A Discrete CVAE for Response Generation on Short-Text Conversation

Jun Gao[1]*, Wei Bi[2]†, Xiaojiang Liu[2], Junhui Li[1], Guodong Zhou[1], Shuming Shi[2]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
imgaojun@gmail.com, {lijunhui,gdzhou}@suda.edu.cn
[2]Tencent AI Lab, Shenzhen, China
{victoriabi, kieranliu, shumingshi}@tencent.com

Abstract

Neural conversation models such as encoder-decoder models are easy to generate bland and generic responses. Some researchers propose to use the conditional variational autoencoder (CVAE) which maximizes the lower bound on the conditional log-likelihood on a continuous latent variable. With different sampled latent variables, the model is expected to generate diverse responses. Although the CVAE-based models have shown tremendous potential, their improvement of generating high-quality responses is still unsatisfactory. In this paper, we introduce a discrete latent variable with an explicit semantic meaning to improve the CVAE on short-text conversation. A major advantage of our model is that we can exploit the semantic distance between the latent variables to maintain good diversity between the sampled latent variables. Accordingly, we propose a two-stage sampling approach to enable efficient diverse variable selection from a large latent space assumed in the short-text conversation task. Experimental results indicate that our model outperforms various kinds of generation models under both automatic and human evaluations and generates more diverse and informative responses.

1 Introduction

Open-domain response generation (Perez-Marin, 2011; Sordoni et al., 2015) for single-round short text conversation (Shang et al., 2015), aims at generating a meaningful and interesting response given a query from human users. Neural generation models are of growing interest in this topic due to their potential to leverage massive conversational datasets on the web. These generation models such as encoder-decoder models (Vinyals and Le, 2015; Shang et al., 2015; Wen et al., 2015), directly build a mapping from the input query to its output response, which treats all query-response pairs uniformly and optimizes the maximum likelihood estimation (MLE). However, when the models converge, they tend to output bland and generic responses (Li et al., 2016a,c; Serban et al., 2016).

Many enhanced encoder-decoder approaches have been proposed to improve the quality of generated responses. They can be broadly classified into two categories (see Section 2 for details): (1) One that does not change the encoder-decoder framework itself. These approaches only change the decoding strategy, such as encouraging diverse tokens to be selected in beam search (Li et al., 2016a,b); or adding more components based on the encoder-decoder framework, such as the Generative Adversarial Network (GAN)-based methods (Xu et al., 2017; Zhang et al., 2018; Li et al., 2017) which add discriminators to perform adversarial training; (2) The second category modifies the encoder-decoder framework directly by incorporating useful information as latent variables in order to generate more specific responses (Yao et al., 2017; Zhou et al., 2017). However, all these enhanced methods still optimize the MLE of the log-likelihood or the complete log-likelihood conditioned on their assumed latent information, and models estimated by the MLE naturally favor to output frequent patterns in training data.

Instead of optimizing the MLE, some researchers propose to use the conditional variational autoencoder (CVAE), which maximizes the lower bound on the conditional data log-likelihood on a continuous latent variable (Zhao et al., 2017; Shen et al., 2017). Open-domain response generation is a one-to-many problem, in which a query can be associated with many valid responses. The CVAE-based models generally assume the latent

*Work done when Jun Gao was interning at Tencent AI Lab.
†Corresponding author

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1898–1908, Hong Kong, China, November 3–7, 2019.

variable follows a multivariate Gaussian distribution with a diagonal covariance matrix, which can capture the latent distribution over all valid responses. With different sampled latent variables, the model is expected to decode diverse responses. Due to the advantage of the CVAE in modeling the response generation process, we focus on improving the performance of the CVAE-based response generation models.

Although the CVAE has achieved impressive results on many generation problems (Yan et al., 2016; Sohn et al., 2015), recent results on response generation show that the CVAE-based generation models still suffer from the low output diversity problem. That is multiple sampled latent variables result in responses with similar semantic meanings. To address this problem, extra guided signals are often used to improve the basic CVAE. Zhao et al. (2017) use dialogue acts to capture the discourse variations in multi-round dialogues as guided knowledge. However, such discourse information can hardly be extracted for short-text conversation.

In our work, we propose a discrete CVAE (DCVAE), which utilizes a discrete latent variable with an explicit semantic meaning in the CVAE for short-text conversation. Our model mitigates the low output diversity problem in the CVAE by exploiting the semantic distance between the latent variables to maintain good diversity between the sampled latent variables. Accordingly, we propose a two-stage sampling approach to enable efficient selection of diverse variables from a large latent space assumed in the short-text conversation task.

To summarize, this work makes three contributions: (1) We propose a response generation model for short-text conversation based on a DCVAE, which utilizes a discrete latent variable with an explicit semantic meaning and could generate high-quality responses. (2) A two-stage sampling approach is devised to enable efficient selection of diverse variables from a large latent space assumed in the short-text conversation task. (3) Experimental results show that the proposed DCVAE with the two-stage sampling approach outperforms various kinds of generation models under both automatic and human evaluations, and generates more high-quality responses. All our code and datasets are available at https://ai.tencent.com/ailab/nlp/dialogue.

2 Related Work

In this section, we briefly review recent advancement in encoder-decoder models and CVAE-based models for response generation.

2.1 Encoder-decoder models

Encoder-decoder models for short-text conversation (Vinyals and Le, 2015; Shang et al., 2015) maximize the likelihood of responses given queries. During testing, a decoder sequentially generates a response using search strategies such as beam search. However, these models frequently generate bland and generic responses.

Some early work improves the quality of generated responses by modifying the decoding strategy. For example, Li et al. (2016a) propose to use the maximum mutual information (MMI) to penalize general responses in beam search during testing. Some later studies alter the data distributions according to different sample weighting schemes, encouraging the model to put more emphasis on learning samples with rare words (Nakamura et al., 2018; Liu et al., 2018). As can be seen, these methods focus on either pre-processing the dataset before training or post-processing the results in testing, with no change to encoder-decoder models themselves.

Some other work use encoder-decoder models as the basis and add more components to refine the response generation process. Xu et al. (2017) present a GAN-based model with an approximate embedding layer. Zhang et al. (2018) employ an adversarial learning method to directly optimize the lower bounder of the MMI objective (Li et al., 2016a) in model training. These models employ the encoder-decoder models as the generator and focus on how to design the discriminator and optimize the generator and discriminator jointly. Deep reinforcement learning is also applied to model future reward in chatbot after an encoder-decoder model converges (Li et al., 2016c, 2017). The above methods directly integrate the encoder-decoder models as one of their model modules and still do not actually modify the encoder-decoder models.

Many attentions have turned to incorporate useful information as latent variables in the encoder-decoder framework to improve the quality of generated responses. Yao et al. (2017) consider that a response is generated by a query and a precomputed cue word jointly. Zhou et al. (2017) uti-

lize a set of latent embeddings to model diverse responding mechanisms. Xing et al. (2017) introduce pre-defined topics from an external corpus to augment the information used in response generation. Gao et al. (2019) propose a model that infers latent words to generate multiple responses. These studies indicate that many factors in conversation are useful to model the variation of a generated response, but it is nontrivial to extract all of them. Also, these methods still optimize the MLE of the complete log-likelihood conditioned on their assumed latent information, and the model optimized with the MLE naturally favors to output frequent patterns in the training data. Note that we apply a similar latent space assumption as used in (Yao et al., 2017; Gao et al., 2019), i.e. the latent variables are words from the vocabulary. However, they use a latent word in a factorized encoder-decoder model, but our model uses it to construct a discrete CVAE and our optimization algorithm is entirely different from theirs.

2.2 The CVAE-based models

A few works indicate that it is worth trying to apply the CVAE to dialogue generation which is originally used in image generation (Yan et al., 2016; Sohn et al., 2015) and optimized with the variational lower bound of the conditional log-likelihood. For task-oriented dialogues, Wen et al. (2017) use the latent variable to model intentions in the framework of neural variational inference. For chit-chat multi-round conversations, Serban et al. (2017) model the generative process with multiple levels of variability based on a hierarchical sequence-to-sequence model with a continuous high-dimensional latent variable. Zhao et al. (2017) make use of the CVAE and the latent variable is used to capture discourse-level variations. Gu et al. (2019) propose to induce the latent variables by transforming context-dependent Gaussian noise. Shen et al. (2017) present a conditional variational framework for generating specific responses based on specific attributes. Yet, it is observed in other tasks such as image captioning (Wang et al., 2017) and question generation (Fan et al., 2018) that the CVAE-based generation models suffer from the low output diversity problem, i.e. multiple sampled variables point to the same generated sequences. In this work, we utilize a discrete latent variable with an interpretable meaning to alleviate this low output diversity problem on short-text conversation.

We find that Zhao et al. (2018) make use of a set of discrete variables that define high-level attributes of a response. Although they interpret meanings of the learned discrete latent variables by clustering data according to certain classes (e.g. dialog acts), such latent variables still have no exact meanings. In our model, we connect each latent variable with a word in the vocabulary, thus each latent variable has an exact semantic meaning. Besides, they focus on multi-turn dialogue generation and presented an unsupervised discrete sentence representation learning method learned from the context while our concentration is primarily on single-turn dialogue generation with no context information.

3 Proposed Models

3.1 DCVAE and Basic Network Modules

Following previous CVAE-based generation models (Zhao et al., 2017), we introduce a latent variable z for each input sequence and our goal is to maximize the lower bound on the conditional data log-likelihood $p(\mathbf{y}|\mathbf{x})$, where $\mathbf{x}$ is the input query sequence and $\mathbf{y}$ is the target response sequence:

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{x}) \geq\ & \mathbb{E}_{z\sim q(z|\mathbf{y},\mathbf{x})}[\log p(\mathbf{y}|\mathbf{x},z)] \\ & -D_{KL}(q(z|\mathbf{y},\mathbf{x})||p(z|\mathbf{x})). \quad (1)\end{aligned}$$

Here, $p(z|\mathbf{x})$/$q(z|\mathbf{y},\mathbf{x})$/$p(\mathbf{y}|\mathbf{x},z)$ is parameterized by the prior/posterior/generation network respectively. $D_{KL}(q(z|\mathbf{y},\mathbf{x})||p(z|\mathbf{x}))$ is the Kullback-Leibler (KL) divergence between the posterior and prior distribution. Generally, z is set to follow a Gaussian distribution in both the prior and posterior networks. As mentioned in the related work, directly using the above CVAE formulation causes the low output diversity problem. This observation is also validated in the short-text conversation task in our experiments.

Now, we introduce our basic discrete CVAE formulation to alleviate the low output diversity problem. We change the continuous latent variable z to a discrete latent one with an explicit interpretable meaning, which could actively control the generation of the response. An intuitive way is to connect each latent variable with a word in the vocabulary. With a sampled latent z from the prior (in testing)/posterior network (in training), the generation network will take the query representation together with the word embedding of this latent

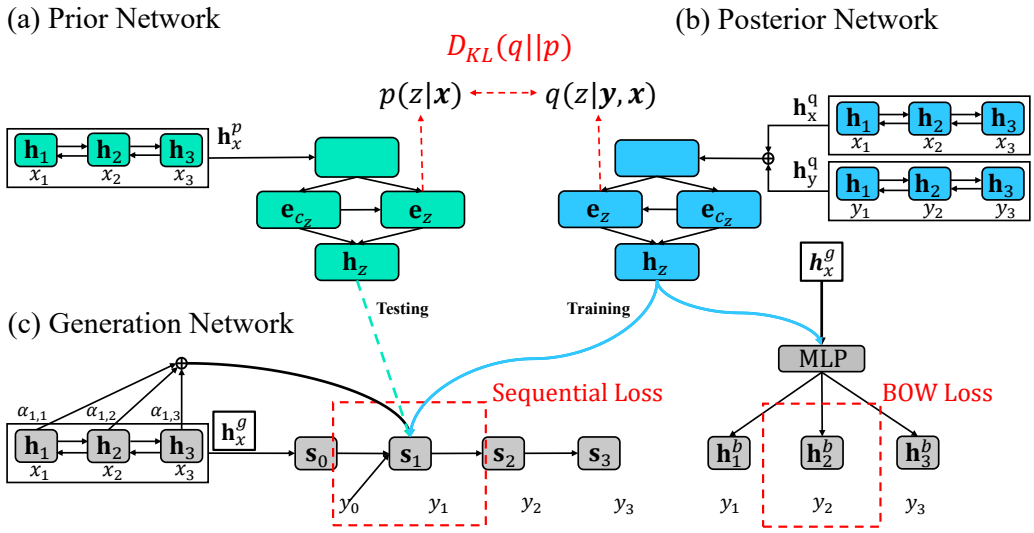


Figure 1: The architecture of the proposed discrete CVAE. $\mathbf{e}_{c_z}$ and $\mathbf{e}_z$ are embeddings of a cluster and a word sampled from the estimated discrete distributions. $\mathbf{e}_{c_z}$ is only applied when the two-stage sampling approach in Section 3.2 is used. If $\mathbf{e}_{c_z}$ is applied, the latent representation $\mathbf{h}_z$ is the sum of $\mathbf{e}_{c_z}$ and $\mathbf{e}_z$; otherwise, $\mathbf{h}_z$ is $\mathbf{e}_z$. α denotes the attention weight. $\oplus$ denotes the sum of input vectors.

variable as the input to decode the response. Here, we assume that a single word is enough to drive the generation network to output diverse responses for short text conversation, in which the response is generally short and compact.

A major advantage of our DCVAE is that for words with far different meanings, their word embeddings (especially that we use a good pre-trained word embedding corpus) generally have a large distance and drive the generation network to decode scattered responses, thus improve the output diversity. In the standard CVAE, z's assumed in a continuous space may not maintain the semantic distance as in the embedding space and diverse z's may point to the same semantic meaning, in which case the generation network is hard to train well with such confusing information. Moreover, we can make use of the semantic distance between latent variables to perform better sampling to approximate the objective during optimization, which will be introduced in Section 3.2.

The latent variable z is thus set to follow a categorical distribution with each dimension corresponding to a word in the vocabulary. Therefore the prior and posterior networks should output categorical probability distributions:

$$p_\theta(z|\mathbf{x}) = \text{softmax}(g_\theta(\mathbf{x})), \quad (2)$$

$$q_\phi(z|\mathbf{y},\mathbf{x}) = \text{softmax}(f_\phi(\mathbf{y},\mathbf{x})), \quad (3)$$

where θ and ϕ are parameters of the two networks respectively. The KL distance of these two distributions can be calculated in a closed form solution:

$$D_{KL}(q(z|\mathbf{y},\mathbf{x})||p(z|\mathbf{x})) = \sum_{z\in Z} q(z|\mathbf{y},\mathbf{x}) \log \frac{q(z|\mathbf{y},\mathbf{x})}{p(z|\mathbf{x})}, \quad (4)$$

where Z contains all words in the vocabulary. In the following, we present the details of the prior, posterior and generation network.

Prior network $p(z|\mathbf{x})$: It aims at inferring the latent variable z given the input sequence x. We first obtain an input representation $\mathbf{h}_\mathbf{x}^p$ by encoding the input query $\mathbf{x}$ with a bi-directional GRU and then compute $g_\theta(\mathbf{x})$ in Eq. 2 as follows:

$$g_\theta(\mathbf{x}) = \mathbf{W}_2 \cdot \tanh(\mathbf{W}_1 \mathbf{h}_x^p + \mathbf{b}_1) + \mathbf{b}_2, \quad (5)$$

where θ contains parameters in both the bidirectional GRU and Eq. 5.

Posterior network $q(z|\mathbf{y},\mathbf{x})$: It infers a latent variable z given a input query $\mathbf{x}$ and its target response $\mathbf{y}$. We construct both representations for the input and the target sequence by separated bi-directional GRU's, then add them up to compute $f_\phi(\mathbf{y},\mathbf{x})$ in Eq. 3 to predict the probability of z:

$$f_\phi(\mathbf{y},\mathbf{x}) = \mathbf{W}_4 \cdot \tanh(\mathbf{W}_3(\mathbf{h}_x^q + \mathbf{h}_y^q) + \mathbf{b}_3) + \mathbf{b}_4, \quad (6)$$

where ϕ contains parameters in the two encoding functions and Eq. 6. Note that the parameters of

the encoding functions are not shared in the prior and posterior network.

Generation network $p(\mathbf{y}|\mathbf{x}, z)$: We adopt an encoder-decoder model with attention (Luong et al., 2015) used in the decoder. With a sampled latent variable z, a typical strategy is to combine its representation, which in this case is the word embedding $\mathbf{e}_z$ of z, only in the beginning of decoding. However, many previous works observe that the influence of the added information will vanish over time (Yao et al., 2017; Gao et al., 2019). Thus, after obtaining an attentional hidden state at each decoding step, we concatenate the representation $\mathbf{h}_z$ of the latent variable and the current hidden state to produce a final output in our generation network.

3.2 A Two-Stage Sampling Approach

When the CVAE models are optimized, they tend to converge to a solution with a vanishingly small KL term, thus failing to encode meaningful information in z. To address this problem, we follow the idea in (Zhao et al., 2017), which introduces an auxiliary loss that requires the decoder in the generation network to predict the bag-of-words in the response $\mathbf{y}$. Specifically, the response $\mathbf{y}$ is now represented by two sequences simultaneously: $\mathbf{y}_o$ with word order and $\mathbf{y}_{bow}$ without order. These two sequences are assumed to be conditionally independent given z and $\mathbf{x}$. Then our training objective can be rewritten as:

$$\begin{aligned} J(\Theta) =& \mathbb{E}_{z\sim q(z|\mathbf{y},\mathbf{x})}[\log p(\mathbf{y}|\mathbf{x}, z)] \\ & - D_{KL}(q(z|\mathbf{y}, \mathbf{x})||p(z|\mathbf{x})) \\ & + \mathbb{E}_{z\sim q(z|\mathbf{y},\mathbf{x})}[\log p(\mathbf{y}_{bow}|\mathbf{x}, z)], \end{aligned} \tag{7}$$

where $p(\mathbf{y}_{bow}|\mathbf{x}, z)$ is obtained by a multilayer perceptron $\mathbf{h}^b = \text{MLP}(\mathbf{x}, z)$:

$$p(\mathbf{y}_{bow}|\mathbf{x}, z) = \prod_{t=1}^{|\mathbf{y}|} \frac{\exp(h^b_{y_t})}{\sum_{j\in V}\exp(h^b_j)}, \tag{8}$$

where $|\mathbf{y}|$ is the length of $\mathbf{y}$, y_t is the word index of t-th word in $\mathbf{y}$, and V is the vocabulary size.

During training, we generally approximate $\mathbb{E}_{z\sim q(z|\mathbf{y},\mathbf{x})}[\log p(\mathbf{y}|\mathbf{x}, z)]$ by sampling N times of z from the distribution $q(z|\mathbf{y}, \mathbf{x})$. In our model, the latent space is discrete but generally large since we set it as the vocabulary in the dataset [1]. The vocabulary consists of words that are similar in syntactic or semantic. Directly sampling z from the categorical distribution in Eq. 3 cannot make use of such word similarity information.

Hence, we propose to modify our model in Section 3.1 to consider the word similarity for sampling multiple accurate and diverse latent z's. We first cluster $z \in Z$ into K clusters $c_1, \ldots, c_K$. Each z belongs to only one of the K clusters and dissimilar words lie in distinctive groups. We use the K-means clustering algorithm to group z's using a pre-trained embedding corpus (Song et al., 2018). Then we revise the posterior network to perform a two-stage cluster sampling by decomposing $q(z|\mathbf{y}, \mathbf{x})$ as :

$$\begin{aligned} q(z|\mathbf{y}, \mathbf{x}) &= \sum_k q(z|\mathbf{x}, \mathbf{y}, c_k)q(c_k|\mathbf{x}, \mathbf{y}) \\ &= q(z|\mathbf{x}, \mathbf{y}, c_{k_z})q(c_{k_z}|\mathbf{x}, \mathbf{y}). \end{aligned} \tag{9}$$

That is, we first compute $q(c_{k_z}|\mathbf{y}, \mathbf{x})$, which is the probability of the cluster that z belongs to conditioned on both $\mathbf{x}$ and $\mathbf{y}$. Next, we compute $q(z|\mathbf{x}, \mathbf{y}, c_{k_z})$, which is the probability distribution of z conditioned on the $\mathbf{x}$, $\mathbf{y}$ and the cluster c_{k_z}. When we perform sampling from $q(z|\mathbf{x}, \mathbf{y})$, we can exploit the following two-stage sampling approach: first sample the cluster based on $q(c_k|\mathbf{x}, \mathbf{y})$; next sample a specific z from z's within the sampled cluster based on $q(z|\mathbf{x}, \mathbf{y}, c_{k_z})$.

Similarly, we can decompose the prior distribution $p(z|\mathbf{x})$ accordingly for consistency:

$$p(z|\mathbf{x}) = p(z|\mathbf{x}, c_{k_z})p(c_{k_z}|\mathbf{x}). \tag{10}$$

In testing, we can perform the two-stage sampling according to $p(c_k|\mathbf{x})$ and $p(z|\mathbf{x}, c_{k_z})$. Our full model is illustrated in Figure 1.

Network structure modification: To modify the network structure for the two-stage sampling method, we first compute the probability of each cluster given $\mathbf{x}$ in the prior network (or $\mathbf{x}$ and $\mathbf{y}$ in the posterior network) with a softmax layer (Eq. 5 or Eq. 6 followed by a softmax function). We then add the input representation and the cluster embedding $\mathbf{e}_{c_z}$ of a sampled cluster c_z, and use another softmax layer to compute the probability of each z within the sampled cluster. In the generation network, the representation of z is the sum of the cluster embedding $\mathbf{e}_{c_z}$ and its word embedding $\mathbf{e}_z$.

Network pre-training: To speed up the convergence of our model, we pre-extract keywords from

[1] Note that we remove special tokens including UNK (unknown token), BOS (start of sentence) and EOS (end of sentence) in the latent space such that our model will only select meaningful words as the latent variables.

each query using the TF-IDF method. Then we use these keywords to pre-train the prior and posterior networks. The generation network is not pre-trained because in practice it converges fast in only a few epochs.

4 Experimental Settings

Next, we describe our experimental settings including the dataset, implementation details, all compared methods, and the evaluation metrics.

4.1 Dataset

We conduct our experiments on a short-text conversation benchmark dataset (Shang et al., 2015) which contains about 4 million post-response pairs from the Sina Weibo [2], a Chinese social platforms. We employ the Jieba Chinese word segmenter [3] to tokenize the queries and responses into sequences of Chinese words. We use a vocabulary of 50,000 words (a mixture of Chinese words and characters), which covers 99.98% of words in the dataset. All other words are replaced with <UNK>. We randomly hold out two subsets as the development and test dataset, each containing 900 pairs.

4.2 Implementation Details

We use single-layer bi-directional GRU for the encoder in the prior/posterior/generation network, and one-layer GRU for the decoder in the generation network. The dimension of all hidden vectors is 1024. The cluster embedding dimension is 620. Except that the word embeddings are initialized by the word embedding corpus (Song et al., 2018), all other parameters are initialized by sampling from a uniform distribution $[-0.1, 0.1]$. The batch size is 128. We use Adam optimizer with a learning rate of 0.0001. For the number of clusters K in our method, we evaluate four different values $(5, 10, 100, 1000)$ using automatic metrics and set K to 10 which tops the four options empirically. It takes about one day for every two epochs of our model on a Tesla P40 GPU, and we train ten epochs in total. During testing, we use beam search with a beam size of 10.

4.3 Compared Methods

In our work, we focus on comparing various methods that model $p(\mathbf{y}|\mathbf{x})$ differently. We compare our proposed discrete CVAE (**DCVAE**) with the two-stage sampling approach to three categories of response generation models:

1. Baselines: **Seq2seq**, the basic encoder-decoder model with soft attention mechanism (Bahdanau et al., 2015) used in decoding and beam search used in testing; **MMI-bidi** (Li et al., 2016a), which uses the MMI to re-rank results from beam search.

2. **CVAE** (Zhao et al., 2017): We adjust the original work which is for multi-round conversation for our single-round setting. For a fair comparison, we utilize the same keywords used in our network pre-training as the knowledge-guided features in this model.

3. Other enhanced encoder-decoder models: Hierarchical Gated Fusion Unit (**HGFU**) (Yao et al., 2017), which incorporates a cue word extracted using pointwise mutual information (PMI) into the decoder to generate meaningful responses; Mechanism-Aware Neural Machine (**MANM**) (Zhou et al., 2017), which introduces latent embeddings to allow for multiple diverse response generation.

Here, we do not compare RL/GAN-based methods because all our compared methods can replace their objectives with the use of reward functions in the RL-based methods or add a discriminator in the GAN-based methods to further improve the overall performance. However, these are not the contribution of our work, which we leave to future work to discuss the usefulness of our model as well as other enhanced generation models combined with the RL/GAN-based methods.

4.4 Evaluation

To evaluate the responses generated by all compared methods, we compute the following automatic metrics on our test set:

1. BLEU: BLEU-n measures the average n-gram precision on a set of reference responses. We report BLEU-n with n=1,2,3,4.

2. Distinct-1 & distinct-2 (Li et al., 2016a): We count the numbers of distinct uni-grams and bi-grams in the generated responses and divide the numbers by the total number of generated uni-grams and bi-grams in the test set. These metrics can be regarded as an automatic metric to evaluate the diversity of the responses.

[2] http://weibo.com
[3] https://github.com/fxsjy/jieba

Method	Automatic Metrics						Human Evaluation		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	dist-1	dist-2	quality	accept	good
Seq2seq	0.63 ± 0.14	0.49 ± 0.16	0.35 ± 0.14	0.21 ± 0.12	0.03	0.08	1.64 ± 0.30	49%	15%
MMI-bidi	0.54 ± 0.17	0.39 ± 0.18	0.28 ± 0.15	0.17 ± 0.13	0.04	0.11	1.71 ± 0.34	52%	19%
CVAE	0.60 ± 0.13	0.43 ± 0.16	0.30 ± 0.14	0.18 ± 0.11	0.03	0.06	1.60 ± 0.33	47%	13%
MANM	0.62 ± 0.14	0.48 ± 0.15	0.34 ± 0.14	0.22 ± 0.12	0.05	0.14	1.73 ± 0.35	53%	21%
HGFU	0.52 ± 0.11	0.38 ± 0.14	0.27 ± 0.12	0.16 ± 0.11	0.08	0.27	1.63 ± 0.39	51%	12%
DCVAE	$\mathbf{0.64 \pm 0.14}$	$\mathbf{0.49 \pm 0.16}$	$\mathbf{0.35 \pm 0.15}$	$\mathbf{0.22 \pm 0.13}$	0.08	0.24	$\mathbf{2.03 \pm 0.34}$	**73%**	**30%**

Table 1: The automatic and human evaluation results of all compared methods. Note that the acceptable ratio is the percentage of responses with 2 or 3 points.

Three annotators from a commercial annotation company are recruited to conduct our human evaluation. Responses from different models are shuffled for labeling. 300 test queries are randomly selected out, and annotators are asked to independently score the results of these queries with different points in terms of their quality: (1) Good (3 points): The response is grammatical, semantically relevant to the query, and more importantly informative and interesting; (2) Acceptable (2 points): The response is grammatical, semantically relevant to the query, but too trivial or generic (e.g.,"我不知道(I don't know)", "我也是(Me too)", "我喜欢(I like it)" etc.); (3) Failed (1 point): The response has grammar mistakes or irrelevant to the query.

5 Experimental Results and Analysis

In the following, we will present results of all compared methods and conduct a case study on such results. Then, we will perform further analysis of our proposed method by varying different settings of the components designed in our model.

5.1 Results on All Compared Methods

Results on automatic metrics are shown on the left-hand side of Table 1. From the results we can see that our proposed DCVAE achieves the best BLEU scores and the second best distinct ratios. The HGFU has the best dist-2 ratio, but its BLEU scores are the worst. These results indicate that the responses generated by the HGFU are less close to the ground true references. Although the automatic evaluation generally indicates the quality of generated responses, it can not accurately evaluate the generated response and the automatic metrics may not be consistent with human perceptions (Liu et al., 2016). Thus, we consider human evaluation results more reliable.

For the human evaluation results on the right-hand side of Table 1, we show the mean and standard deviation of all test results as well as the percentage of acceptable responses (2 or 3 points) and good responses (3 points only). Our proposed DCVAE has the best quality score among all compared methods. Moreover, DCVAE achieves a much higher good ratio, which means it generates more informative and interesting responses. Besides, the HGFU's acceptable and good ratios are much lower than our model indicating that it may not maintain enough response relevance when encouraging diversity. This is consistent with the results of the automatic evaluation in Table 1. We also notice that the CVAE achieves the worst human annotation score. This validates that the original CVAE for open-domain response generation does not work well and our proposed DCVAE is an effective way to improve the CVAE for better output diversity.

5.2 Case Study

Figure 2 shows four example queries with their responses generated by all compared methods. The Seq2seq baseline tends to generate less informative responses. Though MMI-bidi can select different words to be used, its generated responses are still far from informative. MANM can avoid generating generic responses in most cases, but sometimes its generated response is irrelevant to the query, as shown in the left bottom case. Moreover, the latent responding mechanisms in MANM have no explicit or interpretable meaning. Similar results can be observed from HGFU. If the PMI selects irrelevant cue words, the resulting response may not be relevant. Meanwhile, responses generated by our DCVAE are more informative as well as relevant to input queries.

5.3 Different Sizes of the Latent Space

We vary the size of the latent space (i.e., sampled word space Z) used in our proposed DCVAE. Fig-

Input	刚和卫哲通了电话，他心态很平和， 说现在有时间找他打球了。 I just talked to Wei Zhe on the phone and he was calm. I said now I had time to play ball with him.	房子，小院，车子，落英缤纷的门前， 静谧的生活。 The house, the courtyard, the car, the door of fallen petals, quiet life.
Seq2seq	是的，我也是这样。 Yes, me too.	这是在哪啊？在哪里？ Where is this? Where?
MMI-bidi	恭喜恭喜！恭喜恭喜！ Congratulations! Congratulations!	哇，好漂亮，好想去 Wow, it's beautiful, I want to go there
MANM	我也想知道他是谁 I wonder who he is, too	好喜欢，好喜欢。 I like it, I like it.
HGFU	我也想知道他是谁[据说] I wonder who he is, too [reputedly]	好美的地方，好美的地方[学习] What a beautiful place, what a beautiful place [study]
DCVAE	哈哈，我也喜欢看你打球[打球] Haha, I like watching you play ball, too [play ball]	好漂亮的房子，好喜欢[生活] What a beautiful house, I love it [life]
Input	这位雷神长得有点眼熟 This god of thunder looks a little familiar	下午媒体公开课！现在睡了…晚上见 I have an open media class this afternoon! Now I'm going to bed...See you tonight
Seq2seq	哈哈哈哈哈，笑死我了！ Hahahahaha, you are killing me!	辛苦了，注意身体！ You've been working hard, take care of your health!
MMI-bidi	笑死我了 You are killing me	辛苦了，好好休息！ You've been working hard, have a good rest!
MANM	这是要闹哪样啊。 What is this for.	辛苦了，注意身体！ You've been working hard, take care of your health!
HGFU	尼玛，这是要逆天啊！[哇塞] Holy crap, you are going against the world！[wow]	加油加油加油加油加油[尼玛] Come on, come on, come on [holy crap]
DCVAE	雷神长的好漂亮啊。[这位] The god of thunder is so beautiful. [this]	我也想睡了，明天还要上班[下午] I want to sleep too, I need to work tomorrow [afternoon]

Figure 2: Examples of the generated responses. The sampled latent words (z) are showed in the brackets.

ure 3 shows the automatic and human evaluation results on the latent space setting to the top 10k, 20k, all words in the vocabulary. On the automatic evaluation results, if the sampled latent space is getting larger, the BLEU-4 score increases but the distinct ratios drop. We find out that though the DCVAE with a small latent space has a higher distinct-1/2 ratio, many generated sentences are grammatically incorrect. This is also why the BLEU-4 score decreases. On the human evaluation results, all metrics improve with the use of a larger latent space. This is consistent with our motivation that open-domain short-text conversation covers a wide range of topics and areas, and the top frequent words are not enough to capture the content of most training pairs. Thus a small latent space, i.e. the top frequent words only, is not feasible to model enough latent information and a large latent space is generally favored in our proposed model.

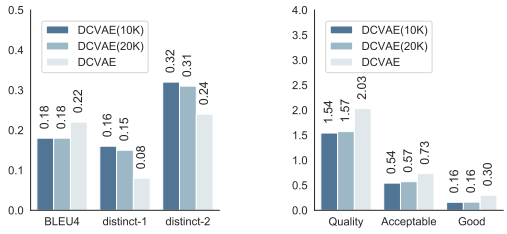
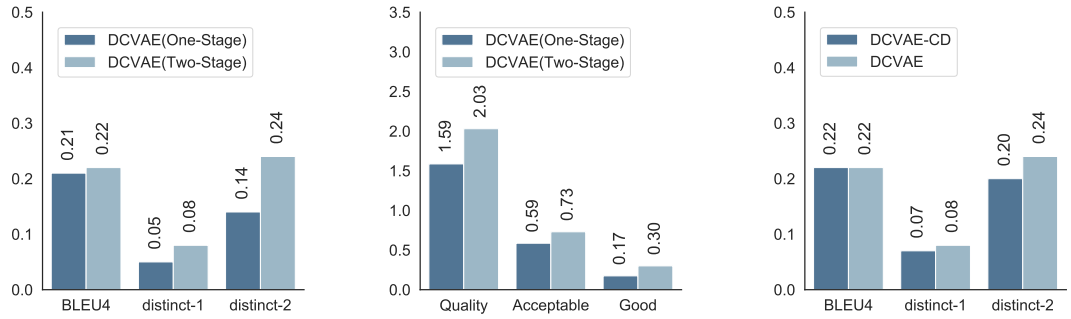


Figure 3: Different sizes of the latent space used in the DCVAE: automatic evaluation (left) and human evaluation (right).

5.4 Analysis on the Two-Stage Sampling

We further look into whether the two-stage sampling method is effective in the proposed DCVAE. Here, the One-Stage method corresponds to the basic formulation in Section 3.1 with no use of the clustering information in the prior or posterior network. Results on both automatic and human evaluation metrics are shown in Figure. 4(a) and 4(b). We can observe that the performance of the DCVAE without the two-stage sampling method drops drastically. This means that the proposed two-stage sampling method is important for the DCVAE to work well.

Besides, to validate the effectiveness of clustering, we implemented a modified DCVAE (DCVAE-CD) that uses a pure categorical distribution in which each variable has no exact meaning. That is, the embedding of each latent variable does not correspond to any word embedding. Automatic evaluation results of this modified model are shown in Figure. 4(c). We can see that DCVAE-CD performs worse, which means the distribution on word vocabulary is important in our model.



(a) (b) (c)

Figure 4: (a)/(b): Automatic/human evaluation on the DCVAE with/without the two-stage sampling approach. (c): Automatic evaluation on our proposed DCVAE and the modified DCVAE that uses a pure categorical distribution (DCVAE-CD) in which each variable has no exact meaning.

6 Conclusion

In this paper, we have presented a novel response generation model for short-text conversation via a discrete CVAE. We replace the continuous latent variable in the standard CVAE by an interpretable discrete variable, which is set to a word in the vocabulary. The sampled latent word has an explicit semantic meaning, acting as a guide to the generation of informative and diverse responses. We also propose to use a two-stage sampling approach to enable efficient selection of diverse variables from a large latent space, which is very essential for our model. Experimental results show that our model outperforms various kinds of generation models under both automatic and human evaluations.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No. 61751206, 61876120).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 2015 International Conference on Learning Representations*.

Zhihao Fan, Zhongyu Wei, Siyuan Wang, Yang Liu, and Xuanjing Huang. 2018. A reinforcement learning framework for natural question generation using bi-discriminators. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1763–1774.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. 2019. Generating multiple diverse responses for short-text conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6383–6390.

Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim. 2019. DialogWAE: Multimodal response generation with conditional wasserstein auto-encoder. In *Proceedings of the 2019 International Conference on Learning Representations*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. 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.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169.

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.

Yahui Liu, Wei Bi, Jun Gao, Xiaojiang Liu, Jian Yao, and Shuming Shi. 2018. Towards less generic responses in neural conversation models: A statistical re-weighting method. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2769–2774.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Ryo Nakamura, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura. 2018. Another diversity-promoting objective function for neural dialogue generation. *arXiv preprint arXiv:1811.08100*.

Diana Perez-Marin. 2011. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices: Techniques and Effective Practices*. IGI Global.

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 *Proceedings of AAAI Conference on Artificial Intelligence*, pages 3776–3784.

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

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1577–1586.

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A conditional variational framework for dialog generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 504–509.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–180.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205.

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

Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. 2017. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems*, pages 5756–5766.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. 2017. Latent intention dialogue models. In *International Conference on Machine Learning*, pages 3732–3741.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 3351–3357.

Zhen Xu, Bingquan Liu, Baoxun Wang, SUN Chengjie, Xiaolong Wang, Zhuoran Wang, and Chao Qi. 2017. Neural response generation via gan with an approximate embedding layer. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 617–626.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2016. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer.

Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. 2017. Towards implicit content-introducing for generative short-text conversation systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2190–2199.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pages 1810–1820.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 55th Annual Meeting of the 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 *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 654–664.

Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 3400–3407.