# Are Training Samples Correlated? Learning to Generate Dialogue Responses with Multiple References

**Lisong Qiu**[1,2], **Juntao Li**[1,2], **Wei Bi**[3], **Dongyan Zhao**[1,2], **Rui Yan**[1,2*]
[1]Center for Data Science, Peking University, Beijing, China
[2]Institute of Computer Science and Technology, Peking University, Beijing, China
[3]Tencent AI Lab, Shenzhen, China
{qiuls,lijuntao,zhaody,ruiyan}@pku.edu.cn
victoriabi@tencent.com

## Abstract

Due to its potential applications, open-domain dialogue generation has become popular and achieved remarkable progress in recent years, but sometimes suffers from generic responses. Previous models are generally trained based on 1-to-1 mapping from an input query to its response, which actually ignores the nature of 1-to-n mapping in dialogue that there may exist multiple valid responses corresponding to the same query. In this paper, we propose to utilize the multiple references by considering the correlation of different valid responses and modeling the 1-to-n mapping with a novel two-step generation architecture. The first generation phase extracts the common features of different responses which, combined with distinctive features obtained in the second phase, can generate multiple diverse and appropriate responses. Experimental results show that our proposed model can effectively improve the quality of response and outperform existing neural dialogue models on both automatic and human evaluations.

## 1 Introduction

In recent years, open-domain dialogue generation has become a research hotspot in Natural Language Processing due to its broad application prospect, including chatbots, virtual personal assistants, etc. Though plenty of systems have been proposed to improve the quality of generated responses from various aspects such as topic (Xing et al., 2017), persona modeling (Zhang et al., 2018b) and emotion controlling (Zhou et al., 2018b), most of these recent approaches are primarily built upon the sequence-to-sequence architecture (Cho et al., 2014; Shang et al., 2015) which suffers from the "safe" response problem (Li et al., 2016a; Sato et al., 2017). This can be ascribed to modeling the response generation process as 1-to-1 mapping, which ignores the nature of 1-to-n mapping of dialogue that multiple possible responses can correspond to the same query.

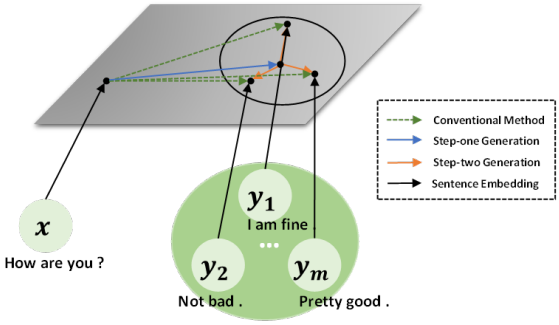


Figure 1: An illustration of the two-step generation architecture. Different from the conventional methods (shown in green color) which model each response from scratch every time, our method first builds a common feature of multiple responses and models each response based on it afterward.

To deal with the generic response problem, various methods have been proposed, including diversity-promoting objective function (Li et al., 2016a), enhanced beam search (Shao et al., 2016), latent dialogue mechanism (Zhou et al., 2017, 2018a), Variational Autoencoders (VAEs) based models (Zhao et al., 2017; Serban et al., 2017), etc. However, these methods still view multiple responses as independent ones and fail to model multiple responses jointly. Recently, Zhang et al. (2018a) introduce a maximum likelihood strategy that given an input query, the most likely response is approximated rather than all possible responses, which is further implemented by Rajendran et al. (2018) with reinforcement learning for task-oriented dialogue. Although capable of generating the most likely response, these methods fail to model other possible responses and ignore the correlation of different responses.

In this paper, we propose a novel response generation model for open-domain conversation, which learns to generate multiple diverse responses with multiple references by considering

*Corresponding author: Rui Yan (ruiyan@pku.edu.cn)



*Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3826–3835
Florence, Italy, July 28 - August 2, 2019. 

the correlation of different responses. Our motivation lies in two aspects: 1) multiple responses for a query are likely correlated, which can facilitate building the dialogue system. 2) it is easier to model each response based on other responses than from scratch every time. As shown in Figure 1, given an input query, different responses may share some common features e.g. positive attitudes or something else, but vary in discourses or expressions which we refer to as distinct features. Accordingly, the system can benefit from modeling these features respectively rather than learning each query-response mapping from scratch.

Inspired by this idea, we propose a two-step dialogue generation architecture as follows. We jointly view the multiple possible responses to the same query as a response bag. In the first generation phase, the common feature of different valid responses is extracted, serving as a base from which each specific response in the bag is further approximated. The system then, in the second generation phase, learns to model the distinctive feature of each individual response which, combined with the common feature, can generate multiple diverse responses simultaneously.

Experimental results show that our method can outperform existing competitive neural models under both automatic and human evaluation metrics, which demonstrates the effectiveness of the overall approach. We also provide ablation analyses to validate each component of our model. To summarize, our contributions are threefold:

- We propose to model multiple responses to a query jointly by considering the correlations of responses with multi-reference learning.
- We consider the common and distinctive features of the response bag and propose a novel two-step dialogue generation architecture.
- Experiments show that the proposed method can generate multiple diverse responses and outperform existing competitive models on both automatic and human evaluations.

## 2 Related Work

Along with the flourishing development of neural networks, the sequence-to-sequence framework has been widely used for conversation response generation (Shang et al., 2015; Sordoni et al., 2015) where the mapping from a query $x$ to a reply $y$ is learned with the negative log likelihood. However, these models suffer from the "safe" response problem. To address this problem, various methods have been proposed. Li et al. (2016a) propose a diversity-promoting objective function to encourage diverse responses during decoding. Zhou et al. (2017, 2018a) introduce a responding mechanism between the encoder and decoder to generate various responses. Xing et al. (2017) incorporate topic information to generate informative responses. However, these models suffer from the deterministic structure when generating multiple diverse responses. Besides, during the training of these models, response utterances are only used in the loss function and ignored when forward computing, which can confuse the model for pursuing multiple objectives simultaneously.

A few works explore to change the deterministic structure of sequence-to-sequence models by introducing stochastic latent variables. VAE is one of the most popular methods (Bowman et al., 2016; Zhao et al., 2017; Serban et al., 2017; Cao and Clark, 2017), where the discourse-level diversity is modeled by a Gaussian distribution. However, it is observed that in the CVAE with a fixed Gaussian prior, the learned conditional posteriors tend to collapse to a single mode, resulting in a relatively simple scope (Wang et al., 2017). To tackle this, WAE (Gu et al., 2018) which adopts a Gaussian mixture prior network with Wasserstein distance and VAD (Du et al., 2018) which sequentially introduces a series of latent variables to condition each word in the response sequence are proposed. Although these models overcome the deterministic structure of sequence-to-sequence model, they still ignore the correlation of multiple valid responses and each case is trained separately.

To consider the multiple responses jointly, the maximum likelihood strategy is explored. Zhang et al. (2018a) propose the maximum generated likelihood criteria which model a query with its multiple responses as a bag of instances and proposes to optimize the model towards the most likely answer rather than all possible responses. Similarly, Rajendran et al. (2018) propose to reward the dialogue system if any valid answer is produced in the reinforcement learning phase. Though considering multiple responses jointly, the maximum likelihood strategy fails to utilize all the references during training with some cases ig-

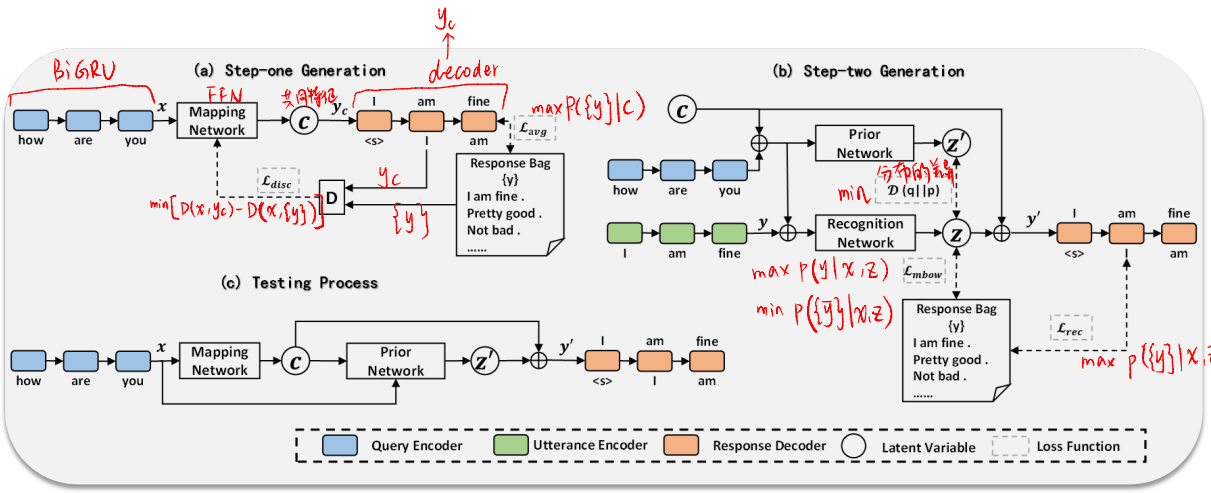


Figure 2: The overall architecture of our proposed dialogue system where the two generation steps and testing process are illustrated. Given an input query $x$, the model aims to approximate the multiple responses in a bag $\{y\}$ simultaneously with the continuous common and distinctive features, i.e., the latent variables $\mathbf{c}$ and $\mathbf{z}$ obtained from the two generation phases respectively.

nored. In our approach, we consider multiple responses jointly and model each specific response separately by a two-step generation architecture.

## 3 Approach

In this paper, we propose a novel response generation model for short-text conversation, which models multiple valid responses for a given query jointly. We posit that a dialogue system can benefit from multi-reference learning by considering the correlation of multiple responses. Figure 2 demonstrates the whole architecture of our model. We now describe the details as follows.

### 3.1 Problem Formulation and Model Overview

Training samples $\{(\mathbf{x}, \{\mathbf{y}\})_i\}_{i=1}^{i=N}$ consist of each query $\mathbf{x}$ and the set of its valid responses $\{\mathbf{y}\}$, where $N$ denotes the number of training samples. For a dialogue generation model, it aims to map from the input query $\mathbf{x}$ to the output response $\mathbf{y} \in \{\mathbf{y}\}$. To achieve this, different from conventional methods which view the multiple responses as independent ones, we propose to consider the correlation of multiple responses with a novel two-step generation architecture, where the response bag $\{\mathbf{y}\}$ and each response $\mathbf{y} \in \{\mathbf{y}\}$ are modeled by two separate features which are obtained in each generation phase respectively. Specifically, we assume a variable $\mathbf{c} \in \mathbb{R}^n$ representing the common feature of different responses and an unobserved latent variable $\mathbf{z} \in \mathbf{Z}$ corresponding to the distinct feature for each $\mathbf{y}$ in the bag. The common feature $\mathbf{c}$ is generated in the first stage given $\mathbf{x}$ and the distinctive feature $\mathbf{z}$ is sampled from the latent space $\mathbf{Z}$ in the second stage given the query $\mathbf{x}$ and common feature $\mathbf{c}$. The final responses are then generated conditioned on both the common feature $\mathbf{c}$ and distinct feature $\mathbf{z}$ simultaneously.

### 3.2 Common Feature of the Response Bag

In the first generation step, we aim to map from the input query $\mathbf{x}$ to the common feature $\mathbf{c}$ of the response bag $\{\mathbf{y}\}$. Inspired by multi-instance learning (Zhou, 2004), we start from the simple intuition that it is much easier for the model to fit multiple instances from their mid-point than a random start-point, as illustrated in Figure 1.

To obtain this, we model the common feature of the response bag as the mid-point of embeddings of multiple responses. In practice, we first encode the input $\mathbf{x}$ with a bidirectional gated recurrent units (GRU) (Cho et al., 2014) to obtain an input representation $\mathbf{h}_x$. Then, the common feature $\mathbf{c}$ is computed by a mapping network which is implemented by a feed-forward neural network whose trainable parameter is denoted as $\theta$. The feature $\mathbf{c}$ is then fed into the response decoder to obtain the intermediate response $\mathbf{y}_c$ which is considered to approximate all valid responses. Mathematically, the objective function is defined as:

$$\max \quad \mathcal{L}_{avg} = \frac{1}{|\{\mathbf{y}\}|} \sum_{\mathbf{y} \in \{\mathbf{y}\}} \log p_\psi(\mathbf{y}|\mathbf{c}) \qquad (1)$$

where $|\{\mathbf{y}\}|$ is the cardinality of the response bag $\{\mathbf{y}\}$ and $p_\psi$ represents the response decoder.

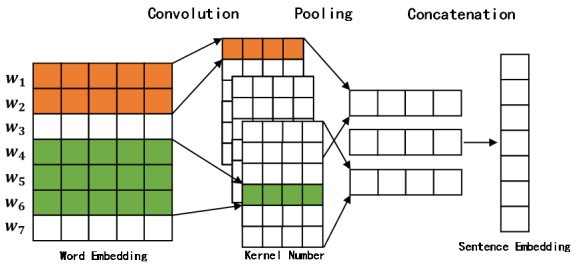


Figure 3: The sentence embedding function of the discriminator in the first generation phase.

Besides, to measure how well the intermediate response $\mathbf{y}_c$ approximates the mid-point response, we set up an individual discriminator and derive the mapping function to produce better results. As to the discriminator, we first project each utterance to an embedding space with fixed dimensionality via convolutional neural networks (CNNs) with different kernels as the process shown in Figure 3. Then, the cosine similarity of the query and response embeddings is computed, denoted as $D_{\theta'}(\mathbf{x}, \mathbf{y})$, where $\theta'$ represents trainable parameter in the discriminator. For the response bag $\{\mathbf{y}\}$, the average response embedding is used to compute the matching score. The objective of intermediate response $\mathbf{y}_c$ is then to minimize the difference between $D_{\theta'}(\mathbf{x}, \mathbf{y}_c)$ and $D_{\theta'}(\mathbf{x}, \{\mathbf{y}\})$:

$$\mathcal{L}_{disc} = \mathbb{E}_{\mathbf{x},\{\mathbf{y}\},\mathbf{y}_c}[D_{\theta'}(\mathbf{x}, \mathbf{y}_c) - D_{\theta'}(\mathbf{x}, \{\mathbf{y}\})] \quad (2)$$

where $\mathbf{y}_c$ denotes the utterance produced by the decoder conditioned on the variable $\mathbf{c}$.

To overcome the discrete and non-differentiable problem, which breaks down gradient propagation from the discriminator, we adopt a "soft" continuous approximation (Hu et al., 2017):

$$\hat{\mathbf{y}}_{c_t} \sim \text{softmax}(\mathbf{o_t}/\tau) \quad (3)$$

where $\mathbf{o_t}$ is the logit vector as the inputs to the softmax function at time-step $t$ and the temperature $\tau$ is set to $\tau \to 0$ as training proceeds for increasingly peaked distributions. The whole loss for the step-one generation is then

$$\mathcal{L}_{first} = \mathcal{L}_{avg} + \mathcal{L}_{disc} \quad (4)$$

which is optimized by a minimax game with adversarial training (Goodfellow et al., 2014).

### 3.3 Response Specific Generation

The second generation phase aims to model each specific response in a response bag respectively. In practice, we adopt the CVAE (Sohn et al., 2015; Yan et al., 2015) architecture, while two prominent modifications remain. Firstly, rather than modeling each response with the latent variable $\mathbf{z}$ from scratch, our model approximates each response based on the bag representation $\mathbf{c}$ with only the distinctive feature of each specific response remaining to be captured. Secondly, the prior common feature $\mathbf{c}$ can provide extra information for the sampling network which is supposed to decrease the latent searching space.

Specifically, similar to the CVAE architecture, the overall objective for our model in the second generation phase is as below:

$$\begin{aligned}\mathcal{L}_{cvae} =& \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y},\mathbf{c})p_\theta(\mathbf{c}|\mathbf{x})}[\log p_\psi(\mathbf{y}|\mathbf{c},\mathbf{z})] \\ &- \mathcal{D}[q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y},\mathbf{c})||p_\varphi(\mathbf{z}|\mathbf{x},\mathbf{c})]\end{aligned} \quad (5)$$

where $q_\phi$ represents the recognition network and $p_\varphi$ is the prior network with $\phi$ and $\varphi$ as the trainable parameters; $\mathcal{D}(\cdot||\cdot)$ is the regularization term which measures the distance between the two distributions. In practice, the recognition networks are implemented with a feed-forward network that

$$\begin{bmatrix}\mu \\ \log(\sigma^2)\end{bmatrix} = \mathbf{W}_q \begin{bmatrix}\mathbf{h}_x \\ \mathbf{h}_y \\ \mathbf{c}\end{bmatrix} + \mathbf{b}_q \quad (6)$$

where $\mathbf{h}_x$ and $\mathbf{h}_y$ are the utterance representations of query and response got by GRU respectively, and the latent variable $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2\mathbf{I})$. For the prior networks, we consider two kinds of implements. One is the vanilla CVAE model (Zhao et al., 2017) where the prior $p_\varphi(\mathbf{z}|\mathbf{x},\mathbf{c})$ is modeled by a another feed-forward network conditioned on the representations $\mathbf{h}_x$ and $\mathbf{c}$ as follows,

$$\begin{bmatrix}\mu' \\ \log(\sigma'^2)\end{bmatrix} = \mathbf{W}_p \begin{bmatrix}\mathbf{h}_x \\ \mathbf{c}\end{bmatrix} + \mathbf{b}_p \quad (7)$$

and the distance $\mathcal{D}(\cdot||\cdot)$ here is measured by the KL divergence. For the other, we adopt the WAE model (Gu et al., 2018) in which the prior $p_\varphi(\mathbf{z}|\mathbf{x},\mathbf{c})$ is modeled by a mixture of Gaussian distributions $\text{GMM}(\pi_k, \mu'_k, \sigma'^2_k\mathbf{I})_{k=1}^K$, where $K$ is the number of Gaussian distributions and $\pi_k$ is the mixture coefficient of the $k$-th component of the GMM module as computed:

$$\pi_k = \frac{\exp(e_k)}{\sum_{i=1}^K \exp(e_i)} \quad (8)$$

and

$$\begin{bmatrix} e_k \\ {\mu'}_k \\ \log {\sigma'}^2_k \end{bmatrix} = \mathbf{W}_{p,k} \begin{bmatrix} \mathbf{h}_x \\ \mathbf{c} \end{bmatrix} + \mathbf{b}_{p,k} \tag{9}$$

To sample an instance, Gumble-Softmax reparametrization trick (Kusner and Hernández-Lobato, 2016) is utilized to normalize the coefficients. The distance here is measured by the Wasserstein distance which is implemented with an adversarial discriminator (Zhao et al., 2018).

Recap that in the second generation phase the latent variable $\mathbf{z}$ is considered to only capture the distinctive feature of each specific response. Hence to distinguish the latent variable $\mathbf{z}$ for each separate response, we further introduce a multi-reference bag-of-word loss (MBOW) which requires the network to predict the current response $\mathbf{y}$ against the response bag:

$$\begin{aligned} \mathcal{L}_{mbow} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y},\mathbf{c})}[\log p(\mathbf{y}_{bow}|\mathbf{x},\mathbf{z}) \\ + \lambda \log(1 - p(\{\bar{\mathbf{y}}\}_{bow}|\mathbf{x},\mathbf{z}))] \end{aligned} \tag{10}$$

where the probability is computed by a feed-forward network $f$ as the vanilla bag-of-word loss (Zhao et al., 2017) does; $\{\bar{\mathbf{y}}\}$ is the complementary response bag of $\mathbf{y}$ and its probability is computed as the average probability of responses in the bag; and $\lambda$ is a scaling factor accounting for the difference in magnitude. As it shows, the MBOW loss penalizes the recognition networks if other complementary responses can be predicted from the distinctive variable $\mathbf{z}$. Besides, since the probability of the complementary term may approach zero which makes it difficult to optimize, we actually adopt its lower bound in practice:

$$\begin{aligned} \log(1 - p(\mathbf{y}_{bow}|\mathbf{x},\mathbf{z})) &= \log(1 - \prod_{t=1}^{|\mathbf{y}|} \frac{e^{f_{\mathbf{y}_t}}}{\sum_j^{|V|} e^{f_j}}) \\ &\geq \log(\prod_{t=1}^{|\mathbf{y}|}(1 - \frac{e^{f_{\mathbf{y}_t}}}{\sum_j^{|V|} e^{f_j}})) \end{aligned} \tag{11}$$

where $|V|$ is vocabulary size.

Totally, the whole loss for the step-two generation is then:

$$\mathcal{L}_{second} = \mathcal{L}_{cvae} + \mathcal{L}_{mbow} \tag{12}$$

which can be optimized in an end-to-end way.

### 3.4 Optimization and Testing

Our whole model can be trained in an end-to-end fashion. To train the model, we first pre-train the word embedding using Glove ((Pennington et al., 2014))[1]. Then modules of the model are jointly trained by optimizing the losses $\mathcal{L}_{first}$ and $\mathcal{L}_{second}$ of the two generation phases respectively. To overcome the vanishing latent variable problem (Wang et al., 2017) of CVAE, we adopt the KL annealing strategy (Bowman et al., 2016), where the weight of the KL term is gradually increased during training. The other technique employed is the MBOW loss which is able to sharpen the distribution of latent variable $\mathbf{z}$ for each specific response and alleviate the vanishing problem at the same time.

During testing, diverse responses can be obtained by the two generation phases described above, where the distinctive latent variable $\mathbf{z}$ corresponding to each specific response is sampled from the prior probability network. This process is illustrated in Figure 2. Capable of capturing the common feature of the response bag, the variable $\mathbf{c}$ is obtained from the mapping network and no intermediate utterance is required, which facilitates reducing the complexity of decoding.

## 4 Experimental Setup

### 4.1 Dataset

Focusing on open-domain dialogue, we perform experiments on a large-scale single-turn conversation dataset Weibo (Shang et al., 2015), where each input post is generally associated with multiple response utterances[2]. Concretely, the Weibo dataset consists of short-text online chit-chat dialogues in Chinese, which is crawled from Sina Weibo [3]. Totally, there are 4,423,160 query-response pairs for training set and 10000 pairs for the validation and testing, where there are around 200k unique query in the training set and each query used in testing correlates with four responses respectively. For preprocessing, we follow the conventional settings (Shang et al., 2015).

### 4.2 Baselines

We compare our model with representative dialogue generation approaches as listed below:

[1] https://nlp.stanford.edu/projects/glove/

[2] More such multi-reference data is widely available, e.g. social media like Twitter. But we adopt Weibo in this work since it is large and publicly available.

[3] https://www.weibo.com/

| Method | Multi-BLEU | | EMBEDDING | | | Intra-Dist | | Inter-Dist | |
|---|---|---|---|---|---|---|---|---|---|
| | BLEU-1 | BLEU-2 | G | A | E | Dist-1 | Dist-2 | Dist-1 | Dist-2 |
| **S2S** | 21.49 | 9.498 | 0.567 | 0.677 | 0.415 | 0.311 | 0.447 | 0.027 | 0.127 |
| **S2S+DB** | 20.20 | 9.445 | 0.561 | 0.682 | 0.422 | 0.324 | 0.457 | 0.028 | 0.130 |
| **MMS** | 21.40 | 9.398 | 0.569 | 0.691 | 0.427 | 0.561 | 0.697 | 0.033 | 0.158 |
| **CVAE** | 22.71 | 8.923 | 0.601 | 0.730 | 0.452 | 0.628 | 0.801 | 0.035 | 0.179 |
| **CVAE+BOW** | 23.12 | 8.420 | 0.605 | 0.741 | 0.456 | 0.687 | 0.873 | 0.038 | 0.194 |
| **WAE** | 24.02 | 9.281 | 0.611 | 0.754 | 0.460 | **0.734** | 0.885 | 0.044 | 0.196 |
| **Ours-First** | 23.68 | 9.240 | 0.619 | 0.762 | 0.471 | 0.725 | **0.891** | **0.045** | **0.199** |
| **Ours-Disc** | **24.22** | 9.101 | 0.617 | 0.754 | 0.465 | 0.670 | 0.863 | 0.036 | 0.184 |
| **Ours-MBOW** | 23.88 | **9.582** | 0.622 | **0.778** | 0.477 | 0.681 | 0.877 | 0.040 | 0.190 |
| **Ours** | 24.04 | 9.362 | **0.625** | 0.771 | 0.480 | 0.699 | 0.876 | 0.042 | 0.190 |
| **Ours+GMP** | 24.20 | 9.417 | 0.618 | 0.769 | **0.482** | 0.728 | 0.889 | 0.044 | 0.198 |

Table 1: Automatic evaluation results of different models where the best results are bold. The **G**, **A** and **E** of **Embedding** represent Greedy, Average, Extreme embedding-based metrics, repsectively.

| Method | Rela. | Divt. | Red. | Overall |
|---|---|---|---|---|
| **Gold** | 3.90 | 4.22 | 3.79 | 3.97 |
| **S2S** | 3.10 | 2.77 | **3.24** | 3.07 |
| **CVAE** | 2.98 | 3.12 | 3.10 | 3.07 |
| **Ours** | **3.22** | **3.19** | 3.23 | **3.21** |

Table 2: Human evaluation results of different models. **Rela.**, **Divt.** and **Red.** represent *Relevance*, *Diversity* and *Readability*, respectively. The Kappa score among different human evaluators is 0.4412, which indicates moderate human agreements.

**S2S**: the vanilla sequence-to-sequence model with attention mechanism (Bahdanau et al., 2014) where standard beam search is applied in testing to generate multiple different responses.

**S2S+DB**: the vanilla sequence-to-sequence model with the modified diversity-promoting beam search method (Li et al., 2016b) where a fixed diversity rate 0.5 is used.

**MMS**: the modified multiple responding mechanisms enhanced dialogue model proposed by Zhou et al. (2018a) which introduces responding mechanism embeddings (Zhou et al., 2017) for diverse response generation.

**CVAE**: the vanilla CVAE model (Zhao et al., 2017) with and without BOW (bag-of-word) loss (**CVAE+BOW** and **CVAE**).

**WAE**: the conditional Wasserstein autoencoder model for dialogue generation (Gu et al., 2018) which models the distribution of data by training a GAN within the latent variable space.

**Ours**: we explore our model **Ours** and conduct various ablation studies: the model with only the second stage generation (**Ours-First**), the model without the discriminator (**Ours-Disc**) and multi-reference BOW loss (**Ours-MBOW**), and the model with GMM prior networks (**Ours+GMP**).

### 4.3 Evaluation Metrics

To comprehensively evaluate the quality of generated response utterances, we adopt both automatic and human evaluation metrics:

**BLEU**: In dialogue generation, BLEU is widely used in previous studies (Yao et al., 2017; Shang et al., 2018). Since multiple valid responses exist in this paper, we adopt multi-reference BLEU where the evaluated utterance is compared to provided multiple references simultaneously.

**Distinctness**: To distinguish safe and commonplace responses, the distinctness score (Li et al., 2016a) is designed to measure word-level diversity by counting the ratio of distinctive [1,2]-grams. In our experiments, we adopt both **Intra-Dist**: the distinctness scores of multiple responses for a given query and **Inter-Dist**: the distinctness scores of generated responses of the whole testing set.

**Embedding Similarity**: Embedding-based metrics compute the cosine similarity between the sentence embedding of a ground-truth response and that of the generated one. There are various ways to obtain the sentence-level embedding from the constituent word embeddings. In our experiments, we apply three most commonly used strategies: $Greedy$ matches each word of the reference with the most similar word in the evaluated sentence; $Average$ uses the average of word embed-

| | | |
|---|---|---|
| **Input Query** | 火山喷发瞬间的一些壮观景象。<br>These are some magnificent sights at the moment of the volcanic eruption. | 再过十分钟就进入win8时代，我是系统升级控。<br>There remain ten minutes before we entering the era of win8. I am a geek of system updating. |
| **Gold** | 大自然才是人类的最终boss。<br>Nature is the final boss of human.<br>真帅，12月份的时候就能亲眼看到了，好开心啊。<br>So cool! I am so happy to see it by myself in December.<br>被惊艳震撼到了。<br>I am deeply surprised and shocked.<br>震撼了，小小人类仰视造物主的强大。<br>Shocked! The imperceptible humanity looks up to the power of the creator. | 问个白痴问题必须正版才能升级吧?<br>May I ask an idiot problem. Does the update require a license?<br>不是给平板电脑用的系统吗?<br>Isn't this system for PAD?<br>已经用了一个多月了，不过还是不喜欢8<br>I have used it for a month but I still don't like it 8<br>好久木用电脑了，想念。<br>Having not used the computer for a long time, I miss it. |
| **CVAE** | 大半夜的不光是白天。<br>It's midnight, not only daytime.<br>一天一天就能看到了。<br>We can see it day after day.<br>天地之间的风景有如此之美。<br>How could there exist such amazing sights.<br>火山喷发瞬间的萤火虫。<br>The glowworm at the moment of volcanic eruption. | 这是要用手机吗?<br>Do you want to use the phone?<br>我是升级了升级版了。<br>I have updated to the upgrade.<br>我还以为是我的电脑。<br>I thought it was my computer.<br>升级版的机器人。<br>The upgraded robot. |
| **Ours** | 好美，这是哪里呀?<br>So amazing! Where is this?<br>好壮观啊一定要保存下来。<br>It's so magnificent that it should be preserved.<br>大白天的不能看到。<br>It can't be seen during the day.<br>如果有机会亲眼所见过。<br>If you have chance to see it yourself.<br>如此这般这般淼小。<br>It is so so imperceptible. | 这是什么软件啊，求解。<br>I am wondering what software it is.<br>我觉得微软的ui还不错。<br>I think the ui of Microsoft is not bad.<br>现在的产品已经不是新产品了。<br>The current product is not the new.<br>这个是什么应用啊。<br>What application is this.<br>我觉得这样的界面更像windows8。<br>I think interface like this looks more like windows8. |

Table 3: Case study for the generated responses from the testing set of Weibo, where the Chinese utterances are translated into English for the sake of readability. For each input query, we show four responses generated by each method and an additional intermediate utterance (marked with underline) for our model.

dings; and $Extreme$ takes the most extreme value among all words for each dimension of word embeddings in a sentence. Since multiple references exist, for each utterance to be evaluated, we compute its score with the most similar reference.

**Human Evaluation with Case Analysis**: As automatic evaluation metrics lose sight of the overall quality of a response (Tao et al., 2018), we also adopt human evaluation on 100 random samples to assess the generation quality with three independent aspects considered: *relevance* (whether the reply is relevant to the query), *diversity* (whether the reply narrates with diverse words) and *readability* (whether the utterance is grammatically formed). Each property is assessed with a score from 1 (worst) to 5 (best) by three annotators. The evaluation is conducted in a blind process with the utterance belonging unknown to the reviewers.

### 4.4 Implementation Details

All models are trained with the following hyperparameters: both encoder and decoder are set to one layer with GRU (Cho et al., 2014) cells, where the hidden state size of GRU is 256; the utterance length is limited to 50; the vocabulary size is 50,000 and the word embedding dimension is 256; the word embeddings are shared by the encoder and decoder; all trainable parameters are initialized from a uniform distribution [-0.08, 0.08]; we employ the Adam (Kingma and Ba, 2014) for optimization with a mini-batch size 128 and initialized learning rate 0.001; the gradient clipping strategy is utilized to avoid gradient explosion, where the gradient clipping value is set to be 5. For the latent variable, we adopt dimensional size 256 and the component number of the mixture Gaussian for prior networks in WAE is set to 5. As to the discriminator, we set the initialized learning rate as 0.0002 and use 128 different kernels for each kernel size in $\{2, 3, 4\}$. The size of the response bag is limited to 10 where the instances inside are randomly sampled for each mini-batch. All the models are implemented with Pytorch 0.4.1 [4].

## 5 Results and Analysis

### 5.1 Comparison against Baselines

Table 1 shows our main experimental results, with baselines shown in the top and our models at the bottom. The results show that our model (Ours) outperforms competitive baselines on various evaluation metrics. The Seq2seq based models (S2S, S2S-DB and MMS) tend to generate

[4]https://pytorch.org

fluent utterances and can share some overlapped words with the references, as the high BLEU-2 scores show. However, the distinctness scores illustrate that these models fail to generate multiple diverse responses in spite of the diversity-promoting objective and responding mechanisms used. We attribute this to that these models fail to consider multiple references for the same query, which may confuse the models and lead to a commonplace utterance. As to the CVAE and WAE models, with the latent variable to control the discourse-level diversity, diverse responses can be obtained. Compared against these previous methods, our model can achieve the best or second best performances on different automatic evaluation metrics where the improvements are most consistent on BLEU-1 and embedding-based metrics, which demonstrates the overall effectiveness of our proposed architecture.

In order to better study the quality of generated responses, we also report the human evaluation results in Table 2. As results show, although there remains a huge gap between existing methods and human performance (the Gold), our model gains promising promotions over previous methods on generating appropriate responses with diverse expressions. With both obvious superiority (readability for S2S and diversity for CVAE) and inferiority (diversity for S2S and relevance for CVAE), the baselines show limited overall performances, in contrast to which our method can output more diverse utterances while maintaining the relevance to the input query and achieve a high overall score.

### 5.2 Ablation Study

To better understand the effectiveness of each component in our model, we further conduct the ablation studies with results shown at the bottom of Table 1. Above all, to validate the effectiveness of the common feature, we remove the first generation stage and get the Ours-First model. As the results of BLEU and embedding-based metrics show, the system can benefit from the common feature for better relevance to the query.

Moreover, pairwise comparisons Ours-Disc vs. Ours and Ours-MBOW vs. Ours validate the effects of the discriminator and modified multi-reference bag-of-word loss (MBOW). As results show, the discriminator facilitates extracting the common feature and yields more relevant responses to the input query afterward. The MBOW loss, similar to the effects of BOW loss in the CVAE, can lead to a more unique latent variable for each response and improve the final distinctness scores of generated utterances. In the experiments, we also observed the KL vanishing problem when training our model and we overcame it with the KL weight annealing strategy and the MBOW loss described above.

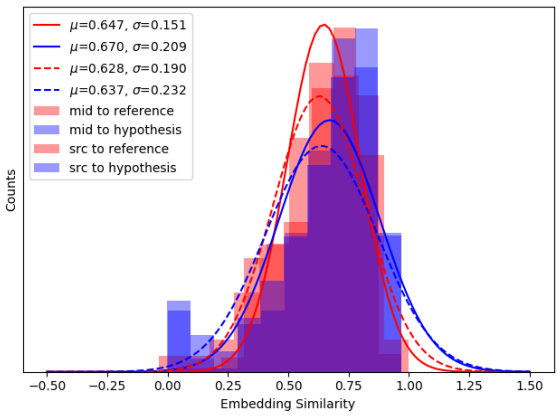


Figure 4: The statistics of distances between the input query/intermediate utterance and gold references/generated responses, where the distance is measured by the cosine similarity of sentence embeddings.

### 5.3 Case Study and Discussion

Table 3 illustrates two examples of generated replies to the input query got from the testing set. Comparing the CVAE and Ours, we can find that although the CVAE model can generate diverse utterances, its responses tend to be irrelevant to the query and sometimes not grammatically formed, e.g. the words "glowworm" and "robot" in the sentences. In contrast, responses generated by our model show better quality, achieving both high relevance and diversity. This demonstrates the ability of the two-step generation architecture. For better insight into the procedure, we present the intermediately generated utterances which show that the feature extracted in the first stage can focus on some common and key aspects of the query and its possible responses, such as the "amazing" and "software". With the distinctive features sampled in the second generation phase, the model further revises the response and outputs multiple responses with diverse contents and expressions.

Recap that the common feature is expected to capture the correlations of different responses and serve as the base of a response bag from which different responses are further generated, as shown

in Figure 1. To investigate the actual performances achieved by our model, we compute the distance between the input query/intermediate utterance and gold references/generated responses and present the results in Figure 4. As shown, intermediate utterances obtained in the first generation phase tend to approximate multiple responses with similar distances at the same time. Comparing the generated responses and the references, we find that generated responses show both high relevant and irrelevant ratios, as the values near 0.00 and 1.00 show. This actually agrees well with our observation that the model may sometimes rely heavily on or ignore the prior common feature information. From a further comparison between the input query and the mid, we also observe that the intermediate utterance is more similar to final responses than the input query, which correlates well with our original intention shown in Figure 1.

## 6 Conclusion and future work

In this paper, we tackle the one-to-many query-response mapping problem in open-domain conversation and propose a novel two-step generation architecture with the correlation of multiple valid responses considered. Jointly viewing the multiple responses as a response bag, the model extracts the common and distinct features of different responses in two generation phases respectively to output multiple diverse responses. Experimental results illustrate the superior performance of the proposed model in generating diverse and appropriate responses compared to previous representative approaches. However, the modeling of the common and distinct features of responses in our method is currently implicit and coarse-grained. Directions of future work may be pursuing better-defined features and easier training strategies.

## 7 Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2017YFC0804001), the National Science Foundation of China (NSFC No. 61672058; NSFC No. 61876196). Rui Yan was sponsored by CCF-Tencent Open Research Fund and Alibaba Innovative Research (AIR) Fund.

## References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21.

Kris Cao and Stephen Clark. 2017. Latent variable dialogue models and their diversity. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 182–187.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Jiachen Du, Wenjie Li, Yulan He, Ruifeng Xu, Lidong Bing, and Xuan Wang. 2018. Variational autoregressive decoder for neural response generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3154–3163.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Xiaodong Gu, Kyunghyun Cho, Jungwoo Ha, and Sunghun Kim. 2018. Dialogwae: Multimodal response generation with conditional wasserstein auto-encoder. *arXiv preprint arXiv:1805.12352*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. *arXiv preprint arXiv:1703.00955*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Matt J Kusner and José Miguel Hernández-Lobato. 2016. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, 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*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Janarthanan Rajendran, Jatin Ganhotra, Satinder Singh, and Lazaros Polymenakos. 2018. Learning end-to-end goal-oriented dialog with multiple answers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3834–3843.

Shoetsu Sato, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2017. Modeling situations in neural chat bots. In *Proceedings of ACL 2017, Student Research Workshop*, pages 120–127.

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 *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1577–1586.

Mingyue Shang, Zhenxin Fu, Nanyun Peng, Yansong Feng, Dongyan Zhao, and Rui Yan. 2018. Learning to converse with noisy data: Generation with calibration. In *IJCAI*, pages 4338–4344.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2016. Generating long and diverse responses with neural conversation models. *openreview*.

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.

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. *arXiv preprint arXiv:1506.06714*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *AAAI*, volume 17, pages 3351–3357.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2015. Attribute2image: Conditional image generation from visual attributes. *arXiv preprint arXiv:1512.00570*.

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.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018a. Tailored sequence to sequence models to different conversation scenarios. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1479–1488.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *International Conference on Machine Learning*, pages 5897–5906.

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 (Volume 1: Long Papers)*, 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 *AAAI*, pages 3400–3407.

Ganbin Zhou, Ping Luo, Yijun Xiao, Fen Lin, Bo Chen, and Qing He. 2018a. Elastic responding machine for dialog generation with dynamically mechanism selecting. In *AAAI Conference on Artificial Intelligence, AAAI*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018b. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhi-Hua Zhou. 2004. Multi-instance learning: A survey. *Department of Computer Science & Technology, Nanjing University, Tech. Rep*.