Guiding Variational Response Generator to Exploit Persona

Bowen Wu[1], Mengyuan Li[2]*, Zongsheng Wang[1], Yifu Chen[3]*, Derek F. Wong[4],
Qihang Feng[1], Junhong Huang[1], Baoxun Wang[1]

[1]Platform and Content Group, Tencent
[2]Peking University, Beijing, China
[3]University of Chinese Academy of Sciences
[4]NLP[2]CT Lab / Department of Computer and Information Science, University of Macau

{jasonbwwu,jasoawang,careyfeng,vincenthuang,asulewang}@tencent.com
limengyuan@pku.edu.cn, chenyifu17@mails.ucas.ac.cn, derekfw@um.edu.mo

Abstract

Leveraging persona information of users in Neural Response Generators (NRG) to perform personalized conversations has been considered as an attractive and important topic in the research of conversational agents over the past few years. Despite of the promising progress achieved by recent studies in this field, persona information tends to be incorporated into neural networks in the form of user embeddings, with the expectation that the persona can be involved via End-to-End learning. This paper proposes to adopt the personality-related characteristics of human conversations into variational response generators, by designing a specific conditional variational autoencoder based deep model with two new regularization terms employed to the loss function, so as to guide the optimization towards the direction of generating both persona-aware and relevant responses. Besides, to reasonably evaluate the performances of various persona modeling approaches, this paper further presents three direct persona-oriented metrics from different perspectives. The experimental results have shown that our proposed methodology can notably improve the performance of persona-aware response generation, and the metrics are reasonable to evaluate the results.

1 Introduction

As an essential research topic in generative conversational agents (a.k.a., chat-bots), Persona Modeling is of great importance for such deep neural network based intelligent interactive systems (Li et al., 2016b; Kottur et al., 2017; Wang et al., 2017). Apparently, user-personality-dependent responses provided by a chat-bot are able to significantly improve the consistency of its conversations, meanwhile, it is possible for users

---
* Contribution during the internship at Tencent.

to flexibly customize the persona of a chat-bot based on some existent dialogues. As for the studies on this topic, with no doubt, incorporating persona factors into End-to-End generative models is an attractive topic with great challenges.

The current studies mainly focus on adopting the explicit meta-data of user profiles (Qian et al., 2018; Chu et al., 2018) or character descriptions (Zhang et al., 2018; Mazare et al., 2018; Song et al., 2019) to generate persona-aware responses. However, on one hand, user profiles are usually highly privacy-related and thus it is difficult to obtain such information from users practically. On the other hand, little correlation can be explicitly observed between such meta-data profiles and persona characteristics of users. Especially, those character descriptions, tailor-made for the persona-aware response generation with the great cost of manual work, are only a variant of user profile innately in terms of different natural language forms.

One of the reasonable and practically executable methodologies for introducing persona factors into conversation models is to adopt the real-valued user representation as a medium (Li et al., 2016b; Kottur et al., 2017; Liu et al., 2018; Al-Rfou et al., 2016). In particular, such user representations can be derived from users' historical dialog utterances with rich linguistic and personality information involved. Taking persona representations as the guidance for generating customized responses becomes a widely accepted methodology due to the recent development of deep latent variable models (Zhao et al., 2017; Shen et al., 2017; Zhou and Wang, 2018). However, for current models, without the explicit learning objectives or constraints, the user representation is adopted in a passive way to reduce the model loss and KL divergence via end-to-end learning. In this case, it is highly possible that the
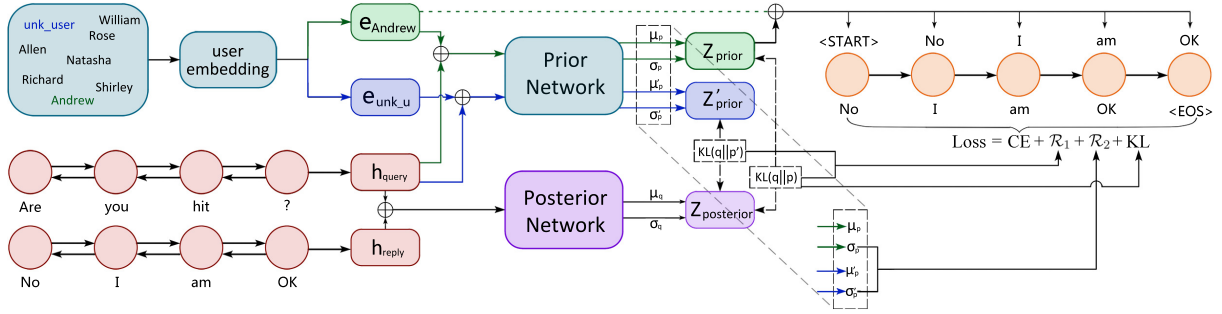
Figure 1: The architecture of the Persona-Aware Variational Response Generator (PAGenerator) described in this paper. $\oplus$ represents the concatenation of inputs and CE denotes the cross-entropy of predictions. The dotted arrow line indicates the connection is optional, and the default model named PAGenerator decodes with the user embedding.

employed embeddings will not work as effectively as expected.

Consequently, it is necessary to employ explicit guidance to help variational response generators sense persona. From observations upon persona-contained dialogs, there exist intuitive characteristics for directing the optimization of the persona-aware variational response generation. Obviously, for a given user, the appropriately modeled and leveraged persona information can help to generate hidden variables semantically relevant with corresponding responses. Besides, since users may have their own linguistic style, the adoption of personal information in NRG aims to have direct influence on the degree of linguistic (e.g. lexical and syntactic) convergence for a specific user.

This paper aims at exploring the explicit guidance to help the variational response generator exploit persona information hidden in the non-structured contents produced by the users, by utilizing intuitive characteristics of personalized conversations for model training. The contributions of this paper can be summarized as follows:

- A persona-aware variational response generator is proposed to exploit persona while modeling the conversations.

- Based on the model, two regularization terms are presented to guide the model in encoding user information into the latent variables and converging to user-specific responses.

- Three discriminative metrics are further introduced to evaluate the capabilities of persona-aware response generators.

2 Approach

Based on the current progress on the development of latent variable models, we propose a persona-aware variational response generator to automatically exploit persona from conversations, and utilize such personal information to model the future conversation. Besides, given that personal information can be exploited as optimization guidance to better modeling persona, we further introduce two regularization terms to guide the model learning. In the following section, we first describe the general structure of PAGenerator, and then explain the two additional regularization terms.

2.1 Persona-Aware Variational Response Generator

Utilizing latent variables in response generation has become a widely accepted methodology in NRG due to their Bayesian essence. It helps to deal with external knowledge efficiently, e.g. Persona. Therefore, our proposed model is built based on the generation model with latent variables. The overall architecture of the single turn persona-aware variational response generator proposed in this paper is illustrated in Figure 1.

Let q, r, u stand for the query, the reply and the corresponding user of r, respectively, and e_u stands for the embedding of user u. A bidirectional LSTM is first employed to encode the query and reply into fixed size vectors h_q and h_r. After that, the prior network (parametrized by θ) takes u_e, h_q as inputs to generate the distribution $p_\theta(z|q, u)$ of latent variable z. Meanwhile, h_q, h_r are fed into a posterior network (parameterized by ϕ) to compute $q_\phi(z|q, r)$. As we adopt the assumption that z follows isotropic Gaussian distribution, $p_\theta(z|q, u)$ and $q_\phi(z|q, r)$ are also normally

54

distributed, such that:

$$p_\theta(z|q,u) \sim \mathcal{N}(\mu_p, \sigma_p^2 \mathbf{I})$$
$$q_\phi(z|q,r) \sim \mathcal{N}(\mu_q, \sigma_q^2 \mathbf{I}) \qquad (1)$$

where the means and variances are computed as follows:

$$\begin{bmatrix} \mu_p \\ \log(\sigma_p^2) \end{bmatrix} = W_p \begin{bmatrix} q \\ u \end{bmatrix} + b_p \qquad (2)$$

$$\begin{bmatrix} \mu_q \\ \log(\sigma_q^2) \end{bmatrix} = W_q \begin{bmatrix} q \\ r \end{bmatrix} + b_q \qquad (3)$$

where W_p, W_q, b_p and b_q are the trainable parameters. A sample of z using the reparametrization trick (Kingma and Welling, 2013) is then fed into the decoder as a part of input at each time step.

In addition, the bag-of-word (BOW) loss (Zhao et al., 2017) is employed to tackle the latent variable vanishing problem, and PAGenerator is trained to maximize the variational lower-bound (Chung et al., 2015; Serban et al., 2017):

$$\begin{aligned} \mathcal{L}(\theta,\phi;q,r,u) &= \mathrm{E}_{q_\phi(z|q,r)}[\log p_\theta(r|z,q,u)] \\ &- KL(q_\phi(z|q,r)\|p_\theta(z|q,u)) \\ &+ \mathrm{E}_{q_\phi(z|q,r)}[\log p(r_{bow}|z,q,u)] \end{aligned} \qquad (4)$$

2.2 User Information Enhancing Regularization

Ideally, we expect that the introduction of user embedding is fully utilized during model training. However, due to the KL vanishing problem, the training of PAGenerator suffers from the hazard that the rapid decrease of $\mathcal{L}$ in Equation 4 might be attributed to the strong fitting capability of the decoder on the training data, rather than the involvement of user embedding. Thus, we introduce a regularization term to promote the usage of user's hidden information in latent variables.

At the beginning, as illustrated in Figure 1, a general unk_u is introduced to represent the case for user unspecified. Subsequently, taking the default user embedding e_{unk_u} as input, we obtain the KL divergence as $KL(q_\phi(z|q,r)\|p_\theta(z|q,unk\_u))$ from the network. In this case, once the real user u is introduced, a regularization term $\mathcal{R}_1(\theta,\phi;q,r,u)$ can be constructed as follows:

$$\begin{aligned} \mathcal{R}_1(\theta,\phi;q,r,u) = \max(&-\gamma_1, \\ &KL(q_\phi(z|q,r)\|p_\theta(z|q,u)) \\ &- KL(q_\phi(z|q,r)\|p_\theta(z|q,unk\_u))) \end{aligned} \qquad (5)$$

where $\gamma_1 \in \mathbb{R}$, $\gamma_1 > 0$, and $p_\theta(z|q,unk\_u) \sim \mathcal{N}(\mu_p', \sigma_p'^2 \mathbf{I})$.

It should be noted that, according to the equation above, the two prior distributions are generated from the same network with partially different inputs (u VS. unk_u), and the regularization constrains the prior distribution with specified user to be closer to the posterior distribution. Thus, the optimization encourages the utilization of user information and correspondingly inhibits the generated results from ignoring the user information. Meanwhile, $\mathcal{R}_1$ in our proposed model also alleviates the KL vanishing problem.

2.3 Variance Controlling Regularization

The BOW loss forces the latent variables to predict the bag-of-words in the response. Therefore, the semantic distribution of z is required to be capable of representing the topics and wording of the target response. Besides, for a given query, the possible replies from a specific user should be more convergent to each other than those from an unknown user, due to each user's unique preference on the topics and wording. Correspondingly, under the assumption that the distribution of z represents the user's language preference, the specification of user information is expected to reduce the entropy of the isotropic Gaussian distribution of z, reflected by a lower standard deviation σ_p. On this basis, we introduce another regularization term $\mathcal{R}_2(\theta,\phi;q,r,u)$ to control the variance:

$$\mathcal{R}_2(\theta,\phi;q,r,u) = \max(-\gamma_2, \sigma_p^2 - \sigma_p'^2) \qquad (6)$$

where $\gamma_2 \in \mathbb{R}$ and $\gamma_2 > 0$. $\mathcal{R}_2$ prefers those z with decrease $\geq \gamma_2$ in standard deviation σ_p after specifying users, and such decrease indicates the latent variables are more semantically convergent.

On this basis, we update the new training objective of PAGenerator as follows:

$$\begin{aligned} \mathcal{L}'(\theta,\phi;q,r,u) &= \mathcal{L}(\theta,\phi;q,r,u) \\ &- \mathcal{R}_1(\theta,\phi;q,r,u) - \mathcal{R}_2(\theta,\phi;q,r,u) \end{aligned} \qquad (7)$$

By employing the two regularization terms to constrain the model training, $\mathcal{L}'(\theta,\phi;q,r,u)$ now also pays attention to the utilization of user information and language preference.

3 Specified Evaluation Metrics of Persona NRG

In the previous section, two regularization terms are proposed to guide the model in the persona exploration. However, we still lack effective persona-focused metrics to quantify how well one

model is on learning persona. The currently applied metrics for persona-aware NRG evaluation, such as perplexity and BLEU, are used to evaluate the plain NRG models (Li et al., 2016b; Kottur et al., 2017). Apparently, such metrics are inadequate to evaluate the capacity of a response generator on capturing persona.

Innately, an effective persona-aware response generator should be able to successfully identify and generate responses for users according to their language styles. Besides, the generated responses from different users should be diversified to each other in wording. Considering these properties, we propose the following metrics to measure the level of persona-aware in response generators.

3.1 Language Style Detection

It is important for a persona-aware response generator to identify a user's response from other user-irrelevant ones, by detecting the user's language style in responses. In this subsection, we propose User-Relative-Rank (*uRank*) to measure such capability. Given a query-response-user triple $\{q, r, u\}$, a pre-trained seq2seq model $S2S$ and a model M to be evaluated, we first generate n user-irrelevant responses $\{r'_i | i \in [1, n]\}$ from $S2S$ using beam search. For a desired persona-aware model M, it is expected to assign the ground truth response r with a higher probability than other user-irrelevant ones $\{r'_i | i \in [1, n]\}$. Thus, taking $S2S$ as reference, we set *uRank* to be 1 if M scores r a higher ranking position among r'_i than $S2S$, specifically:

$$
\begin{aligned}
rank_M &= |\{i | P_M(r'_i) > P_M(r)\}| \\
rank_{S2S} &= |\{i | P_{S2S}(r'_i) > P_{S2S}(r)\}| \\
uRank &= \begin{cases} 1 & \text{if } rank_M < rank_{S2S} \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{8}
$$

where $P_m(r)$ and $P_{s2s}(r)$ are the probabilities of $\{q, r, u\}$ given by M and $s2s$ respectively, $|X|$ presents the cardinal number of a set X, and the lower score of either $rank_M$ or $rank_{S2S}$ indicates a better ranking position. Overall, for model M, its average *uRank* for different queries denotes the rate of rank-promoted ground-truth replies.

3.2 Language Style Imitation

Apart from perceiving users' language styles, an effective persona-aware model should also be able to imitate language styles by generating responses satisfying users' language behaviors.

User-Language-Perplexity (*uPPL*) is proposed to measure this property.

Given a user u_i, to conduct such metric, a statistical language model LM_i is first trained using the user's utterances. After that, for a generated response r', its corresponding *uPPL* is defined as the perplexity of r' given by LM_i. *uPPL* quantifies the power of a persona-aware model on generating responses similar to users' history utterances.

3.3 Diversity between Users

Finally yet importantly, due to the introduction of user information, given a query, we expect that responses for different users from a persona-aware model should be also diversified. Therefore, Users-Distinct (*uDistinct*) is proposed in this paper to capture such property. Given a query q_i and m different users $\{u_j | j \in [1, m]\}$, we generate different responses $\{r'_j | j \in [1, m]\}$ for each user using M. On this basis, Distinct-1 and Distinct-2 (Li et al., 2016a) of the response set $\{r'_j | j \in [1, m]\}$ are utilized to measure the in-group diversity of responses generated by M within users. Li et al. (2016b) also compare models through the case studies from the similar perspective.

4 Experiments

4.1 Datasets

To evaluate the performance of our proposed method, we implement experiments on a Chinese Social Networking Service (SNS) corpus and the Cornell Movie Dialogues corpus (Danescu-Niculescu-Mizil and Lee, 2011). The Chinese SNS corpus is crawled from a Chinese social network service Douban,[1] containing totally 1,022,592 single-turn dialogues from 12,857 users; while the Cornell Movie Dialogues corpus consists of conversations from movie scrips. By cleaning up the Cornell corpus with the open-source script,[2] we obtain 109,952 single-turn dialogues from 9,035 movie characters. The training/test ratios for the two corpora are around 200:1 and 50:1, respectively. Besides, for the Douban corpus, the mean, maximum, minimum, and the standard deviation values of the number of utterances for each user are 80, 1190, 33, and 49, respectively. Meanwhile, these statistics values are 14, 237, 4, and 22, correspondingly.

---

[1]https://www.douban.com/group
[2]https://github.com/suriyadeepan/datasets/

There are two main differences between the two datasets: 1) The scenes of conversations are different. The dialogues in Douban are crawled from an open domain social media. By contrast, since the characters in Cornell movie corpus are assigned with fixed personas, the language styles and habits of users are more templatized. Besides, the language style in Cornell is more oral-like, with many personal pronouns. 2) The average number of utterances for each user of the Douban corpus is around 10 times more than that of Cornell.

4.2 Model Variations

S2SA Vanilla sequence-to-sequence model with attention (Sordoni et al., 2015).

fact_bias S2SA with fact bias for persona modeling (Michel and Neubig, 2018). fact_bias is originally proposed in NMT, it models user information as an additional bias vector learned through a factored model in the softmax layer.

Speaker Model Framework proposed by Li et al. (2016b). This model is similar to S2SA + fact_bias, except that the user information is added as a part of decoder input rather than bias in the softmax layer.

VAE Standard Variational AutoEncoder for response generation (Serban et al., 2017). In our experiment, we replace the utterance with the query only and apply the auxiliary BOW loss (Zhao et al., 2017) in training.

CVAE Conditional Variational AutoEncoder with user information as prior knowledge for modeling persona (Zhao et al., 2017). Similar to VAE, bag-of-words loss is applied in CVAE.

For a fair comparison, we use the same configuration for all models. The size of word embedding and user embedding are respectively set to 300 and 128. All the user embeddings, including that of the unknown user, are initialized randomly and trained during the optimizing. We employ a bi-directional LSTM of hidden size = 256 for encoding, and a LSTM of hidden size = 512 for decoding. For latent models, the dimension of z is set as 128.

All models are optimized using Adam (Kingma and Ba, 2014) with learning rate $= 2e-4$ and batch size = 128. For latent models, we also use KL annealing (Bowman et al., 2016) (400,000 batches for Douban corpus and 100,000 batches for Cornell Movie corpus) to achieve better performance.

4.3 Automatic Evaluation Metrics

To thoroughly evaluate our systems, both standard and persona-focused metrics are employed in our experiments. For standard metrics, we adopt unigram BLEU (BLEU-1) (Papineni et al., 2002) and Word Embedding metrics (Liu et al., 2016) including Embedding Average (Average), Vector Extrema (Extrema) and Greedy Matching (Greedy) to evaluate the semantics of generated responses with regards to ground truths. We use the pretrained word embeddings from (Song et al., 2018) for the Douban corpus and embeddings from (Pennington et al., 2014) for the Cornell movie corpus.

The three proposed metrics (*uRank*, *uPPL* and *uDistinct*) are adopted to measure the performance of capturing persona. For *uPPL*, we use a bi-gram language model for perplexity computation. Since the effectiveness of *uPPL* relies on the quality of constructed user language models, we pretrain the SLM with the whole training data and afterwards finetune it using each user's utterances. Besides, we drop the users with utterances less than 100 in Douban and 30 in Cornell. The value of *uRank*, which depends on the rankings of predicted probabilities of responses, is not stable for latent models due to the randomness on sampling z. Therefore, *uRank* for each latent model is computed by running 10 rounds, so that we obtain 10 ranking results and their corresponding *uRank*. Then we average the obtained 10 *uRank* as the final *uRank* for each latent enhanced model. The later experimental results show that *uRank* for any latent model varies slightly around ± 0.005 for each round.

4.4 The Human Evaluation Criterion

For further comparisons, we also use the crowdsourcing labeling resources of our organization to manually evaluate the relevance and the persona of generated responses. Since the degree of persona reflected in the response is even more difficult to be judged by humans, we simplify the annotation into a "yes or no" task, that is, annotators are only asked to decide whether the response can reflect persona for the given user. Before that, the annotators have to read all the utterances of each user to learn the persona for judging. Moreover, in practice, we limit the number of each user's sample utterances to 100. However, the judgment is inevitably much more subjective. Thus, for each sample, we recruit 11 annotators to label and make the final determination by voting. The evaluation

of relevance is relatively easy. For the evaluation of relevance, each query-response pair is cross-evaluated by 3 annotators, following the labeling criterion used in (Xing et al., 2017; Wang et al., 2018). The details of data sampling and labeling are given in the **Supplementary Material**.

5 Results & Analysis

5.1 Results on the Douban Corpus

We first report the performance on the Douban corpus. The results of automatic evaluating metrics are illustrated in Table 1, numbers in bold mean that the improvement on that metric is statistically significant over other methods (**p-value ≤ 0.01**). It is observed that the BLEU-1 scores of various models are relatively low and close to each other. We attribute this to the fact that the semantics of possible responses for one query is highly diversified in terms of speaking styles and topics, there might be the situation that only a small portion of words share among the responses except those of high-frequency words (Mou et al., 2016; Liu et al., 2016). However, user enhanced models achieve higher BLEU-1 scores due to their capability in considering the preference of a user.

Furthermore, by comparing the performances on embedding metrics, we find that all models obtain decent scores, but none of the models outperform the others significantly. Such phenomenons can also be observed in previous studies (Serban et al., 2017; Wang et al., 2019), since all the models generate responses semantically similar to the ground truths. Despite this, PAGenerator achieves the highest score on average, which suggests the responses generated by PAGenerator are more semantically relevant to the ground truths.

While all models perform more or less the same on standard metrics, their experimental results on persona metrics are quite different. All persona-aware NRG models outperform S2SA and VAE which contain no user information on the *uRank*, while the two variational models with user information significantly exceed the rest models. It shows that persona-aware response generators, especially those exploiting user embeddings to generate latent variables, are more sensitive on identifying users' language styles. Among all models with user modeling, our proposed PAGenerator achieves the highest *uRank*.

The advantage of introducing persona information into NRG is also reflected by *uPPL*. The replies given by the three models employing user embeddings are more consistent with the user's language style, which indicates that user embedding is useful in learning language style automatically in an End-to-End NRG model. By contrast, since S2SA with fact_bias focuses on learning user's bias based on only unigrams, it struggles from achieving a high *uPPL* which scores from bigram perspective. Moreover, comparing the performance of CVAE to Speaker Model, it appears that utilizing latent variables in standard method cannot further improve *uPPL*. By contrast, the two new regularizations proposed for persona modeling can help PAGenerator generating replies with more specific persona, the *uPPL* of which is reduced by 21.2 points compared to CVAE.

As mentioned in previous sections, *uDistinct* measures the diversity of the generated responses between different users. In general, latent models achieve higher *uDistinct* than non-latent ones as the randomness brought by the latent variables. Within latent models, the adoption of user information in CVAE only slightly improves its *uDistinct* compared to VAE without user specification. It indicates that user embeddings are ineffectively utilized in CVAE, and this is the motivation for us to propose new methods for variational response generator. The notable improvement in *uDistinct* can verify their effectiveness in exploiting persona. The cases can further demonstrate such improvements in **Supplementary Material**.

Besides, the comparison among baseline models is consistent with the experiments in previous studies (Li et al., 2016b; Zhou and Wang, 2018), which indicates the proposed metrics are apposite for evaluating the capability of NRG models on capturing persona.

5.2 Human Evaluation

To further evaluate the quality of generated responses from each model more subjectively, we also implement human labeling. As shown in Table 2, adjusting unigram distributions for users by fact_bias reduces the quality of generated responses. By contrast, all other models produce more high-quality replies comparing with S2SA. Moreover, responses from PAGenerator achieve the best human evaluation result, which indicates that the improvement of persona capturing of PAGenerator does not reduce correlation.

Meanwhile, in the last column, the trend of eval-

Methods	BLEU	Embedding			Persona Metrics			
		Average	Extreme	Greedy	uRank	uPPL	uDist-1	uDist-2
S2SA (Sordoni et al., 2015)	0.29	0.834	0.615	0.666	0	200.4	0.115	0.113
fact_bias (Michel and Neubig, 2018)	0.29	0.840	0.618	0.671	0.022	202.3	0.091	0.101
Speaker Model (Liu et al., 2016)	**0.31**	0.837	0.621	**0.674**	0.023	163.6	0.183	0.199
VAE (Serban et al., 2017)	0.30	0.830	0.609	0.659	0.017	225.9	0.367	0.467
CVAE (Zhao et al., 2017)	**0.31**	0.836	0.616	0.668	0.039	174.5	0.377	0.486
PAGenerator	**0.31**	**0.845**	**0.622**	0.670	**0.044**	*153.3*	**0.406**	**0.524**

Table 1: Evaluation results on Douban corpus. *uDist* is the abbreviation for *uDistinct* in the table.

Methods	Human Evaluation			
	0	1	2	Persona
S2SA	60.0%	35.0%	5.0%	1.6%
fact_bias	70.0%	26.7%	3.3%	7.8%
Speaker Model	53.2%	41.6%	5.2%	9.6%
VAE	58.3%	35.0%	6.7%	3.8%
CVAE	55.0%	38.8%	7.2%	12.2%
PAGenerator	*51.7%*	38.3%	**10.0%**	**13.4%**

Table 2: Human labeled results upon generated responses of models trained on the Douban corpus, with the beam width of 10. The Fleiss' kappa (Fleiss and Cohen, 1973) on all annotations is around 0.65, which can be considered as "substantial agreement".

uated results on persona almost consists to those evaluated by proposed automatic evaluation metrics. The PAGenerator outperforms other models, and some particular parts of replies generated by persona-aware models can reflect the personality. Besides, due to the randomness, some responses given by S2SA and VAE are also labeled as persona-aware. However, fewer high-quality responses generated by S2SA compared to VAE, and thus, the proportion of S2SA is even lower.

5.3 Results on the Cornell Corpus

As shown in Table 3, the overall trend of the experimental results on Cornell corpus is consistent with that on Douban corpus. The models that are aware of the specified user outperform others slightly on BLEU and Embedding metrics. Regards to persona metrics, the experimental results on Cornell corpus shows two main differences: a) The Speaker Model does not perform that well on user language style detection and generation, mainly because the training data of each user is less than that in Douban corpus. It is hard to automatically model the informative user embedding via target oriented learning without guidance. By contrast, utilizing the KL divergence as the guidance in CVAE effectively improves the exper-

imental results. b) Due to the individual characteristics of movie characters, the user-embedding-enhanced models generate more diverse responses for different users, specially PAGenerator.

5.4 Human Evaluation Results on the Cornell Corpus

As shown in Table 5, on the English dataset, the comparison results are almost consistent with that in Section 5.2. According to the judgment of annotators, our proposed model outperforms the others from both relevance and persona perspective. However, influenced by insufficient training conversations, the overall quality of generated responses for the Cornell queries is not as good as the ones given for the Douban corpus. We attribute this to the difference in the corpus size and the word distribution, which is described in Section 4.1. In detail, the quality of Cornell is influenced by insufficient training conversations. By contrast, the persona is reflected more obviously with the help of more templatized language styles and habits of Cornell.

5.5 Ablation Study

To get a better intuition about how our proposed method works, we implement the ablation tests to analyze the contribution of each component of PAGenerator in persona exploitation. As illustrated in Table 4, adding the user embeddings as a part of decoder inputs brings positive improvements on all the persona-focused metrics. Without UE, the parameter size of PAGenerator reduces considerably, which is harmful to the model on fitting target data. Besides, without direct constraints from the decoder, user embeddings mainly act on reducing KL divergence rather than providing more informative latent variables. Besides, without UE, PAGenerator also significantly outperforms VAE in all metrics, which demonstrates that $\mathcal{R}_1$ and $\mathcal{R}_2$ are indeed useful for guiding the latent variables

Methods	BLEU	Embedding			Persona Metrics			
		Average	Extreme	Greedy	uRank	uPPL	uDist-1	uDist-2
S2SA (Sordoni et al., 2015)	0.32	0.787	0.503	0.679	0	44.8	0.115	0.079
fact_bias (Michel and Neubig, 2018)	0.30	0.785	0.501	0.676	0.044	39.3	0.127	0.095
Speaker Model (Liu et al., 2016)	**0.33**	0.796	0.510	0.681	0.056	41.7	0.228	0.225
VAE (Serban et al., 2017)	0.25	0.780	0.490	0.670	0.058	45.6	0.122	0.114
CVAE(Zhao et al., 2017)	0.28	0.800	0.502	**0.689**	0.085	37.0	0.223	0.251
PAGenerator	**0.33**	**0.814**	**0.514**	0.687	**0.114**	*32.2*	**0.251**	**0.304**

Table 3: Comparison of different approaches on the Cornell Movie Dialogues corpus.

Methods	uRank	uPPL	uDist-1/2
PAGenerator	0.114	32.2	**0.251** / 0.304
w/o $\mathcal{R}_1$	0.117	*29.6*	0.209 / 0.246
w/o $\mathcal{R}_2$	**0.118**	37.2	**0.251** / **0.319**
w/o UE	0.063	43.5	0.149 / 0.139

Table 4: Ablation tests of PAGenerator on Cornell Movie Dialogue Corpus. "w/o" denotes PAGenerator does not contain the specific component, for example, "w/o UE" means the decoder of PAGenerator does not utilize the user embedding as input.

Methods	Human Evaluation			
	0	1	2	Persona
S2SA	70.6%	27.5%	1.9%	1.4%
fact_bias	72.2%	26.0%	1.8%	14.9%
Speaker Model	62.2%	35.6%	2.2%	16.9%
VAE	65.0%	31.6%	3.4%	1.1%
CVAE	61.7%	34.0%	4.3%	21.6%
PAGenerator	*61.5%*	33.8%	**4.7%**	**22.8%**

Table 5: Human evaluation results on the Cornell Corpus.

to model the semantics under the query and users.

Comparing the ablation results of w/o $\mathcal{R}_1$ with w/o $\mathcal{R}_2$, we can conclude that both regularizations promote *uRank* values. However, PAGenerator w/o $\mathcal{R}_2$ only achieves a mediocre result on *uPPL*, while only utilizing $\mathcal{R}_2$ damages the model's ability in generating diverse responses for different users. We attribute this divergence to the trade-off between a) shared movie-style language between users and b) different language preferences among actors in the movie scripts. Since $\mathcal{R}_1$ promotes the divergence of z between the specified and unspecified users, removing $\mathcal{R}_1$ raises the difficulty for the model to generate diverse responses toward different users, reflected by the low *uDistinct* of w/o $\mathcal{R}_1$. However, promoting diversity will more or less sacrifice the model's learning on the common shared movie-style patterns, which is vital in evaluating the language cohesion. Therefore, the performance of PAGenerator only with $\mathcal{R}_1$ on *uPPL* is less-than-ideal. In contrast, since $\mathcal{R}_2$ emphasizes those patterns often used by a given user, it encourages the distribution of user information to be more aggregate. These differences explain the opposite results of w/o $\mathcal{R}_1$ and w/o $\mathcal{R}_2$.

In conclusion, the user embedding is an important constraint for the PAGenerator, and $\mathcal{R}_1$, $\mathcal{R}_2$ can be considered to deploy for different purposes. Furthermore, utilizing all components of PAGenerator described in Figure 1 guarantees a more bal-

anced and relatively best performance in all three evaluated persona exploiting abilities.

6 Related Work

6.1 Persona-based Neural Models

Persona-based neural conversation models can be categorized into two major research directions. One is to directly train a model from conversational data by considering the persona information (Li et al., 2016b; Kottur et al., 2017; Wang et al., 2017; Madotto et al., 2019), while the other approach makes use of the profiles or side-information of users to generate the aligned responses (Chu et al., 2018; Qian et al., 2018; Zhang et al., 2018; Mazare et al., 2018; Song et al., 2019). The work described in this paper belongs to the first research direction. Li et al. (2016b) and Kottur et al. (2017) enrich the models by training persona vectors directly and incorporating them into the decoder. Wang et al. (2017) propose three strategies to learn the language style instead of introducing new models.

Apart from the development of the Persona-based NRG models, recent researches also attempt to incorporate persona into neural machine translators. Michel and Neubig (2018) propose to learn speaker-specific parameters for the bias term in the output to promote user preferring unigrams, and Wuebker et al. (2018) introduce offset tensors to perform fine-tuning for each user.

6.2 Variational Response Generator

The variational response generators have drawn much attention recently, due to the observation that it can be flexible to include the effect from conditions based on its Bayesian architecture (Zhao et al., 2017; Shen et al., 2017) and naturally promote diversity by involving sampling in the generate stage (Serban et al., 2017; Du et al., 2018; Shen et al., 2018). Zhao et al. (2017) and Shen et al. (2017) introduce frameworks taking various conditions to influence the model learning. Afterwards, Zhou and Wang (2018) include the emoji into the variational NRG model to generate responses with particular emotions. Actually, these models (Zhao et al., 2017; Shen et al., 2017; Zhou and Wang, 2018) can also be deployed to the persona-aware response generation scenario. The main difference is that the speaker of the response is unpredictable based on the query. Thus, we have introduced the architecture proposed by Zhao et al. (2017) and modified it to adapt to the persona-aware generation, for the meaningful comparison. Especially, Song et al. (2019) have utilized persona information into the CVAE architecture, except they focus on modeling and copying users' explicit profiles.

7 Conclusions

In this paper, we proposed a variational neural network to model the conversation as well as the persona of users. On the basis of the network, two regularization terms are designed to guide the model in emphasizing the importance of the hidden user information. In addition, to better reflect the persona characteristics of the response generation model, three metrics have been introduced to quantify the level of persona of the generated responses. Experimental results show that our approach significantly outperforms other baseline models and the proposed metrics are effective in evaluating the capabilities of models on generating persona-aware responses.

Acknowledgments

References

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *arXiv preprint arXiv:1606.00372*.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. *CoNLL 2016*, page 10.

Eric Chu, Prashanth Vijayaraghavan, and Deb Roy. 2018. Learning personas from dialogue with attentive memory networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2638–2646.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

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.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

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

Diederik P Kingma and Max Welling. 2013. Autoencoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003. Association for Computational Linguistics.

Bingquan Liu, Zhen Xu, Chengjie Sun, Baoxun Wang, Xiaolong Wang, Derek F Wong, and Min Zhang. 2018. Content-oriented user modeling for personalized response ranking in chatbots. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(1):122–133.

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

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 5454–5459.

Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318. Association for Computational Linguistics.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358.

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

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.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *IJCAI*, pages 4279–4285.

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, Vera Demberg, and Shuzi Niu. 2018. Improving variational encoder-decoders in dialogue generation. *national conference on artificial intelligence*, pages 5456–5463.

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A conditional variational framework for dialog generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 504–509.

Haoyu Song, Weinan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5190–5196.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180.

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

Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. Steering output style and topic in neural response generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2140–2150.

Zongsheng Wang, Yunzhi Bai, Bowen Wu, Zhen Xu, Zhuoran Wang, and Baoxun Wang. 2018. A prospective-performance network to alleviate myopia in beam search for response generation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3608–3618.

Zongsheng Wang, Zhuoran Wang, Yinong Long, Jianan Wang, Zhen Xu, and Baoxun Wang. 2019. Enhancing generative conversational service agents with dialog history and external knowledge. *Computer Speech & Language*, 54:71–85.

Joern Wuebker, Patrick Simianer, and John DeNero. 2018. Compact personalized models for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 881–886.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 654–664.

Xianda Zhou and William Yang Wang. 2018. Mojitalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137. Association for Computational Linguistics.

A   Details of Human Evaluation

A.1   Labeling Dataset Preparation

For each model with a given query set, three generated responses for each query are randomly sampled from the results given by the beam search with a beam size of 10. Then, a total of 3,000 query-response pairs are prepared for labeling.

A.2   Labeling Criterion of Relevance

The labeling criterion for judging the relevance between the response and the given query is described as follows:

0: the quality of response is poor, it is either irrelevant to the query, or grammatically incorrect.

1: although the response itself is acceptable as a reply, its content is not informative and dull.

2: the response is not only relevant and grammatically correct, but also informative or interesting.

B   Case Studies

As shown in Figure 2, we have selected three users whose utterances can reflect their implicit personal features. For example, the gender of user $U3$ in the first case is probably female. The user $U4$ in the second case is very possible to be an animation fun. According to the conversation history of user $U3$ in the last case, it can be inferred that the user is in the trouble of losing weight. Correspondingly, from the responses generated by PAGenertor, we can observe that such implicit information are adopted by our proposed model to produce persona-aware results.

Figure 3 gives additional cases generated by PAGenertor, CVAE and VAE respectively oriented to the same given query. Apparently, every independent user should have his/her own linguistic and personality characteristic. Thus, the results generated for different users are expected to maintain enough diversity. According to the cases in Figure 3, it can be seen that results of PAGenertor keep obvious diversity among different individuals, indicating its better capability of capturing persona of users.

Q: 告别 了 留 了 十多年 的 **长 直发**

CVAE
U1:我 也 是 **长发 及 腰**
U2:那 我 就 去 尝试 一下
U3:我 已经 **剪 完 了**
U4:我 也 是 **长发 及 腰**
U5:我 也 不 知道 是 什么 颜色 的

PAGenerator
U1:我 也 想 **剪 长发 及 腰**
U2:那 我 就 去 尝试 一下
U3:我 就 **剪 了 刘海**
U4:新 的 一年 新 的 开始
U5:好 **勇敢**

Conversation history of U3:
…
这个 **色号** 我 还 蛮 喜欢 的
…

Q:重温 美人 心计 还是 那样 好看

CVAE
U1: 我 看 的 是 电视剧 版 的
U2: 我 还 以为 你 是 主角 呢
U3: 嗯嗯 我 看 你 看 的 是 《 爱情 公寓 》
U4: … … 我 就 看 了 《 <ALPHA> 》
U5: 我 看 的 是 电视剧 版 的

PAGenerator
U1: 哈哈 我 看 了 几集 就 不 看 了
U2: 其实 我 也 想 看
U3: Emoji_113 我 也 看 了
U4: 其实 我 也 喜欢 看，**很 像 动画 的 剧情**
U5: 恩恩，我 看 了 几集 就 好 了

Conversation history of U4:
よく 「 殿下 」 と 呼 ば れ そ う だ が
哦 看到 **遥 久 美 男子 祭** 突然 就 懂 了
头发 是 不是 不 能 这么 **画**

Q: 小心 莫 变成 小 胖妹 了 哦

CVAE
U1: 哈哈 · · 哈哈 · ·
U2: 哎 呦 喂 你 也 是 哦
U3: 呵呵 你 说 的 对 哦
U4: 嗯嗯 、 我 也 觉得 这句 话 太 对
U5: 呵呵 你 说 的 是 实话 嘛

PAGenerator
U1: 哇 咔 咔 我 就 知道 了
U2: 我 也 不 知道 怎么 搞成 这样 了
U3: 本来 就是 嘛 **我 都 快 疯 了**
U4: 好 吧 我 也 是 个 小 棉袄 而已 啦
U5: 我 也 不 知道 该 怎么 做 的 事 了

Conversation history of U3:
…
我 的 目标 是 还 **减 五斤** 一百二 就 可以 了
我 开车 去 走路 **懒得 走** 啊
老大 你 来 **接 我** 嘛
…

Q: I've kept **long straight hair** for a decade, now it's time to farewell.

CVAE
U1:I have **long hair** as well.
U2:I will give it a shot.
U3:I've **finished my haircut**.
U4:I have **long hair** as well.
U5:I have no idea about what color it is.

PAGenerator
U1:I wanna try **long hair** too.
U2:Then I'll give a try.
U3:I just did a **fringe haircut**.
U4:New year new beginning.
U5:So **brave**.

Conversation history of U3:
…
I kind of like this **cosmetic color**.
…

Q: Just watched *Beauty's Rival in Palace* again, still fascinating.

CVAE
U1: I watched the cable version.
U2: I though you are the starring.
U3: em I think what you watched is *iPartment*
U4: … … I only watched *<ALPHA>*.
U5: I watched the cable version.

PAGenerator
U1: haha I only watched couple of beginning episodes.
U2: In fact I wanna watch it as well.
U3: Emoji_113 I watched as well.
U4: I like it too, **its story is quite like animation**.
U5: hmm,I watched some episodes.

Conversation history of U4:
よく 「 殿下 」 と 呼 ば れ そ う だ が
I immediately realized once I watched **遥か美男子祭り**
Can I do hair **caricature** this way?

Q: Don't be fatty, be aware.

CVAE
U1: haha..haha..
U2: hey, same to you.
U3: Interesting, you said it right.
U4: yeah, I think it's so true.
U5: Interesting, are you telling the truth?

PAGenerator
U1: wakaka I knew it.
U2: I don't know why it happened neither.
U3: Indeed it is, **I'm about to go crazy**.
U4: alright I'm just a sweet heart.
U5: I've no idea about how to do this.

Conversation history of U3:
…
My goal is to **lose 2.5 kg more**, 60 kg works for me.
I plan to drive there. **Hate to walk**.
Pick me a drive, boss.
…

Figure 2: Comparisons of cases generated by CVAE and PAGenerator. Especially, we also give the utterances from conversation histories of some users (Conversation history of U_k, where U_k denotes different users from different cases). The translated English version of the samples are listed on the right.

Q: 时间 过 的 真快 来 公司 都 **七年** **半** 了 ~

PAGenerator
U1: 嗯嗯 我们 都 是 **一个** **星期**
U2: 我 还 以为 你 是 **几** **个** **月** 呢
U3: 你 在 哪里 **上班** 啊
U4: 我 也 是 今天 又 开始 **上班** 了
U5: 有 时间 就 可以 去 了

CVAE
U1: 呵呵 … 我 都 快 过年 了
U2: 是 啊 我 都 快到 家 了
U3: 不是 吧 我 今天 才 到家
U4: 是 啊 我 现在 才 到家
U5: 有 什么 好 的 啊

VAE
A1: 是 啊 我 都 快 一年 了
A2: 是 啊 时间 过得 太快 了
A3: 是 啊 现在 还 没 开始 呢
A4: 是 的 啊 我 都 快 一个 月 了
A5: 是 啊 时间 太快 了

Q: Time flies, I've been worked in this **company** for **seven and a half years**.

PAGenerator
U1: yeah, we are both for **one week**.
U2: I thought you've been here for **several months**.
U3: **Where do you work?**
U4: I start to **work** today as well.
U5: You can go if you have time.

CVAE
U1: interesting...I'm about to celebrate the new year.
U2: Yeah, I'm about to return home.
U3: It can't be, I just came back home today.
U4: Yes, I just came back home now.
U5: Nothing good.

VAE
A1: Yes, even I have been worked for almost one year.
A2: Yes, time flies.
A3: Yes, It hasn't started yet.
A4: Yes, even I have been worked for almost one month.
A5: Yes, time files.

Q: 送点 她 喜欢 的 **小** **礼物** , 说点 暖心 的 话

PAGenerator
U1: 祝 你 妈妈 **生日快乐** 身体健康
U2: 我 还 记得 你 送 我 **礼物** 呢
U3: 你 也 要 **快乐** 哦
U4: 我 也 想 送 你 礼物
U5: 你 的 **愿望** 是 最 棒 的

CVAE
U1: 我 也 想 买 礼物 啊
U2: 我 也 是 这么 想 的
U3: 你 要 给 我 买 糖 吗
U4: 我 也 想 给 你 买 礼物
U5: 你 要 给 我 买 糖 吗

VAE
A1: 我 也 想 给 你 寄 礼物
A2: 你 要 送 我 礼物 吗
A3: 给 你 买 礼物 啊
A4: 我 也 要 给 你 送 礼物
A5: 你 还 记得 我 的 礼物 呢

Q: Send her some **little presents** she likes, say some warming words,

PAGenerator
U1**: Happy birthday** to ur mom, wish her the best health.
U2: I still remember your **gift**.
U3: Be **happy**!
U4: I want to give you a gift too.
U5: Your **wish** is the best.

CVAE
U1: I want to buy some gifts too.
U2: That's exactly what I think.
U3: You wanna buy me some sugar?
U4: I want to buy gifts to you too.
U5: You wanna buy me some sugar?

VAE
A1: I want to send you gifts too.
A2: Do you plan to send me gifts?
A3: Do you want my gift?
A4: I want to send you gifts too.
A5: Still remember my gift?

Q: 快 **九月** 了 有 什么 打算

PAGenerator
U1: 我 也 不 知道
U2: 等 我 回来 了 请 你 吃饭
U3: 过 几天 就 **回去** 了
U4: 我 也 不 知道 什么 时候 **回** **学校** 了
U5: 我 还 没 **毕业** 呢

CVAE
U1: 年 年 都 过 了
U2: 年 年 都 没 结婚 了
U3: 我 也 不 知道 该 怎么办 呢
U4: 我 也 不 知道 是 什么 情况
U5: 我 也 不 知道 该 怎么办 了

VAE
A1: 我 也 不 知道 是 谁 啊
A2: 是 啊 我 都 快 一年 了
A3: 计划 赶不上 变化
A4: 过 几天 就 好 了
A5: 年 年 都 过 了

Q: It's almost **September**, any plan?

PAGenerator
U1: I have no idea.
U2: I'll treat you to a meal when I come back.
U3: I will **come back** soon,
U4: I don't know when I will **return school**.
U5: I haven't **graduated** yet.

CVAE
U1: Time is running.
U2: yet I'm still single.
U3: I don't know what to do.
U4: I don't know what happened.
U5: I don't know what to do.

VAE
A1: I don't know who it is.
A2: Yeah it's almost one year for me.
A3: Changes run faster than plans.
A4: It will be fine soon.
A5: Time is running.

Figure 3: Cases for comparing the PAGenerator, CVAE and VAE. It should be noted that VAE have not adopted user information. The translated English version of the samples are listed on the right.