Exploiting Persona Information for Diverse Generation of Conversational Responses

Haoyu Song[1] , **Wei-Nan Zhang**[1,2] , **Yiming Cui**[1,3] , **Dong Wang**[3] and **Ting Liu**[1,2]

[1]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Joint Laboratory of HIT and iFLYTEK (HFL), iFLYTEK Research, Beijing, China
{hysong, wnzhang, ymcui, tliu}@ir.hit.edu.cn, dongwang4@iflytek.com,

Abstract

In human conversations, due to their personalities in mind, people can easily carry out and maintain the conversations. Giving conversational context with persona information to a chatbot, how to exploit the information to generate diverse and sustainable conversations is still a non-trivial task. Previous work on persona-based conversational models successfully make use of predefined persona information and have shown great promise in delivering more realistic responses. And they all learn with the assumption that given a source input, there is only one target response. However, in human conversations, there are massive appropriate responses to a given input message. In this paper, we propose a memory-augmented architecture to exploit persona information from context and incorporate a conditional variational autoencoder model together to generate diverse and sustainable conversations. We evaluate the proposed model on a benchmark persona-chat dataset. Both automatic and human evaluations show that our model can deliver more diverse and more engaging persona-based responses than baseline approaches.

1 Introduction

Building human-like dialogue agents to pass the Turing Test [Turing, 1950] has been a long-term goal of artificial intelligence. Open-domain conversations need to be diversified [Li *et al.*, 2015] and sustained [Li *et al.*, 2016b], as the primary goal of an open-domain chatbot is to establish a connection with the user and to accompany the user over a long period of time [Shum *et al.*, 2018]. Because of the vast quantity of rational responses for an open-domain conversational message, sequence-to-sequence (Seq2Seq) model [Sutskever *et al.*, 2014] has been widely used for conversation modeling [Vinyals and Le, 2015; Shang *et al.*, 2015; Serban *et al.*, 2016]. Seq2Seq model enables the incorporation of rich context when mapping between consecutive dialogue turns and is trained by predicting one target response in a given dialogue context using the maximum-likelihood estimation (MLE) objective function.

Chatbot's Persona	
1. I am a soccer player	4. Nike cleats are my favorite
2. my number is 42	5. I joined a new team last week
3. I'm a goalie	

	User Input	What do you do for a living?
Response	Non-Specific	I don't know what you mean
	Potential Candidates	r_1: I am a soccer player, and you?
		r_2: Emm... Tell me yours first
		r_3: I am a goalie in the soccer team
		r_4: I just joined a new soccer team

Table 1: Given an input and some persona texts, there exist many non-repetitive responses, e.g., grounded on different persona texts.

Despite the successful application of Seq2Seq in dialogue modeling, it is still quite impossible for current dialogue agents to pass the Turing Test, while one of the reasons is the lack of a coherent persona [Vinyals and Le, 2015]. Endowing an open-domain chatbot with persona is challenging, yet impactful to deliver more realistic and engaging conversations. First, presenting a coherent persona is crucial for a chatbot to gain confidence and trust from users [Li *et al.*, 2016a; Qian *et al.*, 2017]. Second, grounded on predefined persona, a chatbot can be trained to both ask and answer questions about personal topics, thus making users more engaged [Zhang *et al.*, 2018]. Due to these reasons, there has been a growing interest in modeling persona within the Seq2Seq model.

However, human dialogue has a unique property that many other learning tasks do not: given a dialogue context, there may exist **more than one** valid response[1] [Zhao *et al.*, 2017; Rajendran *et al.*, 2018]. As illustrated in Table 1, the potential candidate responses **cover different aspects** and are **non-repetitive**. Such responses can not only provide more candidates for different dialogue strategies [Yu *et al.*, 2016] but also enrich the diversity. Nevertheless, the *one-to-many* property is not taken into account in the present persona dialogue models. Although recent Seq2Seq-based persona models were successful, they still suffer from the *low diversity* problem [Jiang and de Rijke, 2018] brought by Seq2Seq. They tend to generate repetitive and non-specific responses like *"I don't know"* [Li *et al.*, 2015; Li *et al.*, 2016b], regard-

---

[1]Denoted as *one-to-many* property.

less of the conversational context.

In this paper, we follow the definition of persona in Zhang *et al.* [2018] (as exemplified in Table 1), and work on the explicit textual persona. This is because the unstructured persona data is easier to collect and more authentic[2]. We propose a memory-augmented architecture to exploit the explicit persona texts and incorporate a latent variable that can capture response variations under the conditional variational autoencoder model, rather than learn to predict only one response using MLE. To generate diverse persona-based responses, our model draws samples from the latent variable and combines two different decoding strategies to generate words or copy words directly from persona texts. We evaluate our model on the ConvAI2 persona-chat dataset and yield promising results in (i) incorporating persona information into responses and (ii) generating diverse and engaging responses. In summary, the contributions of this paper are three-fold:

First, we propose to address the *one-to-many* property in end-to-end model. To the best of our knowledge, this is the first work that jointly models persona text and response diversity.
Second, we present a novel end-to-end model that incorporates persona texts into diverse responses (Section 3).
Third, experimental results show that our model can generate persona-based responses as well as deliver more diverse and more engaging responses than baselines (Section 4).

2 Related Work

This work is related to the recent advancements in persona-based dialogue models and the efforts of improving diversity in dialogue generation.

In recent years, the dialogue research community has shown a growing interest in the modeling of personality. Li *et al.* [2016a] first incorporated implicit persona in dialogue generation using user embeddings, which projects each user into a dense vector. Kottur *et al.* [2017] proposed a neural dialogue model which conditioned on both speakers and context history. However, these two models rely heavily on speaker-tagged dialogue data, which is more expensive and sparser. To present a coherent personality, Qian *et al.* [2017] defined several profile key-value pairs, including name, gender, age, location, etc., and explicitly expressed a profile value in the response. Recent works brought new persona models as well as high-quality data. Zhang *et al.* [2018] contributed a persona-chat dataset, and they further proposed two generative models, *persona-Seq2Seq* and *Generative Profile Memory Network*, to incorporate persona information into responses. Yavuz *et al.* [2018] explored the use of copy mechanism in persona-based dialogue models. Our work follows the line initiated by Zhang *et al.* [2018] but makes a step further by generating diverse persona-based responses.

On the other hand, large-scale dialogue generation still suffers from the tendency of generating dull and generic responses. The efforts to tackle the low diversity problem fall into two major categories. The first focuses on the objective function of the dialogue generation model, and they argue that the MLE objective is unable to approximate the real-world

goal of the conversation [Li *et al.*, 2015; Li *et al.*, 2016b]. The other line of research tries to augmenting the encoder of the generative model with richer information [Xing *et al.*, 2016; Zhao *et al.*, 2017]. As one of these improvements, Zhao *et al.* [2017] proposed a dialogue model adapted from conditional variational autoencoders [Yan *et al.*, 2016]. The dialogue CVAE learns a latent variable that captures discourse-level variations and generates diverse responses by drawing samples from the learned distribution.

We also borrow ideas from Sukhbaatar *et al.* [2015] in the modeling of external memory. The proposed Persona-CVAE can be seen as an extension of dialogue CVAE, with several main differences: (i) the combination with external persona memory, (ii) the decoding strategies, and (iii) the ability to deliver persona-based responses.
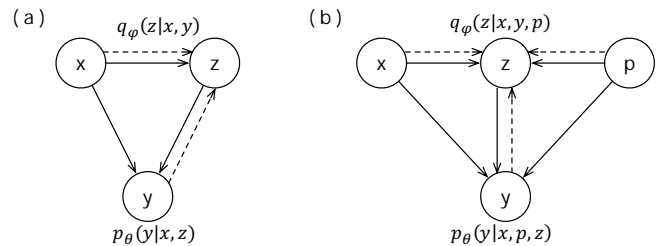

Figure 1: Illustration of the conditional graphical models: (a) dialogue CVAE by Zhao and (b) the proposed Persona-CVAE. Solid lines denote the generative model with parameter θ, dashed lines denote the variational approximation with parameter φ. x, y, z and p denote input, response, latent variable and persona respectively.

3 Persona-CVAE

In this section, we describe the components of the proposed Persona-CVAE model in detail. As illustrated in Figure 1 (b), the conditional graphical model shows the core idea of proposed Persona-CVAE. We assume that the generation of z depends on x, y and p. y relies on x, p and z. The main objective is to learn the conditional probability $q_{\varphi}(z|x,p)$ and $p_{\theta}(y|x,p,z)$, with parameters φ and θ. The variational parameters φ are learned jointly with the generative model parameters θ, as described in Section 3.2.

3.1 Task Definition

The task can be formally defined as: given an input message X and a set of persona texts $P = \{P_1, P_2, ..., P_n\}$, the goal is to generate diverse responses $\hat{Y} = \{\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_m\}$, based on both input message and persona texts.

The **diverse** in responses means that the utterances in $\hat{Y}$ are non-repetitive, cover different aspects and differ in words.

3.2 Persona Based CVAE

The proposed model uses four random variables to represent each dyadic conversation: the dialogue context X, the target response Y, the set of persona texts P and a latent variable z. z is used to capture the latent distribution over the valid responses. Further, X is composed of both input messages

---

[2] For example, 80% of posts on Twitter are about personal emotional state, thoughts or activities [Naaman *et al.*, 2010].
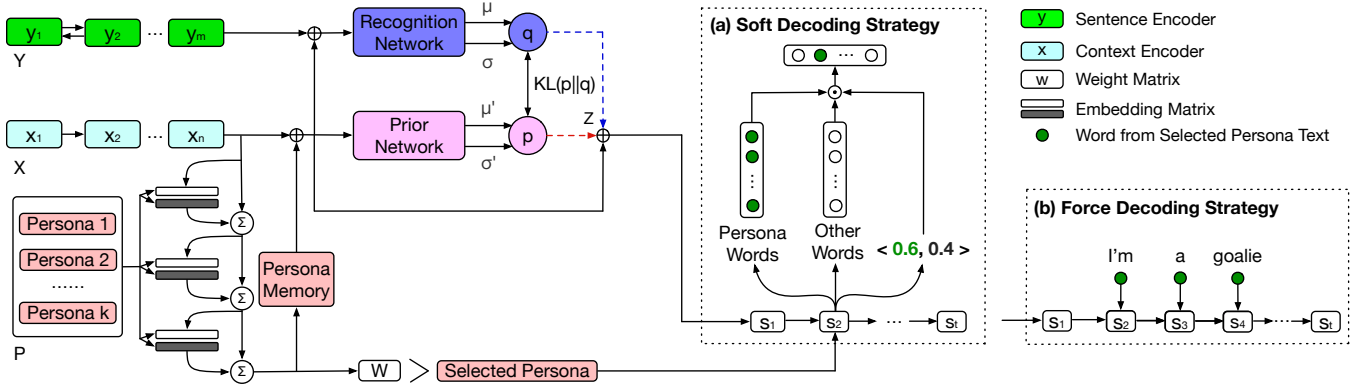
Figure 2: The neural network architecture of the proposed Persona-CVAE model. The persona memory comes from the multi-hops attention over persona texts and dialogue context. The dashed blue connection only appears in the training process, and the dashed red connection only appears in the generation process. The $\oplus$ denotes the concatenation of input vectors.

and responses in dialogue history and can be denoted as $X = \{X_1, X_2, ..., X_n\}$, where $X_i = x_{i,1}x_{i,2}...x_{i,k}$ ($x_{i,j}$ is a single word). Similarly, P consists of several unstructured persona texts: $P = \{P_1, P_2, ..., P_k\}$, where $P_i = p_{i,1}p_{i,2}...p_{i,l}$.

Figure 2 demonstrates an overview of our model. The sentence encoder is a bidirectional recurrent neural network (BRNN) [Schuster and Paliwal, 1997] and is used to encode a single sentence (e.g., the target response). We concatenate the last hidden state of both forward and backward RNN to capture the semantic information from both sides, formulized as $h = [\overrightarrow{h_T}, \overleftarrow{h_T}]$, where T is the sequence length. For the context encoder, we use a hierarchical encoding strategy. First, each sentence X_i in context X is encoded by the sentence encoder to get a latent representation h_{sent_i}. Then a single layer forward RNN is used to encode the sentence representations into a final state $h_{context}$.

Persona differs from dialogue context in two main aspects: (i) it keeps unchanged throughout the conversation, and (ii) it is only the unilateral information (here is for the chatbot), while two sides share the same dialogue context. Therefore, we conjecture that modeling persona differently from dialogue context, with a memory augmented architecture, will be beneficial for the model's performance. The details will be discussed in Section 3.3.

Then we define the conditional distribution over the above random variables. As mentioned, we define the generation process as a conditional distribution $p(y, z|x, p) = p(y|x, p, z)p(z|x, p)$ and our objective is to approximate $p(y|x, p, z)$ and $p(z|x, p)$ with deep neural networks. Following the definition in Zhao *et al.* [2017], we refer to $p(y|x, p, z)$ as the *response decoder* and $p_\theta(z|x, p)$ as the *prior network*. Further, to approximate the true posterior distribution, we refer to $q_\varphi(z|x, y, p)$ as the *recognition network*.

CVAE is trained to maximize the conditional log-likelihood, but this involves an intractable marginalization over the latent variable. Previous works [Sohn *et al.*, 2015; Yan *et al.*, 2016] have shown that CVAE can be efficiently trained with the *Stochastic Gradient Variational Bayes* (SGVB) [Kingma and Welling, 2013] by maximizing the *variational lower bound* of the conditional log-likelihood.

As proposed in Zhao *et al.* [2017], we also assume the latent variable z follows multivariate Gaussian distribution with a diagonal covariance matrix. And the *variational lower bound* of Persona-CVAE can be written as:

$$
\begin{aligned}
\mathcal{L}(\theta, \varphi; x, y, p, z) = &-KL(p_\theta(z|x, p)||q_\varphi(z|x, y, p)) \\
&+ E_{q_\varphi(z|x,y,p)}(\log p(y|x, p, z)) \\
\leq &\log p(y|x, p)
\end{aligned}
\tag{1}
$$

Since we assume latent variable z follows isotropic Gaussian distribution, the recognition network $q_\varphi(z|x, y, p) \sim \mathcal{N}(\mu, \sigma^2 \mathbf{I})$ and the prior network $p_\theta(z|x, p) \sim \mathcal{N}(\mu', \sigma'^2 \mathbf{I})$. We sample z either from $\mathcal{N}(\mu, \sigma^2 \mathbf{I})$ in training or $\mathcal{N}(\mu', \sigma'^2 \mathbf{I})$ in testing. To make the sampling operation differentiable, we use the *reparametrization trick* [Kingma and Welling, 2013] and we have:

$$
\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = W_{recog} \begin{bmatrix} x \\ y \\ p \end{bmatrix} + b_{recog}
\tag{2}
$$

$$
\begin{bmatrix} \mu' \\ \sigma'^2 \end{bmatrix} = W_{piror} \begin{bmatrix} x \\ p \end{bmatrix} + b_{piror}
\tag{3}
$$

We project the concatenation of x,p and z to a vector, to serve as the initial state of *response decoder*. Then the *response decoder* predicts words under the Decoding Strategies (Section 3.4).

3.3 Persona Memory

The persona memory module has two roles: first is to encode the persona texts $P = \{p_1, p_2, ..., p_k\}$ into a dense vector (persona memory) and sencond is to select a persona p_i to be expressed in the response. Notice that persona is not always needed in the response, and thus model selects persona from $\{p_1, p_2, ..., p_k\} \cup \{None\}$.

The **first role** of the persona memory module is defined by multi-hops attention over persona texts and dialogue context. We convert the persona texts into memory vectors $\{m_1, m_2, ..., m_k\}$ of dimension d, computed by embedding each p_i in a continuous space using an embedding matrix A of size $d \times V$. We use the embedded dialogue context $h_{context}$

(with the same dimensions d) as an input. We compute the match between context vector $h_{context}$ and each memory vector m_i in the embedding space by:

$$prob_i = \textbf{softmax}(h_{context}^{\mathrm{T}} m_i) \qquad (4)$$

where $\textbf{softmax}(x_i) = e^{x_i} / \sum_j e^{x_j}$. Meanwhile, each persona p_i has an output vector c_i (by another embedding matrix C). A single hop output o is computed by a sum over the output vector c_i, weighted by the probability from equation (4):

$$o = \sum_i prob_i c_i \qquad (5)$$

And k-hops attention are stacked in the following way:

$$u^k = u^{k-1} + o^{k-1} \qquad (6)$$

where $u^0 = h_{context}$. For the embedding matrix, we use an adjacent strategy, i.e. $A^{k+1} = C^k$. Finally, the u^3 is used as our persona memory[3].

The **second role** of the persona memory module is to decide which persona should be expressed in the generated response. This is implemented as:

$$\alpha_i = \textbf{softmax}(\mathbf{W_p}[u^3, z]) = \textbf{MLP}([u^3, z]) \qquad (7)$$

where $\mathbf{W_p}$ is a weight matrix, and u^3 is the persona memory. z is the sampled latent variable, as described in Section 3.2. The latent variable captures some information that is not presented in the dialogue context and is beneficial for selecting the optimal persona. As aforementioned, in some cases, there is no need to incorporate persona information, so we need an MLP here to make the decision. The model selects a persona with the maximal probability, i.e. $\hat{p} = p_j$, where $j = \textbf{argmax}_i(\alpha_i)$. Then the words from selected persona are further used in *response decoder*.

3.4 Decoding Strategy

We propose two decoding strategies (general purpose and special cases) to better express persona information.

Soft Decoding Strategy (SDS)

The *soft decoding strategy* assumes that at each decoding step t, there is a distribution α^t over the two types {*persona, other*}. Similarly, according to the selected persona words, the entire vocabulary set is also divided into two parts, i.e. {*persona words, other words*}.

At time step t, the *response decoder* first estimates word generation probabilities over the two vocabulary sets independently, denoted as $\mathcal{P}_{per}^t$ and $\mathcal{P}_{other}^t$:

$$\mathcal{P}_{per}^t = \textbf{softmax}_{per}(w_{dec}s_t + b_{dec}) \qquad (8)$$

$$\mathcal{P}_{other}^t = \textbf{softmax}_{other}(w_{dec}s_t + b_{dec}) \qquad (9)$$

and then computes the type distributions:

$$\begin{bmatrix} \alpha_{per}^t \\ \alpha_{other}^t \end{bmatrix} = \textbf{softmax}(w_{sds} \begin{bmatrix} s_t \\ u^3 \end{bmatrix} + b_{sds}) \qquad (10)$$

where s_t is the hidden state of decoder RNN cell. The final probability of generating a word is a mixture of type-specific generation distributions where the coefficients are type probabilities:

$$\mathcal{P}(y_t | Y_{<t}, X, P) = \alpha_{per}^t \mathcal{P}_{per}^t \cup \alpha_{other}^t \mathcal{P}_{other}^t \qquad (11)$$

[3]In our experiments, $k = 3$ performs better than $k = 1$ or 2, but there is no significant performance change when $k = 4$ or 5.

Force Decoding Strategy (FDS)

We propose *force decoding strategy* from the observation that persona texts can directly serve as a valid response under some circumstances. In the decoding process, if the last words of the partially decoded sequence are the same as the first words of selected persona, the input words to decoder RNN cell will come from the selected persona words successively in next few steps:

$$h_{t+1} = \textbf{RNN}_{dec}(h_t, \textbf{embedding}(p_i)) \qquad (12)$$

where p_i is from the selected persona. At time step t, p_i is also the output word. The decoder uses FDS only once in a decoding process. Though this strategy is simple, we found it works well in some situations.

3.5 Training and Optimization

The Persona-CVAE is trained to maximizing the *variational lower bound* of the conditional log-likelihood in Formula (1). Persona memory module and type distribution are trained by two standard cross-entropy loss function.

As addressed in Bowman *et al.* [2015], the *KL annealing* trick is necessary for the training of RNN-based CVAE. Besides, another essential technique for the model to work is the *bag-of-word loss* [Zhao *et al.*, 2017]. Finally, the Persona-CVAE model is trained with the sum of these losses and optimized through backpropagation.

4 Experiment

4.1 Persona-Chat Dataset and Preparation

We perform experiments on the recently released ConvAI2 benchmark dataset, which is an extended version (with a new test set) of persona-chat dataset [Zhang *et al.*, 2018]. The conversations are obtained from crowdworkers who were randomly paired and asked to act the part of a given persona.

This dataset contains 164,356 utterances in over 10,981 dialogues and has a set of 1,155 personas, each consisting of at least four profile texts. The testing set contains 1,016 dialogues and 200 never seen before personas. We set aside 800 dialogues together with its profile texts from the training set for validation. The final data have 9,181/800/1,016 dialogues for train/validate/test.

To train the persona memory module, we label each utterance with its corresponding persona. We first compute word inverse document frequency: $idf_i = 1/(1 + \log(1 + tf_i))$, where tf_i is from the GloVe index via Zipf's law [Zhang *et al.*, 2018]. If the utterance has the highest *tf-idf* similarity with a profile sentence and the similarity is higher than a threshold, then we label the utterance with this profile sentence. Otherwise, the utterance's persona label is none. We also label the position that shares a word with profile sentence to learn the type distribution.

4.2 Baselines

We compared the proposed Persona-CVAE model with five state-of-the-art generative baseline models. The five models fall into two categories, i.e., persona-free model and persona-based model:

Methods	N = 1			N = 5			N = 10		
	Dtinct-1	Dtinct-2	P. Cover	Dtinct-1	Dtinct-2	P. Cover	Dtinct-1	Dtinct-2	P. Cover
Seq2Seq	.0125	.0464	.0026	.0031	.0142	.0057	.0018	.0089	.0071
CVAE	.0366	.2080	.0021	.0090	.0875	.0029	.0050	.0663	.0048
Per.-Seq2Seq	.0159	.0745	.0091	.0036	.0213	.0217	.0021	.0139	.0193
GPMN	.0179	.0738	.0080	.0045	.0195	.0184	.0027	.0103	.0178
Oracle. Copy	.0203	.0830	**.0181**	.0050	.0276	.0273	.0031	.0189	.0247
Per.-CVAE	**.0383**★	**.2088**	.0167	**.0120**★	**.1037**★	**.0410**★	**.0075**★	**.0779**★	**.0395**★

Table 2: Automatic evaluation results in diversity (*Dtinct-1&2*) and persona integration (*P. Cover*). N is the number of generated responses in each turn. Numbers with the ★ mean that the result is statistically significant over all baselines (2-tailed t-test, p-value<0.01).

Seq2Seq. a general persona-free model with context attention mechanism [Shang *et al.*, 2015].

Dialogue CVAE. a persona-free model that uses a latent variable to learn a distribution over potential conversational intents and improves the discourse-level diversity of responses [Zhao *et al.*, 2017].

Persona-Seq2Seq. a persona-based model that prepends persona texts to the input sequence x, i.e., $x = \forall p \in P||x$,where $||$ denotes concatenation [Zhang *et al.*, 2018].

Generative Profile Memory Network (GPMN). a persona-based generative model that encodes each profile sentence as individual memory representations in a memory network [Zhang *et al.*, 2018].

Oracle Seq2Seq+Copy (Oracle. Copy). a persona-based model with copy mechanism. This model is similar to Persona-Seq2Seq, but it selects the persona text whose tokens have the highest unigram *tf-idf* similarity to the **ground truth** response (the *Oracle*). It achieves the best results in a variety of copy-based persona models [Yavuz *et al.*, 2018]. The aim of this experiment is to have a better comparison with the copy-based persona models.

Among these baselines, we fine-tuned Seq2Seq, Dialogue CVAE and Oracle Seq2Seq+Copy on the ConvAI2 persona-chat dataset, while the other two methods are using the latest released models from Zhang *et al.* [2018] in ParlAI. Moreover, to make different models comparable, we generate N responses from the Seq2Seq based models by sampling from the softmax. For the latent variable based models, we sample N times from the latent z to generate responses.

4.3 Experimental Settings

In our experiments[4], the RNN is two-layer GRU with a 500-dimensional hidden state. The dimension of word embedding is set to 300, and thus the persona memory size is also 300. The vocabulary size is limited to 20,000. The latent variable size is set to 100. KL annealing steps are set to 10,000. We train the model with a minibatch size of 32 and use Adam optimizer with an initial learning rate of 0.001. All parameters are initialized by sampling from a uniform distribution.

4.4 Automatic Evaluation

Automatically evaluating an open-domain generative dialogue model is still an unsolved research challenge. Although metrics such as BLEU and perplexity have been

---

[4]Code available at: https://github.com/vsharecodes/percvae .

used for dialogue quality evaluation [Vinyals and Le, 2015; Serban *et al.*, 2016], recent work has found that BLEU shows very weak correlation with human judgment [Liu *et al.*, 2016]. Further, perplexity is not computed on a per-response basis. Since the goal of the proposed model is not to predict only one target response, but rather diverse persona-based responses, we do not employ BLEU or perplexity for evaluation. Following the *one-to-many* property, we use two metrics to evaluate how well our model enhancing diversity and incorporating persona information:

Distinct-K (Dtinct-K). this metric calculates the number of distinct k-grams in generated responses and is scaled by the total number of generated tokens to avoid favoring long responses [Li *et al.*, 2015]. The Distinct-1 and Distinct-2 are thus the token ratios for unigrams and bigrams. This metric is an indicator of word-level diversity for generated responses.

Persona Coverage (P. Cover). we propose this metric to evaluate how well persona information is expressed. We assume that there are M predefined persona texts $\{p_1, p_2, ..., p_M\}$ and N generated responses $\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_N\}$. The response-level persona coverage is defined as:

$$\mathcal{C}_{per} = \frac{\sum_i^N \mathbf{max}_{j \in [1,M]} \mathcal{S}(\hat{y}_i, p_j)}{N} \quad (13)$$

where $\mathcal{S}(\hat{y}_i, p_j)$ is the counts of shared words weighted by words' *idf* (as described in Section 4.1). Assume that the words' set shared by $\hat{y}_i$ and p_j are W and the *inverse document frequency* for $w_k \in W$ is f_k, then we have:

$$\mathcal{S}(\hat{y}_i, p_j) = \frac{\sum_{w_k \in W} (f_k)}{|W|} \quad (14)$$

and $|W|$ denotes the number of words in W.

The final scores of two metrics are averaged over the entire test dataset, and we report the performance in Table 2. As can be seen, even if N=1, our model still outperforms all the baselines in diversity but is inferior to the Oracle model in the *Persona Coverage*. When N=5 and 10, the proposed model obtains the best performance in all metrics compared with all the baseline models. Due to the number of persona texts in each dialogue (generally 4 or 5), generating too many responses leads to a small decline in the *Persona Coverage*.

The results show that under the different number of N, our model always generates diverse responses as well as effectively incorporates persona information into multiple responses. To better evaluate the diversity of generated responses, the following experiments are all based on N=5.

Methods	Engage.	Variety	Per. Detection
Seq2Seq	48.91%	26.07%	44.57%
CVAE	55.43%	88.04%	41.30%
Per.-Seq2Seq	64.13%	29.35%	69.57%
GPMN	67.39%	30.43%	65.76%
Oracle. Copy	63.26%	32.61%	72.83%
Per.-CVAE	**71.74%**	**90.22%**	**78.26%**

Table 3: Human evaluation results (N=5). As a reference, the persona detection rate of the sampled testing data (human) is 90.83%.

Pref. (%)	Per.-Seq2Seq	GPMN	Per.-CVAE
Per.-Seq2Seq	-	41.86	32.56
GPMN	58.14	-	39.53
Per.-CVAE	**67.44**	**60.47**	-

Table 4: Pairwise preference of three persona-based models (N=5).

4.5 Human Evaluation

We also explore two settings for human evaluation. In both settings, we present persona texts, input message, as well as N generated responses.

In the first setting, we employ judges to evaluate a random sample of 100 items per model according to three metrics, based on a 1/0 scoring schema, similar as Qian *et al.* [2017]:

Engagingness (Engage). a response is engaging only when it is appropriate, interesting and easy to answer. This is the overall score of N generated responses.

Variety. this metric measures the variety of N generated responses. Score 1 indicates that there is a significant difference in the linguistic patterns and wordings and score 0 otherwise.

Persona Detection. we measure the model's ability to incorporate persona information by displaying two possible persona texts: one is the true persona that the model used and another is randomly sampled from the rest of persona texts. Then the judges will determine which is more likely to be the persona used by model according to N generated responses.

We calculated the Fleiss' kappa to measure inter-rater consistency. Fleiss' kappa for *Engagingness*, *Variety* and *Persona Detection* is 0.5149, 0.8281 and 0.6656, indicating "moderate agreement", "almost perfect" and "substantial agreement" respectively.

Results in Table 3 support: **First**, Our model performs better than all baselines in all subjective metrics. **Second**, Our model successfully incorporates persona information. **Third**, Our model delivers diverse and engaging responses.

In addition, we performed a **Preference Test** in three persona-based models. In this setting, the judges are presented with pairwise items and are asked to decide which of the two N responses is of higher quality. And a partial ordering relation about the overall quality is given: best response quality > response variety > persona coverage. The judges first need to do a re-rank in mind and decide which model offers a better top response. If no difference then considering variety in responses, otherwise whether persona text is expressed. As shown in Table 4, the proposed model is significantly (2-tailed t-test, $p < 0.01$) preferred than the other

	Chatbot's Persona:	
Mem	I like tacos I have one brother	
	I love dogs I have three sisters	
1st Turn	**U:** Hi, how's it going this afternoon ?	
	r_1: I have been doing pretty good . You?	
	r_2: I am tired from working a day.	
	r_3: Good. I just got done *eating some tacos*.	
	r_4: I am watching Battlestar Galactica. How about you ?	
2nd Turn	**U:** Good. Tired from walking dogs early in the morning.	
	r_1: Wow! What are you into?	
	r_2: I *love dogs* and I work in the fields.	
	r_3: I am good at that. Do you have a job?	
	r_4: Ok. I *like tacos*, and I am in Florida. Where are you from?	

Table 5: Generated examples of diverse responses in two consecutive turns. The italics show the incorporated persona. Mem denotes persona memory (in textual description) and U denotes user input.

Methods	Dtinct-1	Dtinct-2	P. Cover
Per.-CVAE	.0120	**.1037***	**.0410***
w/o SDS	**.0131**	.0921	.0051
w/o FDS	.0107	.1021	.0358
w/o SDS & FDS	.0119	.0907	.0042

Table 6: The ablation tests (N=5). Results in *Dtinc-2* and *P.Cover* are significantly different from the ablated models (p-value<0.01).

two baselines.

Finally, we present some generated examples in Table 5.

4.6 Ablation Tests

In order to investigate the influence of different decoding strategies, we also conducted ablation tests where one or two decoding strategies were removed from Persona-CVAE. The results are shown in Table 6. As we can see, after removing the SDS, the *Persona Coverage* decreases the most, indicating the SDS contributes to a higher persona information coverage. However, SDS also shows a negative effect on *Distinct-1*. This may be because the SDS encourages the use of words from profile sentences, which leads to low-frequency words more difficult to appear.

5 Conclusion and Future Work

In this paper, we focus on the diverse generation of conversational responses based on chatbot's persona and propose a memory-augmented architecture named Persona-CVAE. We conduct experiments on the ConvAI2 dataset. Experimental results show that our model outperforms the state-of-the-art methods, especially in terms of diversity and persona integration. For future work, we will explore modeling the persona information of users in open-domain conversations.

Acknowledgments

References

[Bowman *et al.*, 2015] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *Computer Science*, 2015.

[Jiang and de Rijke, 2018] Shaojie Jiang and Maarten de Rijke. Why are sequence-to-sequence models so dull? understanding the low-diversity problem of chatbots. *arXiv preprint arXiv:1809.01941*, 2018.

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

[Kottur *et al.*, 2017] Satwik Kottur, Xiaoyu Wang, and Vítor Carvalho. Exploring personalized neural conversational models. In *IJCAI*, pages 3728–3734, 2017.

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

[Li *et al.*, 2016a] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[Li *et al.*, 2016b] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[Liu *et al.*, 2016] Chia Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

[Naaman *et al.*, 2010] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM, 2010.

[Qian *et al.*, 2017] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Assigning personality/identity to a chatting machine for coherent conversation generation. *arXiv preprint arXiv:1706.02861*, 2017.

[Rajendran *et al.*, 2018] Janarthanan Rajendran, Jatin Ganhotra, Satinder Singh, and Lazaros Polymenakos. Learning end-to-end goal-oriented dialog with multiple answers. *arXiv preprint arXiv:1808.09996*, 2018.

[Schuster and Paliwal, 1997] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

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

[Shang *et al.*, 2015] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.

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

[Sohn *et al.*, 2015] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, 2015.

[Sukhbaatar *et al.*, 2015] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *NIPS*, pages 2440–2448, 2015.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[Turing, 1950] A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

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

[Xing *et al.*, 2016] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic augmented neural response generation with a joint attention mechanism. arxiv preprint. *arXiv preprint arXiv:1606.08340*, 2(2), 2016.

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

[Yavuz *et al.*, 2018] Semih Yavuz, Abhinav Rastogi, Guanlin Chao, and Dilek Hakkani-Tür. Deepcopy: Grounded response generation with hierarchical pointer networks. *NIPS*, 2018.

[Yu *et al.*, 2016] Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. Strategy and policy learning for non-task-oriented conversational systems. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 404–412, 2016.

[Zhang *et al.*, 2018] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th ACL*, pages 2204–2213, 2018.

[Zhao *et al.*, 2017] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th ACL*, pages 654–664, 2017.