includes the proposed **latent structure** consisting of three kinds of latent variables, i.e., $c = [c_1, c_2, \cdots, c_n]$, $z^I = [z_1^I, z_2^I, \cdots, z_n^I]$ and z^S. Specifically, c depicts the flow of a conversation, and each $c_i \in \{1, \cdots, N\}$ is a discrete latent variable with N as a hyper-parameter. It is worth noting that c is designed for *interpretability*: by interpreting these discrete variables, humans are

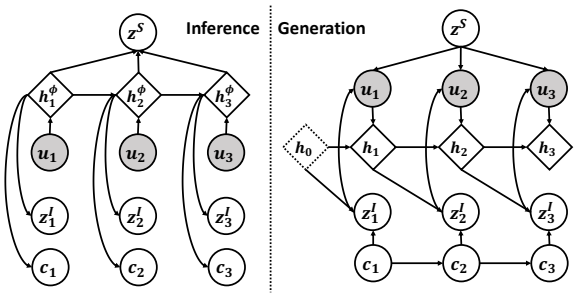

Figure 1: Graphical illustrations of generation and inference processes. Left: inference of the approximate posterior as described in Section 3.3. Right: generation of u and computation of prior as described in Section 3.2. The latent structure consists of c, z^I and z^S. The initial hidden state h_0 is a trainable parameter.

able to understand the logical flow of the conversation as to be shown in Section 5.4. Moreover, z^S and z^I are two disentangled latent variables to capture the context-sensitive information and context-independent information in a dialogue session respectively. In this way, through disentangling z^S and z^I with tailor-designed self-supervised learning objectives (as will be described in Section 3.4), our model is able to capture intrinsic conversation flow for better generalization to different domains (i.e., *transferability*).

With our designed latent structure, given a conversational context $u_{<t}$, the generation of the next utterance u_t can be roughly decomposed into two steps: (1) *infer* the conversation flow $[c_1, c_2, \cdots, c_{t-1}]$ and the context-sensitive variable z^S based on context information, as shown in Figure 1 (left). (2) compute the priors of c_t and z_t^I, and then *generate* the next utterance u_t with z_t^I and z^S, as shown in Figure 1 (right).

3.2 Generation

Context Encoding. We first obtain the contextualized representations of utterances through pre-trained language models (PLMs). Specifically, we exploit GPT-2 (Radford et al., 2019), which is pre-trained using the causal language modeling objective and achieves state-of-the-art results on a range of text generation tasks, as the backbone of our model. Note that our technical novelty lies in the proposal of a disentangled structure learning framework that injects a transferable dialogue structure into PLMs. Given a dialogue session $X = (u_1, u_2, \cdots, u_n)$, we first construct the input I by concatenating all utterances as a single consecutive token sequence:

$$I = [\text{BOS}]u_1[\text{EOS}]u_2[\text{EOS}]\ldots[\text{EOS}]u_n[\text{EOS}], \tag{1}$$

where [BOS] and [EOS] are special tokens designed to separate sentences. The input I is then fed into the PLM and the contextualized representation for X is defined as the hidden states at the last layer:

$$h_{1,1}, \cdots, h_{t,i}, \cdots, h_{n,m} = f_{trans}(I) \in \mathbb{R}^{mn \times d}, \tag{2}$$

where $f_{trans}(\cdot)$ denotes the transformer model (Vaswani et al., 2017) and $h_{t,i} \in \mathbb{R}^d$ denotes the hidden state corresponding to token $w_{t,i}$. It's notable that we use uni-directional attention since the learning objectives are applied to all utterances (as will be illustrated in Section 3.4) and a bi-directional architecture will leak the future information.

The vector representation of the t-th utterance is obtained through attentive pooling (Wu et al., 2020), which is defined as follows:

$$h_t = \sum_{j=1}^{m} \alpha_{t,i} h_{t,i}, \quad \alpha_{t,i} = \frac{e^{q \cdot h_{t,i}}}{\sum_{i=1}^{m} e^{q \cdot h_{t,i}}}, \tag{3}$$

where $q \in \mathbb{R}^d$ is the attention query vector.

Prior of Discrete Latent Variable. The discrete latent variables $[c_1, c_2, \cdots, c_n]$ are used to automatically discover the structural representation in dialogues, which is beneficial to analyze how conversation flow from one utterance to the next one and promotes interpretability. We exclude the impact of $u_{<t}$ on c_t since there is usually a domain discrepancy between the pre-trained and downstream data, which limits the transferability of the learned conversation flow. As a result, we directly model the influence of $c_{<t}$ on c_t in the prior. We employ the transformer model with uni-directional attention to generate the contextualized representation of $c_{<t}$:

$$h_1^c, \cdots, h_{t-1}^c = f_{c-trans}([c_1, \cdots, c_{t-1}]) \in \mathbb{R}^{(t-1) \times d}. \tag{4}$$

Then the probability of predicting c_t is defined as:

$$p(c_t | c_{<t}) = \text{Softmax}(f_{c-mlp}(h_{t-1}^c)), \tag{5}$$

where $f_{c-mlp}(\cdot)$ denotes a MLP network. Different from Shi et al. (2019), our model preserves the capacity to represent n-gram transition probability, which is superior for capturing long-term dependency in the conversation flow of open-domain dialogues.

Priors of Context-Sensitive and Context-Independent Variables. Despite the fact that the discrete latent variables can intuitively characterize conversation flow, the complexity of open-domain conversation necessitates a large number of dialogue states to address fine-grained semantics (Xu et al., 2021), making model training highly challenging and resulting in poor generalization capacity. To alleviate the aforementioned difficulties, we introduce two latent variables to decouple the conversation flow and contextual information, namely the context-sensitive latent variable z^S and the context-independent latent variable z^I.

The prior of context-sensitive latent variable z^S is defined as a standard Gaussian distribution:

$$p(z^S) = \mathcal{N}(0, \mathbf{I}), \tag{6}$$

where $\mathbf{I}$ denotes the unit matrix.

The context-independent latent variable is responsible for capturing dynamic information in each utterance. To achieve this, we condition the prior of z_t^I on both contextualized representation of the previous utterance and the predicted discrete state for the current utterance:

$$\begin{aligned} p(z_t^I|u_{<t}, c_t) &= \mathcal{N}(\mu_t^I, \sigma_t^I \mathbf{I}), \\ \mu_t^I, \sigma_t^I &= f_{I-mlp}([h_{t-1}; e(c_t)]), \end{aligned} \tag{7}$$

where $f_{I-mlp}(\cdot)$ denotes a MLP network, $e(\cdot)$ is the embedding of a latent state and $[\cdot;\cdot]$ denotes vector concatenation. We employ the Gumbel trick (Jang et al., 2017) to handle the discrete and undifferentiable process of sampling c_t.

Decoding. Given the contextualized representation $h_{t,i}$ for token $w_{t,i}$, the original GPT-2 model calculates the pre-softmax logit vector through a linear head, i.e., $p_{t,i} = \mathbf{W}_v h_{t,i}$, where $\mathbf{W}_v$ is a learnable parameter. To explicitly guide the generation through the context-independent latent variable z_t^I and the context-sensitive latent variable z^S, we first project them into the space of $h_{t,i}$ and then calculate two pre-softmax logit vectors similar to $p_{t,i}$:

$$p_t^I = \mathbf{W}_v \mathbf{W}_v^I z_t^I, \quad p^S = \mathbf{W}_v \mathbf{W}_v^S z^S, \tag{8}$$

where $\mathbf{W}_v^I$ and $\mathbf{W}_v^S$ are learnable parameters. We employ the reparameterization trick (Kingma and Welling, 2013) to allow gradient passing through the sampling of z_t^I and z^S. The probability of generating the next token is then defined as:

$$p(w_{t,i+1}|u_{<t}, w_{t,<i+1}, z_t^I, z^S) = \text{Softmax}(p_{t,i} + p_t^I + p^S). \tag{9}$$

The parameterization of the generative model results in the following factorization:

$$\begin{aligned} &p(u_{\le n}, c_{\le n}, z_{\le n}^I, z^S) \\ &= p(z^S) \prod_{t=1}^{n} \left(p(u_t|u_{<t}, z_t^I, z^S) p(z_t^I|u_{<t}, c_t) p(c_t|c_{<t}) \right), \end{aligned} \tag{10}$$

where the probability of generating u_t is formulated as: $p(u_t|u_{<t}, z_t^I, z^S) = \prod_{i=1}^{m} p(w_{t,i}|u_{<t}, w_{t,<i}, z_t^I, z^S)$.

3.3 Inference

For the inference of latent variables, we employ a lightweight transformer that is initialized by the first 6 layers of the GPT-2 model. The vector representation of utterance u_t is denoted as h_t^ϕ and is defined in the same way as Eq.3. We introduce three auxiliary distributions $q_\phi(c|X)$, $q_\phi(z^I|X)$ and $q_\phi(z^S|X)$ which approximate to the posteriors of the discrete latent variable c, the context-independent latent variable z^I and the context-sensitive latent variable z^S respectively.

Since the context-sensitive latent variable z^S captures the holistic information about the whole session, we construct the posterior distribution $q_\phi(z^S|X)$ by summarizing the representations for each utterance:

$$\begin{aligned} q_\phi(z^S|X) &= \mathcal{N}(\hat{\mu}^S, \hat{\sigma}^S \mathbf{I}), \\ \hat{\mu}^S, \hat{\sigma}^S &= f'_{S-mlp}(h^\phi), \end{aligned} \tag{11}$$

where $h^\phi = \frac{1}{n}\sum_{t=1}^{n} h_t^\phi$ and $f'_{S-mlp}(\cdot)$ denotes a MLP network. Similarly, the posterior distribution $q_\phi(z^I|X)$ is defined as:

$$\begin{aligned} q_\phi(z^I|X) &= \prod_{t=1}^{n} q_\phi(z_t^I|X), \\ q_\phi(z_t^I|X) &= \mathcal{N}(\hat{\mu}_t^I, \hat{\sigma}_t^I \mathbf{I}), \\ \hat{\mu}_t^I, \hat{\sigma}_t^I &= f'_{I-mlp}(h_t^\phi), \end{aligned} \tag{12}$$

where h_t^ϕ encodes the contextual information of u_t. The posterior distribution $q_\phi(c|X)$ could be factorized as $\prod_{t=1}^{n} q_\phi(c_t|X)$ where $q_\phi(c_t|X)$ is a Categorical distribution parameterized by $\text{Softmax}(f'_{c-mlp}(h_t^\phi))$. The inference process is depicted in Figure 1.

3.4 Learning

The log-likelihood of the conversation session X is maximized using variational approximation, yielding the evidence lower bound objective

(ELBO) (Hoffman et al., 2013):

$$\begin{aligned}\mathcal{L}_{ELBO} = &\sum_{t=1}^{n} \mathbb{E}_{q_\phi(z^S|X)q_\phi(z_t^I|X)} \log p(u_t|u_{<t}, z_t^I, z^S) \\ &- \sum_{t=1}^{n} D_{\mathrm{KL}}\left(q_\phi(c_t|X) \| p(c_t|c_{<t})\right) \\ &- \sum_{t=1}^{n} \mathbb{E}_{q_\phi(c_t)|X} D_{\mathrm{KL}}\left(q_\phi(z_t^I|X) \| p(z_t^I|u_{<t}, c_t)\right) \\ &- D_{\mathrm{KL}}\left(q_\phi(z^S|X) \| p(z^S)\right),\end{aligned} \tag{13}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ refers to Kullback–Leibler divergence. Detailed derivations are presented in Appendix C.

In addition to optimizing the ELBO objective, we also exploit disentanglement to distill the holistic information into the context-sensitive latent variable while keeping the dynamic information in the context-independent latent variable and the discrete latent variable as demonstrated in the rest of this section.

Holistic Information Discrimination. The latent variable z^S is supposed to only focus on the holistic information that refers to the time-invariant factors (e.g., interlocutor persona) and remains consistent throughout the whole dialogue session. To achieve this, we design a self-supervised task to eliminate the dynamic information from z^S. Ideally, the holistic information of a session X is insensitive to a random shuffle of its internal utterances (i.e., X_{shuf}), but varies across randomly picked different dialogue sessions (i.e., X_{neg}). Thus we could maximize the following objective:

$$\mathcal{L}_{HID} = \log \frac{e^{f_{sim}(z^S, z^S_{shuf}))}}{e^{f_{sim}(z^S, z^S_{shuf})} + e^{f_{sim}(z^S, z^S_{neg})}}, \tag{14}$$

where z^S_{shuf} and z^S_{neg} denote the context-sensitive latent variables of X_{shuf} and X_{neg} respectively, and $f_{sim}(\cdot,\cdot)$ is implemented as the cosine similarity between two vectors.

Dynamic Information Restoration. Since each utterance contains some utterance-specific features that are independent of the conversational context, it is reasonable to encourage the context-independent latent variable z^I to be aware of the dynamic information flow in a session. Specifically, we design a surrogate task to recover the verbs in an utterance u_t given the corresponding context-independent latent variable z_t^I:

$$\mathcal{L}_{DIR} = \sum_{t=1}^{n} \sum_{i=1}^{m} \delta_{t,i} \log p(w_{t,i}|z_t^I), \tag{15}$$

where $\delta_{t,i} = 1$ if the token $w_{t,i}$ is a verb, otherwise $\delta_{t,i} = 0$. $p(w_{t,i}|z_t^I) = \mathrm{Softmax}(\mathbf{W}_{verb} z_t^I)$ outputs a probability distribution over all verbs in the vocabulary with $\mathbf{W}_{verb}$ a learnable parameter.

Mutual Information Minimization. Since the task of Dynamic Information Restoration can not guarantee that the holistic information is exclusive in z^I, we further introduce the mutual information objective as a regularization to minimize the relationship between z^S and z^I:

$$\mathcal{L}_{MIM} = -\sum_{t=1}^{n} [H(z^S) + H(z_t^I) - H(z^S, z_t^I)], \tag{16}$$

where $H(\cdot)$ denotes the entropy which is estimated through minibatch-weighted sampling (Chen et al., 2019; Zhu et al., 2020):

$$\begin{aligned}H(z) &\equiv -\mathbb{E}_{q(z)}[\log q(z)] \\ &\approx -\frac{1}{B}\sum_{i=1}^{B}\left[\log \frac{1}{MB}\sum_{j=1}^{B} q(z(u_t^{(i)})|u_t^{(j)})\right],\end{aligned} \tag{17}$$

for $z = z_t^I$, z^S or (z^S, z_t^I), where $u_t^{(i)}$ denotes u_t in the i-th data point, $z(u_t^{(i)})$ is a sample from $q(z|u_t^{(i)})$, M and B are the data size and minibatch size respectively.

The final learning objective is defined as:

$$\mathcal{L} = \mathcal{L}_{ELBO} + \alpha(\mathcal{L}_{HID} + \mathcal{L}_{DIR} + \mathcal{L}_{MIM}), \tag{18}$$

where α is a hyper-parameter to balance the objective of evidence lower bound and those related to disentanglement.

4 Experimental Setup

4.1 Datasets

We follow Zhang et al. (2020) to adopt the Reddit comments as our pre-training data. We evaluate our model on two benchmark datasets for multi-turn dialogue generation, including DailyDialog (Li et al., 2017) and ConvAI2 (Dinan et al., 2020). More details about all datasets are provided in Appendix A.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Distinct-1	Distinct-2
Zero-resource setting										
DialoGPT	4.29	1.52	0.74	0.41	9.01	2.12	8.63	3.93	6.97	31.79
DialogBERT	6.91	1.22	0.22	0.02	5.54	0.21	4.87	4.35	5.12	30.84
PLATO-2	5.38	1.86	0.85	0.47	8.90	2.08	8.49	3.72	4.88	19.50
DialoFlow	4.76	1.40	0.52	0.20	9.07	1.36	8.52	5.08	5.35	20.23
Ours	**9.58**	**3.40**	**1.64**	**0.90**	**11.99**	**2.50**	**11.35**	**5.31**	5.11	23.86
Full-resource setting										
DialoGPT	18.97	8.67	4.95	3.09	15.97	4.26	14.38	9.81	2.71	15.75
DialogBERT	11.27	2.20	0.62	0.18	4.56	0.23	4.21	4.20	3.91	24.86
PLATO-2	14.10	6.08	3.32	2.04	14.48	3.74	13.49	6.34	2.64	10.31
DialoFlow	17.78	8.51	5.21	3.56	18.14	6.06	16.54	10.14	2.82	18.15
Ours	**24.87**	**16.70**	**13.69**	**12.16**	**22.29**	**11.89**	**21.47**	**11.36**	**4.88**	**28.28**

Table 1: Automatic evaluation results on the test set of DailyDialog. Numbers in bold mean that the improvement to the best-performing baseline is statistically significant (t-test with p-value< 0.05).

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Distinct-1	Distinct-2
Zero-resource setting										
DialoGPT	7.23	3.14	1.64	0.90	12.90	2.79	12.40	4.99	6.26	31.60
DialogBERT	6.53	1.24	0.27	0.09	6.15	0.22	5.48	4.54	5.10	30.75
PLATO-2	7.85	3.38	1.78	1.02	11.04	2.71	10.58	4.73	4.49	18.94
DialoFlow	8.17	3.27	1.43	0.63	12.01	2.54	11.22	5.74	6.92	27.95
Ours	**11.16**	**4.62**	**2.34**	**1.29**	**14.73**	**3.25**	**14.11**	**5.82**	5.24	27.79
Full-resource setting										
DialoGPT	14.92	7.32	4.01	2.34	17.07	4.61	15.53	10.78	1.28	7.33
DialogBERT	12.69	3.11	0.92	0.34	8.20	0.54	7.55	4.66	1.03	5.67
PLATO-2	15.21	7.25	4.11	2.63	13.90	4.07	13.13	6.94	1.56	5.94
DialoFlow	18.57	8.83	4.69	2.70	18.35	4.64	16.95	9.76	1.85	9.96
Ours	**20.15**	**9.41**	**5.17**	**3.09**	**19.19**	**5.04**	**17.38**	10.13	1.69	8.74

Table 2: Automatic evaluation results on the test set of ConvAI2. Numbers in bold mean that the improvement to the best-performing baseline is statistically significant (t-test with p-value < 0.05).

4.2 Evaluation Metrics

Automatic Evaluation. We choose three commonly used reference-based metrics including BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) and ROUGE (Lin, 2004), where BLEU and METEOR are computed with an open source NLG evaluation tool available at `https://github.com/Maluuba/nlg-eval`, and ROUGE is calculated with the code published at `https://github.com/bckim92/language-evaluation`. We report the F1 scores for ROUGE-1, ROUGE-2 and ROUGE-L. We also use Distinct (Li et al., 2015) to evaluate the lexical diversity with Distinct-1 and Distinct-2 denoting ratios of distinct unigrams and bigrams in responses, respectively.

Human Evaluation. We randomly sample 300 examples from the test sets of DailyDialog and ConvAI2 respectively, and recruit 6 well-educated native speakers to do qualitative analysis on the responses generated by our model and all competitive baselines, which are randomly shuffled to hide identification. The annotators judge the quality of the responses from four aspects: (1) *Fluency*: whether the response is fluent without any grammatical errors; (2) *Relevance*: whether the response is coherent with the context; (3) *Informativeness*: whether the response contains informative content; (4) *Engagement*: how much does the annotator like the response. Each annotator assigns a score from $\{0, 1, 2\}$ (representing "bad", "fair" and "good" respectively) to each response for each aspect. Each response obtains four scores for the aforementioned four aspects, and the agreement among all annotators is measured via Fleiss' kappa (Fleiss, 1971).

4.3 Baseline Models

The following models are selected as baselines: (1) **DialoGPT.** A model that is pre-trained on the Reddit comments and attains a performance close to human in single-turn dialogues (Zhang et al., 2020). We adopt the *medium*-sized model which achieves the best performance in the original paper. (2) **DialogBERT.** A model that encodes the dialogue context with a hierarchical transformer architecture (Gu et al., 2021). (3) **PLATO-2.** A model that learns a fine-grained one-to-many generation with the advent of a discrete latent variable (Bao et al., 2021). It is notable that the per-

Models	DailyDialog					ConvAI2				
	Fluency	Relevance	Informativeness	Engagement	Kappa	Fluency	Relevance	Informativeness	Engagement	Kappa
DialoGPT	1.67	1.71	1.69	1.75	0.61	1.73	1.65	1.63	1.79	0.74
DialogBERT	1.43	1.39	1.33	1.37	0.77	1.36	1.35	1.31	1.27	0.75
PLATO-2	1.74	1.68	1.73	1.65	0.70	1.72	1.64	1.72	1.70	0.64
DialoFlow	1.71	1.72	1.71	1.76	0.65	1.79	1.70	1.78	1.81	0.69
Ours	**1.81**	**1.79**	1.74	**1.86**	0.68	1.83	**1.76**	1.80	**1.89**	0.70

Table 3: Human evaluation results on DailyDialog and ConvAI2. Numbers in bold mean that the improvement to the best-performing baseline is statistically significant (t-test with p-value < 0.05).

formance of PLATO-2 is superior to PLATO (Bao et al., 2020) by introducing more parameters and training data. (4) **DialoFlow.** A model that is pre-trained on the Reddit comments and incorporates a dynamic flow mechanism to model the context flow in dialogues (Li et al., 2021). We adopt the *large*-sized model which achieves the best performance in the original paper.

All the baselines are taken from their open-source implementations. We continue to train DialogBERT and PLATO-2 on the Reddit comments for the sake of fairness. The parameter sizes of all baselines are shown in Table 4. We provide more implementation details in Appendix B.

5 Results and Discussion

5.1 Main Results

In this section, we will compare the performance of various models on DailyDialog and ConvAI2, as well as provide some further analyses. We conduct experiments in two different settings, including zero-resource and full-resource, both of which are commonly employed by pre-trained language models. All models solely use the Reddit comments during training in the zero-resource scenario, however, in the full-resource situation, all models are pre-trained on the Reddit comments and then fine-tuned on downstream tasks. Table 1 and Table 2 show the performance of our model on DailyDialog and ConvAI2 respectively. From the results, we can observe that: (1) Although our model is much smaller than the other baseline models, it achieves the best performance on appropriateness-related metrics (i.e., BLEU, ROUGE and METEOR) and performs comparably on distinctness-related metrics (i.e., Distinct) at the same time, demonstrating the effectiveness of a context-free dialogue structure. Additionally, our model takes advantage of z^S and z^I to capture both the time-invariant and time-varying factors and generate a coherent response. (2) DialoFlow outperforms DialoGPT on most metrics after fine-tuning, but not as good as ours. This verifies the necessity of capturing the dialogue flow in PLMs, and the proposed context-free dialogue structure is more competent. (3) On the DailyDialog, our model outperforms baselines by a larger margin than that on the ConvAI2. This is possibly due to the introduction of dialogue act flow in the construction of DailyDialog, which has a similar effect to the dialogue structure.

Human Evaluation. Table 3 shows the results of the human evaluation. While our model achieves comparable performance to the others in terms of *Fluency* and *Informativeness*, it outperforms them on both *Relevance* and *Engagement*, agreeing with the results of automatic evaluation. All kappa values are more than 0.6, indicating that the annotators are in agreement.

Models	Parameter Size	Decoding Speed (ms)			
		DailyDialog		ConvAI2	
		zero	full	zero	full
DialoGPT	345M	34.93	36.72	39.01	39.45
DialogBERT	338M	56.10	104.96	104.66	49.47
PLATO-2	310M	482.02	401.19	479.39	347.88
DialoFlow	941M	94.80	96.10	103.29	82.94
Ours	207M	15.45	17.34	17.13	16.72

Table 4: Evaluation results about decoding speed. "zero" and "full" are abbreviations for zero-resource setting and full-resource setting respectively.

Speed Test. We further compare our model with baselines in terms of decoding speed. Specifically, we calculate the average prediction time per word in response generation in both zero-resource and full-resource settings utilizing all dialogues in the test sets. The experiments are conducted on an RTX 3090. Table 4 shows the speed comparison results. The discrete variable c learns a general transition pattern from the entire corpus, which compensates for the small parameter scale. As a consequence, our model significantly outperforms all competitive baselines thanks to its lightweight architecture.

Models	DailyDialog					ConvAI2				
	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	ROUGE-L
Full Model	9.58	3.40	11.99	2.50	11.35	11.16	4.62	14.73	3.25	14.11
-c	4.55	1.62	8.66	2.08	8.30	7.01	2.65	10.30	2.58	10.13
-z^S	6.35	2.16	10.73	2.53	10.27	7.11	2.88	11.30	2.61	10.80
-z^I	7.27	2.50	11.10	2.51	10.65	10.08	3.95	13.34	2.97	12.72
-disentangle	6.26	2.19	1.08	2.41	9.74	7.78	3.06	12.37	2.77	11.80

Table 5: Ablation study on DailyDialog and ConvAI2.

5.2 Ablation Study

To understand the impact of different variables on model performance and the effect of disentanglement, we compare the full model with the following variants: (1) -c: the discrete latent variable is removed; (2) -z^S: the context-sensitive latent variable is removed; (3) -z^I: the context-independent latent variable is removed; (4) -disentangle: the model is only trained with $\mathcal{L}_{ELBO}$. All models are evaluated under the zero-resource setting to gain a full grasp of the transferability of our model. Table 5 reports the evaluation results. We can conclude that: (1) The discrete latent variable c plays a crucial role in both datasets, as eliminating the variable causes a dramatic performance drop. It is reasonable since our model can capture state transitions between utterances thanks to the latent structure. (2) Though the removal of the context-sensitive or the context-independent variables both results in a performance drop, the context-sensitive latent variable z^S is much more beneficial because it can eliminate context-independent information from the dialogue structure, allowing the model to be more transferable. (3) The self-supervised tasks designed for disentanglement are effective because removing them leads to a decline in performance.

Models	DailyDialog			ConvAI2		
	BLEU-1	BLEU-2	ROUGE-L	BLEU-1	BLEU-2	ROUGE-L
Full Model (w/o freeze)	24.87	16.70	21.47	20.15	9.41	17.38
Full Model (freeze)	25.10	16.18	20.47	20.25	9.57	17.27
-z^S & z^I (w/o freeze)	18.07	9.11	15.98	18.05	8.17	16.28
-z^S & z^I (freeze)	16.60	7.19	13.70	16.61	8.08	16.25

Table 6: Evaluation results about the transferability on DailyDialog and ConvAI2.

5.3 Further Analysis on Transferability

This part will move one step further to understand the transferability of the dialogue structure learned from the large-scale corpus. A dialogue structure with strong transferability is supposed to be well adapted to downstream tasks even without fine-tuning structure-related parameters, which is much more challenging. Therefore, to further verify the transferability of our methods, we freeze the parameters of $f_{c-trans}$ (in Eq.4) and f_{c-mlp} (in Eq.5), and only fine-tune the remaining parameters on downstream tasks. We additionally provide a variant in which the context-sensitive and context-independent latent variables are removed. Table 6 reports the results. It can be seen that our model still performs well once some parameters are frozen, thanks to the proposed disentangled structure learning framework. On the other hand, when context-sensitive and context-independent latent variables are removed, freezing the parameters will result in a significant performance drop. The reason is that our proposed approach can disentangle the holistic information from the dialogue structure, allowing it to be used for various downstream tasks.

Context	
ask questions (0.67); give opinions (0.12); offer service(0.08)	**U1**: Do you know there are several catagories of Chinese tea?
give opinions (0.45); ask questions (0.37); give suggestions (0.06)	**U2**: Yes, I believe there are green teas, black teas and scented teas. Any others?
ask questions (0.71); give an example (0.16); raise a new topic (0.07)	**U3**: Well, have you ever heard of oulong tea and compressed tea?
give opinions (0.51); give suggestions (0.11); explain reasons (0.08)	**U4**: Oh, yeah. Oulong tea is good for one's health. isn't it?
Response	
show approval (0.43); give comments (0.27); give praise (0.21)	**Ground-truth**: You surely know a lot about Chinese tea.
show approval (0.41); ask questions (0.35); describe characteristics (0.09)	**Ours**: Sure, you know much about it. What's it like ?
N/A	**DialoFlow:** I am not sure they are called as qingcha in China.
N/A	**DialoGPT**: Yes, it is. But it's also very expensive.
N/A	**DialogBERT**: The brewing technique is complex.
N/A	**PLATO-2**: I don't know possibly because it contains caffeine.

Figure 2: A case from the Test set of DailyDialog.

5.4 Case Study on Iterpretability

Figure 2 shows an example from the test set of DailyDialog. The left column gives the states of the discrete latent variable along with the probabilities predicted by our model and the right column shows the corresponding utterances. We employ

human experts to consistently interpret each state by going through utterances assigned to the same state. We only present the top-3 states due to space constraints. We can observe that: (1) our model can accurately infer the states of context utterances and the ground-truth response given the posterior distribution of c; (2) thanks to the dialogue structure our model can give a more appropriate response to catch up with the context than baseline models.

6 Conclusion

We propose a novel dialogue model with structural bias which is explainable to humans and easily transferable to general dialogue tasks. Empirical experiments on two benchmark datasets indicate that our model with only 22% parameters outperforms the strongest baseline DialoFlow in both decoding speed and response quality measured by automatic and human evaluations. We further show that the learned latent structure enjoys superior transferability and interpretability compared to the conventional methods.

Limitations

In this paper, we propose a dialogue pre-training model that featured a discrete transition structure. By introducing a series of latent variables into the pre-training process, the pre-trained model could be easily adapted into downstream application scenarios in a transparent and interpretable way. However, all technologies built upon the large-scale PLM more or less inherit their potential harms (Bender et al., 2021). Besides, we identify some limitations within our work and describe them below:

(1) Although the conversation flow is discrete and interpretable, it is laborious to interpret the implication of each state in the conversation flow by going through utterances assigned to the same state. Besides, large-scale human evaluation in our experiments is also costly and time-consuming.

(2) The vocabulary size of the latent conversation flow N is an important parameter that requires to be carefully tuned, especially when the model is agnostic to the downstream task. The optimum N may vary according to the different languages or different domains. It is particularly challenging to shift to uncommon languages because of its reliance on large-scale pre-training corpus.

(3) In this paper, we use discrete latent variables to model the conversation flow for better interpretability. From another perspective, we are distributing utterances into clusters according to their latent flow variable c. However, this may incur high intra-cluster diversity, especially when generalizing to out-of-distribution data. A possible remedy is to introduce some regularity in training f_{c-mlp} to push its predicted distribution towards one-hot distribution.

(4) The adoption of our method can lead to better dialogue systems that improve the quality of life for many people. But our method could also affect the human interlocutors in a negative way if used for malicious intent. We advise that any plan to apply our method should consider carefully all potential groups of stakeholders as well as the risk profiles of applied domains to maximize the overall positive impacts.

Ethics Statement

This paper studies open-domain dialogue pre-training and proposes a disentangled structure learning framework that allows the transformer architecture to capture the prior knowledge about state transition in a large-scale dialogue corpus. There are no ethical issues with this research. The datasets we used are commonly utilized by other researchers and are typically accessible to the public. The proposed approach does not introduce ethical or societal prejudice.

Acknowledgements

We appreciate the anonymous reviewers for their constructive comments. This work was supported by National Natural Science Foundation of China (NSFC Grant No. 62122089 and No. 61876196), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China. This work was also supported in part by Independent Research Fund Denmark under agreement 8048-00038B. This work was also supported by the National Key Research and Development Program of China (No. 2020AAA0106600). We wish to acknowledge the support provided and contribution made by Public Policy and Decision-making Research Lab of RUC. Rui Yan is supported by Beijing Academy of Artificial Intelligence (BAAI) and Tencent Collaborative Research Fund.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525, Online. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. 2019. Isolating sources of disentanglement in vaes. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2615–2625.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021. DialogBERT: Discourse-aware response generation via learning to recover and rank utterances.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5).

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pages 228–231.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *NAACL*, 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 Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 128–138, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Liang Qiu, Yizhou Zhao, Weiyan Shi, Yuan Liang, Feng Shi, Tao Yuan, Zhou Yu, and Song-chun Zhu. 2020. Structured attention for unsupervised dialogue structure induction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1889–1899.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *ACL*, pages 1577–1586.

Weiyan Shi, Tiancheng Zhao, and Zhou Yu. 2019. Unsupervised dialog structure learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1797–1807.

Yajing Sun, Yong Shan, Chengguang Tang, Yue Hu, Yinpei Dai, Jing Yu, Jian Sun, Fei Huang, and Luo Si. 2021. Unsupervised learning of deterministic dialogue structure with edge-enhanced graph autoencoder. In *Proceedings of the Thirty-Fifth Conference on Association for the Advancement of Artificial Intelligence (AAAI)*, pages 13869–13877.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, pages 4418–4424.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

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

Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to ask questions in open-domain conversational systems with typed decoders. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2193–2203.

Chuhan Wu, Fangzhao Wu, Tao Qi, Xiaohui Cui, and Yongfeng Huang. 2020. Attentive pooling with learnable norms for text representation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2961–2970.

Can Xu, Wei Wu, Chongyang Tao, Huang Hu, Matt Schuerman, and Ying Wang. 2019. Neural response generation with meta-words. *arXiv preprint arXiv:1906.06050*.

Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Discovering dialog structure graph for coherent dialog generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1726–1739.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2018. Learning to control the specificity in neural response generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1108–1117.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, pages 654–664.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-resource knowledge-grounded dialogue generation. In *International Conference on Learning Representations*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.

Yizhe Zhu, Martin Renqiang Min, Asim Kadav, and Hans Peter Graf. 2020. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6538–6547.

A Details of Datasets

We follow Zhang et al. (2020) to adopt the Reddit comments as our pre-training data, which contains various domains and topics. We crawl the online discussions over a period spanning from 2011 through 2016, and there are $60,579,645$ and $685,881$ dialogues in the training set and the validation set respectively. Each dialogue has 7.9 utterances on average, with each utterance including 27.5 words.

We evaluate our model on two benchmark datasets for multi-turn dialogue generation. (1) **DailyDialog Dataset.** This dataset is manually labeled and contains conversations about daily life (Li et al., 2017). This dataset is split into training set, validation set, and test set by the data owners. (2) **ConvAI2 Dataset.** This dataset is collected by having two workers at Amazon Mechanical Turk chat with each other based on their assigned profiles (Dinan et al., 2020). The profiles define speakers' personas and provide characteristic knowledge for dialogues. Since the test set of ConvAI2 has not been made public, we randomly select 5% sessions from the original training set as our validation set and use the original validation set as our test set.

To facilitate reproducibility, we adopt the datasets shared at ParlAI[1] and conduct preprocessing with the code available there. More statistics of the two datasets are shown in Table 7.

Statistics	DailyDialog			ConvAI2		
	Train	Val	Test	Train	Val	Test
# Sessions	22,236	2,000	2,000	16,985	893	1,000
# Turns	87,170	8,069	7,740	124,877	6,561	7,801
# Turns / Session	3.9	4.0	3.9	7.4	7.3	7.8
# Words / Turn	13.0	12.9	13.1	11.1	11.0	11.7

Table 7: Statistics of the two datasets.

B More Implementation Details

The total number of discrete latent states (i.e., N) is set as 100 for all experiments. The dimension of context-sensitive and context-independent latent variables are both set as 768. The embedding size of discrete latent states is set as 768. The MLP network in f_{c-mlp} has two layers with the input, hidden and output dimensions being 768, 100 and 100 respectively. We choose GPT-2 (117M) as the backbone of our model. All models are learned with Adam (Kingma and Ba, 2015) optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set the initial temperature, the minimum temperature, and the anneal rate of Gumbel Softmax as 1.0, 0.5, and $4e-5$ respectively. In the training phase, the batch size is set as 64, and the learning rate is set as $2e-5$. In the test phase, we employ beam search in response decoding with beam size $= 5$. Early stopping on validation is adopted as a regularization strategy. In our experiments, the profiles in ConvAI2 are concatenated as a long sequence and serve as the first sentence in a session. In both DailyDialog and ConvAI2, 7 turns before an utterance are used as conversation history. All utterances are padded to a maximum length of 32 tokens.

[1] https://github.com/facebookresearch/ParlAI/tree/main/parlai/tasks

C Derivation of ELBO

$$\mathcal{L}_{ELBO} = \sum_{t=1}^{n} \mathbb{E}_{q(z_t^I, z^S)} \log p(u_t|u_{<t}, z_t^I, z^S) - D_{\mathrm{KL}}(q(c, z^I, z^S)\|p(c, z^I, z^S)). \tag{19}$$

According to the mean-field approximation, $q(c, z^I, z^S) \sim q(c)q(z^I)q(z^S)$. Therefore, the last term can be re-written as:

$$\begin{aligned}
&D_{\mathrm{KL}}(q(c, z^I, z^S)\|p(c, z^I, z^S)) \\
&= \sum q(c|X) \int q(z^I|X)q(z^s|X) \log \frac{q(c|X)q(z^I|X)q(z^s|X)}{p(c)p(z^I)p(z^s)} \Delta z^I \Delta z^s \\
&= \sum q(c|X) \log \frac{q(c|X)}{p(c)} + \sum q(c|X) \int q(z^I|X) \log \frac{q(z^I|X)}{p(z^I)} \Delta z^I \\
&+ \int q(z^S|X) \log \frac{q(z^S|X)}{p(z^S)} \Delta z^S \\
&= \sum_{t=1}^{n} D_{KL}(q(c_t|X)\|p(c_t|c_{<t})) + D_{KL}(q(z^s|X)\|p(z^s)) \\
&+ \sum_{t=1}^{n} \mathbb{E}_{q(c_t)} D_{KL}(q(z_t^I|X)\|p(z_t^I|u_{<t}, c_t))
\end{aligned} \tag{20}$$