Modeling Personalization in Continuous Space for Response Generation via Augmented Wasserstein Autoencoders

Zhangming Chan[1,3,*], **Juntao Li**[1,3,*], **Xiaopeng Yang**[4], **Xiuying Chen**[1,3],
Wenpeng Hu[2,3], **Dongyan Zhao**[1,3] and **Rui Yan**[1,3,†]
[1] Center for Data Science, AAIS, Peking University
[2] School of Mathematical Sciences, Peking University
[3] Wangxuan Institute of Computer Technology, Peking University
[4] David R. Cheriton School of Computer Science, Faculty of Mathematics,
University of Waterloo
{zhangming.chan,ljt,ruiyan}@pku.edu.cn

Abstract

Variational autoencoders (VAEs) and Wasserstein autoencoders (WAEs) have achieved noticeable progress in open-domain response generation. Through introducing latent variables in continuous space, these models are capable of capturing utterance-level semantics, e.g., topic, syntactic properties, and thus can generate informative and diversified responses. In this work, we improve the WAE for response generation. In addition to the utterance-level information, we also model user-level information in latent continue space. Specifically, we embed user-level and utterance-level information into two multimodal distributions, and combine these two multimodal distributions into a mixed distribution. This mixed distribution will be used as the prior distribution of WAE in our proposed model, named as PersonaWAE. Experimental results on a large-scale real-world dataset confirm the superiority of our model for generating informative and personalized responses, where both automatic and human evaluations outperform state-of-the-art models.

1 Introduction

Over the past decade, a myriad of conversational systems have been proposed in the field of artificial intelligence and achieved remarkable success in various industry scenarios, such as e-commerce assistant (Li et al., 2017) and chit-chat machine XiaoIce (Shum et al., 2018). Based on the domains involved in previous research, existing work can be categorized into two groups, i.e., vertical-domain (Glas et al., 2015) and open-domain (Zhao et al., 2017), where the former group pursues to complete a specific target with limited domain knowledge while the latter one involves massive topics in conversations. In this work, we focus on the latter one and intend to generate a natural and meaningful response for a given conversation context. Most recent works build upon the sequence to sequence model (Bahdanau et al., 2014) and can generate a fluent response. But they suffer from the notorious "universal response" issue, i.e., generating safe and uninformative responses (e.g., I don't know) (Li et al., 2015).

To address the aforementioned shortcoming, advanced conversational systems propose to capture and incorporate extra information from two different levels, i.e., utterance-level and user-level. As for utterance-level information modeling, previous works mainly construct models upon variational autoencoders (VAE) (Kingma and Welling, 2014). By doing so, responses with diverse and informative words can be generated by introducing a latent variable for modeling utterance-level information such as topic, and syntactic structure (Bowman et al., 2015). It is verified in various open-domain response generation situations that conditional variational autoencoders (CVAE) (Serban et al., 2017; Zhao et al., 2017) are effective for addressing the "universal response" issue. In user-level information modeling, existing models either implicitly learn user information from training data such as learning user embedding (Li et al., 2015) or explicitly collect user profiles as the accurate personalization (Zhang et al., 2017; Yang et al., 2018; Zhang et al., 2018). Although obtaining user profiles is more effective and accurate than user embeddings, it is time-consuming and economically costly, or even impossible under the condition of protecting user privacy.

We propose the PersonaWAE model, a novel conversational system which simultaneously captures user-level personalization and utterance-level information as extra hints for generating better responses. Our model is motivated by following two points: 1) existing embedding based per-

[*] Equal contribution. Ordering is decided by a coin flip.
[†] 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 1931–1940,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

sona modeling methods cannot discover the common properties among users and train the embedding for different user independently, which cause (equal to learning) a very high-dimensional persona embedding and thus have a low data utilization efficiency or require a large amount of training data for each user. 2) benefited by the semantic capturing ability of WAEs, plenitude persona information can be gathered into the continuous space (Li et al., 2019). To this end, we build our model upon the state-of-the-art conversation model WAE (Gu et al., 2019) to model utterance-level and the user-level information. In the case, user embeddings are utilized as the condition of the prior distribution of the latent variable to formulate a WAE conditional prior. Meanwhile, to further model and fuse the utterance and user-level information, we extend these simple prior distributions to the Gaussian Mixture Distributions (GMDs, more details of the reasons in Section 2.2). After obtaining two GMDs, we combine them into a mixed distribution and regard this mixed distribution as the prior distribution of PersonaWAE. To evaluate the effectiveness of our proposed personalized conversational system, we collect a large dataset with user identifications. Experimental results on both automatic and human evaluation demonstrate that our proposed model can outperform several strong methods, and generate personalized responses for different users.

In a nutshell, our contributions can be summarized as follows:

- We proposed a novel personalized Wasserstein autoencoder (PersonaWAE) for open-domain response generation, which incorporates both utterance-level and user-level information;
- We proposed to mix two different types of Gaussian mixture distribution as the prior distribution of our model for scaling up the capability of the latent variable;
- Experiments performed on a large dataset demonstrate the effectiveness of our proposed model and achieves the new state-of-the-art results.

2 Preliminaries

2.1 VAE and WAE

Conditional VAE (CVAE) is a popular framework for dialogue generation (Zhao et al., 2017; Shen et al., 2017, 2018). CVAE, as an extension of VAE, supervises the generation process under an extra condition c. To train a CVAE model, the log-likelihood objective $\log p_\theta(x|c)$ is maximized through pushing up its variational lower bound:

$$\begin{aligned}\mathbb{O}(\theta, \phi, x, c) = &- \mathbf{KL}(q_\phi(z|x,c)||p_\theta(z|c)) \\ &+ \mathbf{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z,c)]\end{aligned} \tag{1}$$

where $q_\theta(z|x,c)$ and $p_\theta(z|c)$ represent the approximated conditional posterior and the conditional prior respectively, $\log p_\theta(x|z,c)$ represents the probability of reconstructing x conditioned on both z and c. Herein $KL(\cdot)$ represents the KL-divergence term, which serves as the regularization for encouraging the approximated posterior $q_\phi(z)$ to approach the prior $p_\theta(z)$, i.e. a standard Gaussian distribution. $E[\cdot]$ is the term of reconstruction loss, reflecting how well the decoder performs. The KL-divergence can be replaced by Wasserstein distance which is implemented by Arjovsky et al. (2017) and is proved to be superior to KL-divergence by many experiments. The conditional VAE based on Wasserstein distance is called conditional WAE.

2.2 Gaussian Mixture Model

In VAE/WAE frameworks, a variable in latent space is introduced for modeling information in datasets. As demonstrated in Figure 1, each gray point represents a data sample while colorized points refer to noise data. If the latent variable obeys a Gaussian distribution, noise samples (colorized points) will result in inferior responses. Alternatively, a Gaussian mixture distribution can model the datasets more accurately. Conventional VAE and WAE models usually set the prior distribution of the latent variable to a multivariate Gaussian distribution, formulated as

$$z \sim \mathcal{N}\{\mu, \sigma^2 I\} \tag{2}$$

where μ and σ^2 represent the mean and variance of $\mathcal{N}$. In our model, we utilize a Gaussian mixture distribution as the prior distribution of the latent variable z, written by

$$z \sim \mathcal{N}\{\pi_k, \mu_k, \sigma_k^2 I\}_{k=1}^K \tag{3}$$

where μ_k and σ_k^2 is the parameter for the k-th gaussian distribution in this multimodal distribution. π_k is the weight.

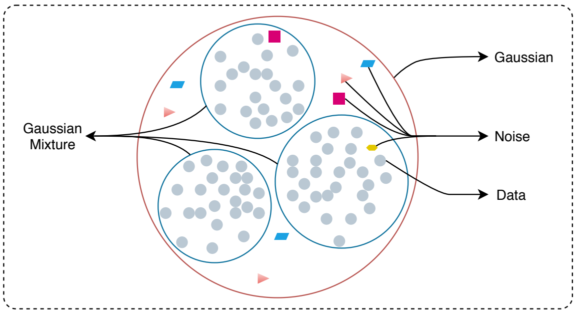

Figure 1: Distributions on latent space. The pink circle represents a Gaussian distribution while three small blue circles refer to a Gaussian mixture distribution, where gray points represent positive samples, and colorized points represent noise data.

2.3 Problem Formulation

We follow the conventional personalized conversation generation research (Li et al., 2016) and formulate the response generation task with the following necessary notations. A dataset with user dialogue history content $\mathcal{D} = \{(c_i, r_i, m_i)\}_{i=1}^{N}$ is firstly given, where c_i, r_i, m_i represent dialogue context, response candidate, and user specific dialogue utterance respectively. Note that we treat the user dialogue utterance for extracting personalization information in multi-turn response generation. Herein, the context is formulated by: $c_i = (s_1, s_2, \cdots, s_j, \cdots, s_{n_i})$ where s_j represents an utterance in the j-th turn of dialogue context and there are n_i utterances in the dialogue context. m_i denotes the user specific dialogue utterance. $r_i = \{r_{i,1}, r_{i,2}, \cdots, r_{i,n_r}\}$, where n_r is the length of a target response r_i. Then, our task is defined as learning a mapping function $f(\cdot)$ from the given dataset that can yield a personalized response according to the given dialogue context and the user dialogue history.

3 Proposed Model

As in Figure 2, our proposed personalized Wasserstein autoencoder (PersonaWAE) consists of user personalization modeling and WAE, where details are elaborated as follows.

3.1 User Personalization Modeling

Personalization Gaussian Mixture Distribution. To model the user-level information in the continue space, we build the Personalization Gaussian Mixture Distribution (Personalization GMD).

We train vector representations of users (Li et al., 2016) from user dialog history $\mathcal{M} = \{m_i\}_{i=1}^{N}$ as the user personalizations to facilitate personalized response generation. We denote the trained user embeddings as $\mathcal{U} = \{\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_i}\}$ where $\mathbf{u_i}$ represents the vector representations of i-th user (User $\mathbf{i}$).

Based on the user embeddings as $\mathcal{U}$, we utilize learned user personalizations in the latent space. Specifically, the conditional prior distribution of WAE part is a Gaussian mixture distribution (GMD) conditioned on the learned user embeddings, namely personalization GMD. We formulate the conditional prior as:

$$p(z_u|\mathbf{u_i}) = \sum_{k=1}^{K} v_k \mathcal{N}(z_u; \mu_k, \sigma_k^2 I) \tag{4}$$

where $\{\pi_k, \mu_k, \sigma_k^2 I\}_{k=1}^{K}$ represent the GMD (the distribution will deprecate to a Gaussian distribution when K=1) and the parameters of k-th Gaussian distribution are $\{\mu_k, \sigma_k^2\}$. v_k is a component indicator with class probabilities $\pi_1, \pi_2, \ldots, \pi_K$, where π_k is the mixture coefficient of the k-th component of the GMD. We follow (Gu et al., 2019) and compute these parameters as:

$$\begin{gathered}\begin{bmatrix} a_k \\ \mu_k \\ \log \sigma_k^2 \end{bmatrix} = W_k(\mathbf{u_i}) + b_k \\ \pi_k = \frac{e^{a_k}}{\sum_{i=1}^{K} e^{a_k}}\end{gathered} \tag{5}$$

To obtain v_k, we use the Gumbel-Softmax reparametrization to replace the exact sampling:

$$\begin{aligned} g_i &= -\log(-\log(b_i)) \\ v_k &= \frac{e^{(a_k+g_k)/\tau}}{\sum_{i=1}^{K} e^{(a_i+g_i)/\tau}}\end{aligned} \tag{6}$$

where b_i is a sample from $U(0,1)$, and τ is the softmax temperature to control the sampling process.

Fusion of Personalization in Decoder. We also incorporate the user embeddings into the decoder. Concretely, user personalization is used as the input of each updating step to obtain user-specific information for generating personalized responses. Meanwhile, $\mathbf{u_i}$ is updated by back-propagating loss signal during training. As user personalizations are high-level representations, we further introduce a gating strategy to dynamically balance the user personalization and the current conversation information.

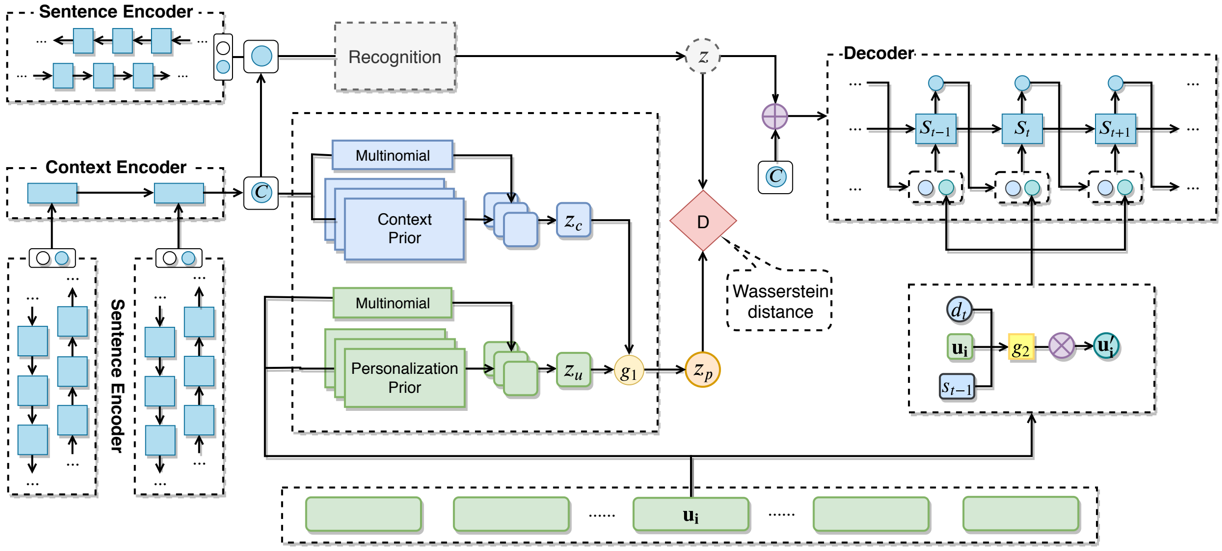

Figure 2: The detailed architecture of our proposed PersonaWAE model.

3.2 Personalized Wasserstein Autoencoder

Our proposed PersonaWAE consists of encoder, prior and recognition networks, and decoder.

Encoder. The encoder encodes a given context by a bidirectional RNN with GRU cells following (Chen et al., 2018). Through the encoder, the context $c_i = (u_1, u_2, \cdots, u_j, \cdots, u_{n_i})$ is represented as concatenated forward and backward[1] vectors $V_c = (v_1, v_2, \cdots, v_j, \cdots, v_{n_i})$, where $v_i = [\overrightarrow{v}_i, \overleftarrow{v}_i]$. Similarly, the target response r_i is represented by the concatenation of states from another bi-directional RNN with GRU cells, denoted as $v_{r,i}$. The vectors sequence V_c is further processed by an RNNs and yields a vector representation $v_{c,i}$. Note that $v_{c,i}$ refers to c in Equation 1 while $v_{r,i}$ represents x.

Recognition and Prior Networks. We use a recognition network to learn the posterior $q_\theta(z|x,c)$, we hypothesize that the approximated variational posterior follows an isotropic multivariate Gaussian distribution $\mathcal{N}(\mu, \sigma^2 I)$, where I represents the diagonal covariance. Thus modeling $q_\theta(z|x,c)$ is converted to learn μ and $\log \sigma^2$:

$$\begin{bmatrix} \mu \\ \log \sigma^2 \end{bmatrix} = W_o(\begin{bmatrix} x \\ c \end{bmatrix}) + b \tag{7}$$

which is presented as the recognition network in Figure 2. W_o and b are trainable parameters.

To approach the prior distribution, we superpose two conditional GMD, where the first one is personalization GMD as mentioned before while another conditional GMD that called context GMD is performed on context c. Resemble to the personalization GMD, the parameters of the context GMD $p_\phi(z_c|c)$ are defined as a'_k, μ'_k and $\log \sigma'^2_k$, which is learned by:

$$\begin{bmatrix} a'_k \\ \mu'_k \\ \log \sigma'^2_k \end{bmatrix} = W_{r_k}(c) + b'_k \tag{8}$$

Fusion of two GMDs. In fusing personalization and context GMD, we use the weighted addition strategy to superpose these two distributions, where the resulted new distribution is the prior distribution of PersonaWAE (which is also a GMD).

$$\begin{bmatrix} w_c \\ w_u \end{bmatrix} = softmax(W_f(\begin{bmatrix} c \\ \mathbf{u_i} \end{bmatrix})) \tag{9}$$
$$z_p = w_c \cdot z_c + w_u \cdot z_u$$

where W_f is a trainable parameter.

Decoder. The decoder is a one-layer GRU network to output the sentence in the generation, which is shown in the right hand of Figure 2. Taking the generation of response r_i as an example, the initial state of the decoder is calculated as:

$$s_{i,0} = W_d(\begin{bmatrix} z_p \\ c \end{bmatrix}) + b_d \tag{10}$$

where W_d is a trainable matrix for dimension transformation. To facilitate the combination of user personalization $\mathbf{u_i}$ and decoder hidden states,

[1] → and ← refer to forward and backward, respectively.

we incorporate a gate module (Tu et al., 2018) in our model:

$$g = f(\mathbf{U}s_{t-1} + \mathbf{V}d_t + \mathbf{W}\mathbf{u_i})$$
$$o_t = \mathbf{GRU}(s_{t-1}, \begin{bmatrix} d_t \\ g \cdot \mathbf{u_i} \end{bmatrix}) \tag{11}$$

where f is the sigmoid funtion and o_t refers to the decoder output in time step t. After processing o_t with the softmax operation, the response r_i is generated.

Training. To train our proposed model, we launch the following objective to simultaneously minimize the Wasserstein distance between $p_\theta(z|c, \mathbf{u_i})$ and $q_\phi(z|x, c)$, and maximize the reconstructing probability of x:

$$\begin{aligned} \mathbb{L}(\theta, \phi; c, x, \mathbf{u_i}) = & - \mathbf{E}_{q_\phi(z|x,c,\mathbf{u_i})} \log p(x|z, c, \mathbf{u_i}) \\ & + W(q_\phi(z|x, c)||p_\theta(z|c, \mathbf{u_i})) \end{aligned} \tag{12}$$

4 Experiments

4.1 Dataset

To evaluate the effectiveness of our proposed personalized WAE model (PersonaWAE), we collect a dataset from an open online chatting forum, i.e., Weibo [2], which contains massive multi-turn conversation sessions and user identification information. Overall, there are 31,128,520 utterances in the raw dataset with corresponded user identifications. To construct the personalized conversation systems, we retrieve users with more than 14 utterances from the raw Weibo corpus. We also filtrate conversation sessions with less than 2 turns for training multi-turn conversation systems. We use a sliding window with a size of 3 to construct each dialogue session and there are 3 utterances in each dialogue session. By doing so, there are 336,342 conversation sessions in the cleaned corpus. We remove emojis in utterances and utilize NLTK for tokenization. Then, we randomly split the Weibo corpus into 335,342/5,000/5,000 sessions as training/validation/testing sets. For each session, the last utterance is the target response for generation while other utterances are treated as context.

4.2 Baselines

In our experiments, we compare our proposed method with the following highly related and strong baselines.

Readability	Is the response grammatically formed and smooth ?
Informativeness	Does the response contains informative words ?
Personalization	Does the response resembles with any user history?

Table 1: Criteria of human evaluation.

Seq2Seq, the vanilla schema of the sequence to sequence model with attention mechanism (Bahdanau et al., 2014), which is widely used in various generation tasks.

Persona, a typical and recent neural personalized conversation system, which incorporates user-level representations in the generation process (Li et al., 2016).

Adaptation, the domain adaptation solution for building personalized conversation systems (Zhang et al., 2017). We adapt the model in our scenario and we use the tf-idf to obtain the personal words as the user information.

CVAE, which is the conventional CVAE model and trained by KL-divergence. We change our model to use KL-divergence as the training loss.

RL-Persona, the personalized conversational system (Yang et al., 2018), which takes the advantages of deep reinforcement learning. We apply the method into our scenario as same as Adaptation.

DiaWAE-GMD, where the former is the state-of-the-art model for open-domain conversation generation (Gu et al., 2019). DiaWAE-GMD employs the Gaussian mixture prior to WAE.

4.3 Settings

The dimension of word embeddings is set to 200, which is initialized with pre-trained word2vec vectors [3]. The vocabulary is comprised of the most frequent 31,000 words. The sentence encoder and the context encoder in our PersonaWAE model are two bi-directional RNN with the GRU cells, respectively. The decoder consists of a one-layer RNN with GRUs. The hidden state sizes of both GRU encoder and decoder are set to 256. Each user is allocated a user-level vector representation with dimension size 512. We set the mini-batch size to 100. The SGD optimizer is used to train the autoencoder module with the initial learning rate 1.0, and the learning rate decay strategy is employed. We use RMSprop optimizer (Hinton et al., 2012) to update the parameters of the generator and the discriminator, where the initial learn-

[2] http://www.weibo.com/

[3] https://github.com/tmikolov/word2vec

Models	Embedding Metrics			BLEU			Human Evaluation		
	Extrema	Greedy	Average	Recall	Precision	F1	Read.	Info.	P-score
Seq2Seq	0.1640	0.4098	0.4911	0.1646	0.1646	0.1646	2.30	2.16	0.49
Persona	0.1631	0.3982	0.4871	0.1646	0.1646	0.1646	2.31	2.15	0.50
Adaptation	0.1722	0.4038	0.5113	0.1689	0.1689	0.1689	2.29	1.93	0.47
CVAE	0.2643	0.2911	0.5759	0.1931	0.1636	0.1771	2.02	2.33	0.45
RL-Persona	0.1694	0.4536	0.4906	0.1723	0.1723	0.1723	2.21	2.22	0.63
DiaWAE-GMD	0.4387	0.4752	0.7573	**0.3409**	0.1710	0.2277	2.31	2.35	0.50
PersonaWAE	**0.4542**	**0.5914**	**0.7585**	0.3365	**0.1806**	**0.2350**	**2.33**	**2.37**	**0.66**
ground-truth							2.73	2.66	0.86

Table 2: The results of both automatic evaluations and human evaluation. Read., Info., P-score refer to readability, informativeness, personalization scores. The kappa value between human annotators is 0.41, which indicates a moderate inter-rater.

# Distributions in GMD	Embedding Metrics			BLEU-F1
	Extrema	Greedy	Average	
1	0.4114	0.3110	0.7026	0.1547
2	0.4228	0.4494	0.7135	0.1543
3	**0.4542**	**0.5914**	**0.7585**	0.2350
4	0.4013	0.5764	0.7256	0.2150
5	0.4382	0.5680	0.7490	**0.2480**

Table 3: The results of different settings of k in personalization GMD, where k denotes number of distributions. When $k = 1$, it is a Gaussian distribution.

Models	Embedding Metrics			BLEU-F1
	Extrema	Greedy	Average	
PersonaWAE	**0.4542**	**0.5914**	0.7585	**0.2350**
w/o prior	0.4360	0.4848	0.7479	0.2256
w/o fusion	0.4293	0.4813	**0.7588**	0.2244

Table 4: The results of Ablation Experiments. w/o denotes without. Fusion represents fusion of personalization in decoder.

ing rates are set to $5e$-5 and $1e$-5, respectively. The gradient penalty is used for training discriminator (Gulrajani et al., 2017). The value of τ in Gumbel softmax is set to 0.1.

4.4 Evaluation Metrics

To evaluate the results of the generated responses, we adopt the following metrics widely used in existing research.

Embedding Metrics. To capture the semantic matching degrees between generated responses and ground-truth, we perform evaluations on embedding space. In consistent with previous study (Gu et al., 2019), we compute the similarity between the bag-of-words (BOW) embeddings representations of generated results and reference. In particular, we calculate three metrics:1) *Greedy* (BOW-Greedy), i.e., greedily matching words in two utterances based on the cosine similarities, and the total scores is then averaged across all words (Rus and Lintean, 2012); 2) *Average* (BOW-Average), cosine similarity between the averaged word embeddings in the two utterances (Mitchell and Lapata, 2008); 3) *Extrema* (BOW-Extrema), cosine similarity between the largest extreme values among the word embeddings in the two utterances (Forgues et al., 2014). We report the maximum BOW embedding scores of the 10 sampled responses for each testing context.

Overlap-based Metric. We utilize BLEU score (Papineni et al., 2002) to measure n-grams overlaps between ground-truth and generated response. Specifically, we follow the conventional setting in previous work (Gu et al., 2019) to compute BLEU scores using smoothing techniques (smoothing 7) [4]. For each testing context, we sample 10 responses from the models and compute their BLEU scores, i.e., n-gram precision (BLEU-Precision), n-gram recall (BLEU-Recall), and n-gram F1 (BLEU-F1).

Human Evaluation. We also employ human evaluation to assess the responses generated by our model and the baselines. Three well-educated annotators are hired to evaluate the quality of generated responses, where the evaluation is conducted in a double-blind fashion. Totally, 200 randomly

[4] http://www.nltk.org/_modules/nltk/translate/bleu_score.html

-这孩子，太稚嫩了，真想踹一脚 (This child is so immature that I really want to kick him.) -wow，你大粗腿，能抬多高啊 (Wow, how high can you get with those big legs.)	
ground truth	你要不要试试？见证一下我能办得到！ (Try it ? Prove I can do it.)
RL-Persona	哈哈，那就好了 (Haha, it's all right.)
DiaWAE-GMD	啊，完了，我就不敢说 (Game over, I can't talk it.)
PersonaWAE	**我拿小拳拳捶你** **(I'll punch you with my fist.)**
-你什么时候回香港? (When are you going back to HongKong?) -签证好了就回去(when visa is ready.)	
ground truth	要半个月嗬 (About half a month.)
RL-Persona	我也不知道，我也不知道 (I don't know, I don't know.)
DiaWAE-GMD	是因为我没在你宿舍里 (Because I'm not in your dormitory.)
PersonaWAE	**哈哈，回来约，约个时间** **(Ha, play togther when you are back.)**

Table 5: Responses generated by baselines and our model.

-求种草水乳，性价比高点的。 (Please recommend cost-effective make-up water and lotion to me.) -我水乳用的老慢啦哈哈，感觉两年用一套 (I use make-up water and lotion so slow that one can be used for 2 years.)	
User1	[UNK]，谢谢，我的小U ([UNK], thank you, my xiaoU (name))
User2	哈哈，么么哒。 (Haha, thank you and I love you.)
User3	小姐姐我买了(Ok, I buy it)
User4	是，哈哈(Yes, Haha)
-这孩子，太稚嫩了，真想踹一脚 (This child is so immature that I really want to kick him.) -wow，你大粗腿，能抬多高啊 (Wow, how high can you get with those big legs.)	
User1	我被风吹走了(I am so thin that i'm almost blown away by the wind.)
User2	我今天，小仙女(Today, I am a fairy !)
User3	哈哈，好(Haha, OK.)
User4	哈哈，我瘦了(Haha, I am thin !.)
-哈哈演不下去了 (Haha! I can't play anymore.) -早已看穿一切 (I have seen everything before.)	
User1	我看了一遍，我觉得我很帅 (I watch it again, I think I am handsome.)
User2	哈哈，好哒好，等我 (Haha, get it, wait for me.)
User3	哈哈，服(Haha, come on.)
User4	哈哈，太可怕(Haha, it's terrible.)

Table 6: Responses generated by our model for four users. We fix the context and test the effects of different user personalizations.

sampled responses generated by each model are rated by each annotator with three different aspects, i.e., readability, informativeness, and personalization. Details of the criteria are illustrated in Table 1. Note that it is very difficult to judge whether a generated response resembles with the style of the corresponding user history utterances, and thus we rate the personalization with {0,1}, representing bad or normal. Other criteria are scored from 1 to 3, i.e., bad, normal, and good.

5 Results and Analysis

In this section, we perform automatic evaluations and human evaluation to measure the quality of the generated responses quantitatively. Meanwhile, we also conduct a qualitative study to intuitively analyze the generated results. Table 2 presents the results of automatic and human evaluation.

5.1 The Effect of WAE

WAE can effectively improve the quality of responses but fails to capture personalization. As we mentioned before, we intend to improve the response quality by using WAE. Unsurprisingly, Seq2Seq gets the worst performance. Comparing DiaWAE-GMD with CVAE, we can observe that DiaWAE-GMD significantly improves BLEU scores and BOW scores upon CVAE, which is shown in Table 2. Such results indicate that the Wasserstein distance and the adversarial training can enhance model learning and address KL-vanishing issue in VAEs, as a result of which achieves better results of generated responses, which is also confirmed in the previous research (Gu et al., 2019). Besides, human evaluation results in Table 4 further illustrate that DiaWAE-GMD fails to model personalization of different users insomuch as DiaWAE-GMD lacks user-level information learning.

The number of distributions in conditional Gaussian mixture distribution significantly alter model performance. Table 3 presents the ablation results of the influence of k value in personalization Gaussian mixture distribution. It is observed that when $k \leq 3$, model performance improves with the increasing of k, which suggests that more distributions in GMD are helpful for modeling user personalization. However, for $k > 3$, model performance slightly drops with the increasing of k, The potential reason is GMD with three distributions is effective enough for modeling personalization, and sophisticated GMD might suffer scarce datasets for training.

5.2 The Influence of Personalization Modeling

User embeddings substantially improve the quality of generated response and introduce personalizations for different users. Through conducting the comparison between PersonaWAE and DiaWAE-GMD. We can learn that incorporating user personalization in decoding step substantially enhance the personalization score of human evaluations, which means user embeddings and the combination in decoder has a positive influence on response quality.

Incorporating personalization in the conditional GMD prior is more effective than combing personalization in decoder. As shown in Table 2, Persona model only achieves comparable results with Seq2Seq in terms of BLEU scores and BOW scores. For PersonaWAE and DiaWAE-GMD, incorporating personalizations in both decoder and the latent space yields a performance improvement. For the BLEU-Recall, which PersonaWAE does not outperform than DiaWAE-GMD, a possible explanation for this might be that PersonaWAE model the personalization information and generation may be more limited.

5.3 Discussion

Overall, PersonaWAE outperforms all other baselines on both automatic and human evaluations. Especially for personalization modeling, PersonaWAE achieves a noticeable achievement over the strong baselines DiaWAE-GMD and RL-Persona. These results support that our proposed PersonaWAE is effective in generating personalized response. We also observe that fusing personalized GMD and context GMD as the conditional prior is also useful, which is proven by the results shown in Table 4.

5.4 Case Study

Table 5 illustrates the generated response of different models for a given context. We can observe that our proposed model can generate responses with readability and personalization information. Table 6 shows a few example responses generated by altering the user personalization information. With different user representations, the generated responses change, which supports that personalization representation introduced in our model helps learn user-level information. Although it is difficult to evaluate the personalization of generated response and there exists a gap between generated responses and human-comprised ones, quality improvement of responses is achieved. Moreover, we observe that our proposed model might generate a too long and repeated sentence in extreme cases. The potential reason might be the relative short dialog history for each user. Besides, explicitly extracting knowledge and user personalization from conversation history is also promising. These results point out the direction of future work.

6 Related Work

Constructing an automatic conversation system is an attractive and prevalent task within the community of artificial intelligence. Previous studies mainly focus on vertical domains by applying rule- and template-based models (Pieraccini et al., 2009). Later on, with the explosive growth of data, the application of open-domain conversation model is promising. Conventional methods in vertical domains have obstacles to scale to open domain. Given this, various data-driven approaches have been proposed for modeling open-domain conversation, including retrieval-based methods (Yan et al., 2016; Tao et al., 2019), statistical machine translation model (Ritter et al., 2011), and neural networks (Serban et al., 2015; Hu et al., 2019).

Recently, building a personalized conversation system has been attached more attention, e.g., implicitly learning user personalizations from dialog history (Li et al., 2015), explicitly collecting and modeling user profiles as personalizations for generating personalized responses (Zhang et al., 2017, 2018). To improve wording diversity, CVAE models (Serban et al., 2017; Zhao et al., 2017; Shen et al., 2018) are well-investigated for open-domain response generation. As the extension of

CVAE, Wasserstein autoencoder (Gu et al., 2019) is also used for open-domain response generation to solve the issues of posterior collapse and vanishing latent variables. We build our model upon both the advantages of WAE and personalization modeling for personalized response generation.

7 Conclusion

Open domain response generation is a challenging task, which involves automatically comprising a response with informative words and personalization. Although prompting progress has been made in wording informativeness, there still exists a noticeable gap between generated response and those created by humans, especially in personalization modeling. To fill this gap, we propose a personalized Wasserstein autoencoder (PersonaWAE) for response generation, where the WAE module improves informativeness by using a continuous latent variable with GMD and user vector representations learned from dialog history is used for introducing personalization information. Experimental results on a large dataset indicate that our proposed model can generate better responses, and outperforms existing models under both automatic and human evaluations.

Acknowledgments

We would like to thank the 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. 61876196 and NSFC No. 61672058). Rui Yan was sponsored by Alibaba Innovative Research (AIR) Grant. Rui Yan is the corresponding author.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *Computer Science*.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Xiuying Chen, Shen Gao, Chongyang Tao, Yan Song, Dongyan Zhao, and Rui Yan. 2018. Iterative document representation learning towards summarization with polishing. *EMNLP*.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, workshop*, volume 2.

Nadine Glas, Ken Prepin, and Catherine Pelachaud. 2015. Engagement driven topic selection for an information-giving agent. In *SemDial*.

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

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *NIPS*, pages 5767–5777.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. 2012. Rmsprop: Divide the gradient by a running average of its recent magnitude. *Neural networks for machine learning, Coursera lecture 6e*.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. Gsn: A graph-structured network for multi-party dialogues. *arXiv preprint arXiv:1905.13637*.

Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. *stat*, 1050:10.

Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, Lei Wang, et al. 2017. Alime assist: an intelligent assistant for creating an innovative e-commerce experience. In *CIKM*, pages 2495–2498.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *ACL*.

Juntao Li, Lisong Qiu, Bo Tang, Dongmin Chen, Dongyan Zhao, and Rui Yan. 2019. Insufficient data can also rock！ learning to converse using smaller data with augmentation.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. *NAACL-HLT*, pages 236–244.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

Roberto Pieraccini, David Suendermann, Krishna Dayanidhi, and Jackson Liscombe. 2009. Are we there yet? research in commercial spoken dialog systems. In *International Conference on Text, Speech and Dialogue*, pages 3–13.

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

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162. ACL.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Building end-to-end dialogue systems using generative hierarchical neural network models. (4):3776–3783.

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.

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. *arXiv preprint arXiv:1705.00316*.

Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018. Improving variational encoder-decoders in dialogue generation. *arXiv preprint arXiv:1802.02032*.

Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology &amp; Electronic Engineering*, 19(1):10–26.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1–11.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association of Computational Linguistics*, 6:407–420.

Rui Yan, Song Yiping, and Wu Hua. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *SIGIR*, pages 55–64.

Min Yang, Qiang Qu, Kai Lei, Jia Zhu, Zhou Zhao, Xiaojun Chen, and Joshua Z Huang. 2018. Investigating deep reinforcement learning techniques in personalized dialogue generation. In *SIAM*, pages 630–638.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, pages 2204–2213.

Wei-Nan Zhang, Qingfu Zhu, Yifa Wang, Yanyan Zhao, and Ting Liu. 2017. Neural personalized response generation as domain adaptation. *WWW*, pages 1–20.

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