A Semi-Supervised Stable Variational Network for Promoting Replier-Consistency in Dialogue Generation

Jinxin Chang[1,2,*], **Ruifang He**[1,2,3,*†], **Longbiao Wang**[1,2,†], **Xiangyu Zhao**[1,2], **Ting Yang**[1,2], **and Ruifang Wang**[1,2]

[1]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, China.
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China.
[3]State Key Laboratory of Cognitive Intelligence, iFLYTEK, P.R. China.
{changjinxin,rfhe,longbiao_wang}@tju.edu.cn
{zhaoxiangyu009,16622898776,wrf276224255}@163.com

Abstract

Neural sequence-to-sequence models for dialog systems suffer from the problem of favoring uninformative and non replier-specific responses due to lacking of the global and relevant information guidance. The existing methods model the generation process by leveraging the neural variational network with simple Gaussian. However, the sampled information from latent space usually becomes useless due to the KL divergence vanishing issue, and the highly abstractive global variables easily dilute the personal features of replier, leading to a non replier-specific response. Therefore, a novel **S**emi-Supervised **S**table **V**ariational **N**etwork (SSVN) is proposed to address these issues. We use a unit hypersperical distribution, namely the von Mises-Fisher (vMF), as the latent space of a semi-supervised model, which can obtain the stable KL performance by setting a fixed variance and hence enhance the global information representation. Meanwhile, an unsupervised extractor is introduced to automatically distill the replier-tailored feature which is then injected into a supervised generator to encourage the replier-consistency. Experimental results on two large conversation datasets show that our model outperforms the competitive baseline models significantly, and can generate diverse and replier-specific responses.

1 Introduction

Dialog systems, aiming at generating relevant and fluent responses in the replier-consistent way, have received increasing attention due to its numerous applications (Grosz, 2016; Chen et al., 2017a). Recently, Seq2Seq neural networks (Sutskever et al., 2014) have demonstrated excellent results on open-domain conversation (Shang et al., 2015; Sordoni et al., 2015; Vinyals and V. Le, 2015; Yao et al., 2015). However, due to lacking of the global and relevant information guidance, they inherently tend to generate trivial and uninformative responses (e.g., "I don't know"), rather than meaningful and replier-specific ones (Serban et al., 2016; Li et al., 2016).

The existing methods based on neural variational methods with Gaussian (Kingma and Welling, 2014; Kingma et al., 2014), are proposed to use a latent variable as the global information in decoder to strengthen the generation (Serban et al., 2017; Zhao et al., 2017; Chen et al., 2018). However, they face the problems of (1) **latent space futility** and (2) **replier-consistency decay**.

(1) The model tends to select more gain from a lower Kullback-Leibler (KL) divergence during training, which encourages the approximate posterior close to Gaussian prior, rendering the latent space of the former unused. Thus, the latent variables on this space become worthless global guidance for decoder. To address this issue, most previous work (Xie et al., 2017; Yang et al., 2017; Chen et al., 2017) has suggested a weaker decoder to match the Gaussian samples, which essentially sacrifice the generative capacity. (2) The speakers in a dyadic conversation have different linguistic characteristics, sentiments and personalities. However, the latent variable is learned conditioned on the holistic context without any distinction between speakers, especially the replier. This will dilute the personal features of replier and lead to a decrease in replier-consistency. Current methods (Li et al., 2016; Zhang et al., 2018) normally recur to artificially scheduled personal information to promote the replier-consistency, but they cannot be migrated to the other datasets.

Inspired by the effectiveness of vMF distribution in solving the KL-vanishing in the unsupervised scene (Xu and Durrett, 2018) and the suc-

* Equal contribution.
† 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 1920–1930, Hong Kong, China, November 3–7, 2019.

cess of Variational Auto-Encoder (VAE) in capturing latent feature of the real data (Davidson et al., 2018), we propose a **S**emi-Supervised **S**table **V**ariational **N**etwork (SSVN) framework to address the above issues. It consists of an unsupervised personal feature extractor (a VAE with vMF) and a supervised information-enhanced generator (a CVAE with vMF). To maintain the consistency of replier features, the extractor only encodes the previous utterances from the replier and produces a personally tailored latent variable. On the top of this, the generator fuses the replier-tailored latent variable and the self vMF distributed global information to facilitate the diverse and replier-specific responses.

In general, our contributions are as follows:

- A semi-supervised stable variational network is proposed to solve the latent space futility issue and promote the replier-consistency.
- To the best of our knowledge, our model is the first to use the vMF distribution in a semi-supervised framework for dialogue generation, which can enhance the global information by alleviating the KL divergence vanishing problem.
- An unsupervised personal feature extractor is designed to acquire the replier-specific features automatically.
- The experimental results on two large conversation datasets validate the effectiveness of our model.
- It is shown that the different roles of vMF on extractor and generator. We suprisingly find that the extractor can alleviate the KL-vanishing to some extent.

2 Related Work

2.1 Neural Variational Network

Variational autoencoder (VAE) (Kingma and Welling, 2014; Rezende et al., 2014) is one of the most popular generative models. The principle idea is to encode the data x to learn a probability distribution z, then sample the latent variables from z and inject them into a directed decoder network to reconstruct x. The model parameters are optimized by maximizing a reparameterized variational lower bound. Based on this process, the conditional VAE (CVAE) (Sohn et al., 2015) can be conditioned on certain attributes to improve diversity. In diaog generation task, Serban et al. (2017) employs the CVAE to acquire a global latent variable as a holistic representation in a hierarchical setting. Zhao et al. (2017) regards the latent variable as a global dialog act information and directly feed it to the decoder to control the dialog act of a response. To maintain the long-term memory of the previous utterances, Chen et al. (2018) utilizes the higher-level abstract variable to retrieve and update memory cells.

2.2 Latent Space Futility

As for the latent space futility issue, also called KL-vanishing in Shen et al. (2018), most previous work has suggested a weaker decoder to encourage the simple Gaussian samples to be leveraged, such as a word drop-out technique in decoder (Xie et al., 2017; Zhao et al., 2017) or a practice of replacing RNN decoder with a CNN counterpart (Yang et al., 2017; Chen et al., 2017; Semeniuta et al., 2017). These methods are contrary to our origiral intention due to generation capacity descending. Other efforts focus on changing prior and posterior: Rezende and Mohamed (2015) and Kingma et al. (2016) utilize a normalizing flow to transform the sampled variables; Shen et al. (2018) introduces an AE module to complicate the quondam distribution. Extending the latter direction yet without increasing the model complexity, we only leverage the vMF distribution instead of the simple Gaussian to strengthen the KL term. Unlike the single vMF-based VAEs implemented in other cases (Davidson et al., 2018; Guu et al.; Xu and Durrett, 2018), we apply vMF into a semi-supervised dialog model to generate diverse and replier-specific responses.

2.3 Replier-Consistency Decay

In order to emphasize the replier-consistency, Li et al. (2016) captures personas' characteristics of Twitter users by encapsulating background information and speaking style into the distributed embeddings (one per user), which are used to improve consistency for the same person. Zhang et al. (2018) presents a persona-provided dialogue dataset and trains dialog models conditioned on their given configurable, but persistent profile information. However, the above work relies heavily on manually scheduled persona information and has difficulties in migrating to other common conversation corpora. In contrast to this, our work

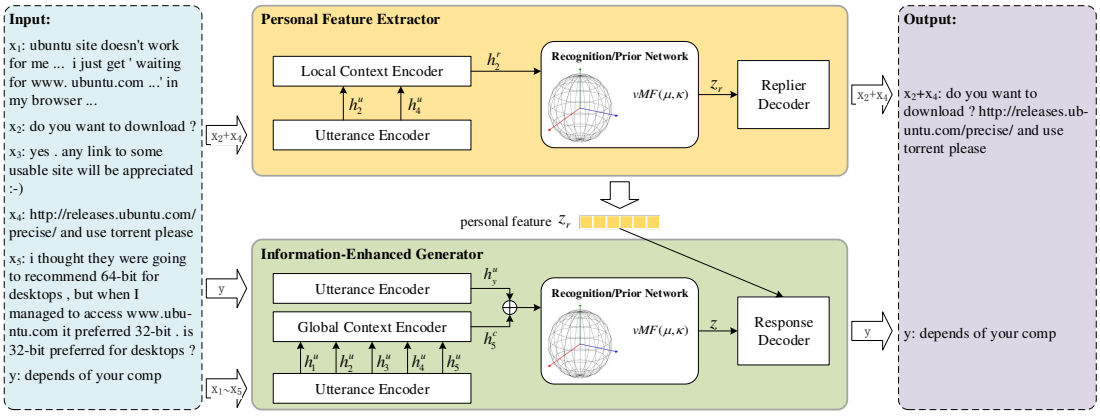

Figure 1: The SSVN framework.

focuses on automatically extracting the individual features of replier from the original conversation text, to enhance the replier-consistency of responses, without any corpus restriction.

3 Model

3.1 Task Description

Given a series of dialogue context utterances $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$, where $\boldsymbol{x}_i = (w_{i,1}, w_{i,2}, ..., w_{i,N_i})$, our task is to generate a response $\boldsymbol{y} = (w_{y,1}, w_{y,2}, ..., w_{y,N_y})$ that not only rely on the global information but also consider the personally special features from the replier. In this paper, we employ the vMF distribution to stimulate the potential of latent space, impelling the extractor to condense a feature-augmented individual information and the generator to generalize a useful global guidance. The overview of SSVN is illustrated in Figure 1.

3.2 von Mises-Fisher

The von Mises-Fisher (vMF) places a distribution over points on the unit hypersphere, parameterized by a direction vector $\mu \in \mathbb{R}^d$ indicating the mean direction and a concentration parameter $\kappa \in \mathbb{R}_{\geq 0}$. The PDF of the vMF distribution for the unit vector $z \in \mathbb{R}^d$ is defined as:

$$f_d(x; \mu, \kappa) = \mathbb{C}_d(\kappa) \exp(\kappa \mu^T x) \quad (1)$$

$$\mathbb{C}_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \quad (2)$$

where $\|\mu\| = 1$, $\mathbb{C}_d$ is the normalization constant, and I_ρ stands for the modified Bessel function of the first kind at order ρ.

3.3 Personal Feature Extractor

To enhance the replier-consistency, the personal feature extractor, implemented by a VAE with vMF, encodes rustically the context utterances from replier $\boldsymbol{x}^r = (\boldsymbol{x}^r_1, \boldsymbol{x}^r_2, ..., \boldsymbol{x}^r_l)$ (e.g., $\boldsymbol{x}^r = (\boldsymbol{x}^r_1, \boldsymbol{x}^r_2) = (\boldsymbol{x}_2, \boldsymbol{x}_4)$ in Figure 1 into a random latent variable z_r, based on which the decoder reconstructs $\boldsymbol{x}^r$. Due to an intractable high-dimensional integral problem over the latent variable z_r, we set a recognition network $q_{\phi_e}(z_r|\boldsymbol{x}^r)$ as a variational approximation to the true posterior $p(z_r|\boldsymbol{x}^r)$, then apply variational inference to optimize the evidence lower bound (ELBO) as:

$$\begin{aligned}\mathcal{L}_E &= -KL(q_{\phi_e}(z_r|\boldsymbol{x}^r)\|p_{\theta_e}(z_r)) \\ &\quad + \mathbb{E}_{q_{\phi_e}(z_r|\boldsymbol{x}^r)} \log p_{\theta_e}(\boldsymbol{x}^r|z_r) \\ &\leq \log \int_{z_r} p_{\theta_e}(z_r) p(\boldsymbol{x}^r|z_r) dz_r = \log p(\boldsymbol{x}^r)\end{aligned} \quad (3)$$

Utterance & Local Context Encoder Concretely, we employ a hierarchical encoder to encode $\boldsymbol{x}^r$: the utterance encoder based on bidirectional RNN (Schuster and Paliwal, 1997) determinitically reads each utterance $\boldsymbol{x}^r_i$ and output a size-fixed real-valued $h^u_i = [\overrightarrow{h^u_i}, \overleftarrow{h^u_i}]$, which the local context encoder takes as input to obtain the final hidden state h^r_l as the summary of $\boldsymbol{x}^r$.

Prior/Posterior Distribution Since we assume the latent space follows vMF distribution, the prior $p_{\theta_e}((z_r) \sim vMF(\cdot, \kappa^e_{prior} = 0)$ and the variational posterior $q_{\phi_e}(z_r|\boldsymbol{x}^r) \sim vMF(\mu^e_{pos}, \kappa^e_{pos})$ where μ^e_{pos} is the output of the recognition net-

work and κ^e_{pos} is set to a constant.

$$\tilde{\mu^e_{pos}} = f^e_{pos}(h^r_l) \tag{4}$$

$$\mu^e_{pos} = \tilde{\mu^e_{pos}}/\|\tilde{\mu^e_{pos}}\| \tag{5}$$

where $f^e_{pos}(\cdot)$ is a linear transformation, and $\|\cdot\|$ stands for 2-norm to ensure the normalization.

With the uniform distribution as our prior, the KL divergence can be computed as:

$$\begin{aligned} &KL(vMF(\mu^e_{pos}, \kappa^e_{pos})\|vMF(\cdot, 0)) = \\ &\left(1 - \frac{d}{2}\right)\log 2 - \log I_{d/2-1}(\kappa^e_{pos}) - \log\Gamma\left(\frac{d}{2}\right) \\ &\quad + \kappa^e_{pos}\frac{I_{d/2}(\kappa^e_{pos})}{I_{d/2-1}(\kappa^e_{pos})} + \left(\frac{d}{2} - 1\right)\log\kappa^e_{pos} \end{aligned} \tag{6}$$

Since Eq. 6 only depends on fixed constant κ^e_{pos}, not on μ^e_{pos}, this term can resolve the latent space futility problem by averting the KL-zeroing.

Replier Decoder During reconstruction, the decoder receives the concatenation of replier's context h^r_l and personal latent variable z_r as the initial hidden state, then generates tokens sequentially under the following probability distribution:

$$p_{\theta_e}(\boldsymbol{x}^r|z_r) = \prod_{i=1}^{l}\prod_{j=1}^{N_i} p(w_{i,j}|\boldsymbol{x}^r_{<i}, w_{i,<j}) \tag{7}$$

where l is the number of turns of the replier's context $\boldsymbol{x}^r$; N_i is the length of the i-th utterance ($\boldsymbol{x}^r_i$) in $\boldsymbol{x}^r$; $w_{i,j}$ is the j-th token in $\boldsymbol{x}^r_i$.

3.4 Information-Enhanced Generator

Similar to the extractor, the information-enhanced generator based on CVAE also employs a recognition network $q_{\phi_g}(z|\boldsymbol{x}, \boldsymbol{y})$ to approximate the true posterior $p(z|\boldsymbol{x}, \boldsymbol{y})$, correspondingly, its ELBO can be calculated as:

$$\begin{aligned} \log p(\boldsymbol{y}|\boldsymbol{x}) &= \log\int_z p(\boldsymbol{y}|\boldsymbol{x}, z)p(z|\boldsymbol{x})dz \\ &\geq -KL(q_{\phi_g}(z|\boldsymbol{x}, \boldsymbol{y})\|p_{\theta_g}(z|\boldsymbol{x})) \\ &\quad + \mathbb{E}_{q_{\phi_g}(z|\boldsymbol{x},\boldsymbol{y})}\log p_{\theta_g}(\boldsymbol{y}|\boldsymbol{x}, z) \end{aligned} \tag{8}$$

when considering an external personal feature z_r, the ELBO in generator would be rewritten as:

$$\begin{aligned} \mathcal{L}_G &= -KL(q_{\phi_g}(z|\boldsymbol{x}, \boldsymbol{y}, z_r)\|p_{\theta_g}(z|\boldsymbol{x}, z_r)) \\ &\quad + \mathbb{E}_{q_{\phi_g}(z|\boldsymbol{x},\boldsymbol{y},z_r)}\log p_{\theta_g}(\boldsymbol{y}|\boldsymbol{x}, z, z_r) \\ &\leq \log\int_z p(\boldsymbol{y}|\boldsymbol{x}, z, z_r)p(z|\boldsymbol{x}, z_r)dz \\ &= \log p(\boldsymbol{y}|\boldsymbol{x}, z_r) \end{aligned} \tag{9}$$

Notice that z_r only participates in the generation process $p_{\theta_g}(\boldsymbol{y}|\boldsymbol{x}, z, z_r)$, the approximate posterior $q_{\phi_g}(z|\boldsymbol{x}, \boldsymbol{y}, z_r) \sim vMF(\mu^g_{pos}, \kappa^g_{pos})$ is conditioned on dialog context $\boldsymbol{x}$ and the corresponding response $\boldsymbol{y}$, and the prior $p_{\theta_g}(z|\boldsymbol{x}, z_r) \sim vMF(\mu^g_{prior}, \kappa^g_{prior})$ depends on $\boldsymbol{x}$ [1].

Utterance & Golbal Context Encoder The hierarchical encoder in this part utilizes the shared utterance encoder from extractor to encode utterances $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n, \boldsymbol{y}$ into the corresponding representations $h^u_1, h^u_2, ..., h^u_n, h^u_y$ orderly. Thereafter, the utterance vectors $h^u_1, h^u_2, ..., h^u_n$ are fed to the global context encoder to compute the representation of the whole dialog context h^c_n. Based on these, the approximate posterior and prior can be determined by the following operations:

$$\tilde{\mu^g_{pos}} = f^g_{pos}([h^c_n, h^u_y]) \tag{10}$$

$$\mu^g_{pos} = \tilde{\mu^g_{pos}}/\|\tilde{\mu^g_{pos}}\| \tag{11}$$

$$\tilde{\mu^g_{prior}} = f^g_{prior}(h^c_n) \tag{12}$$

$$\mu^g_{prior} = \tilde{\mu^g_{prior}}/\|\tilde{\mu^g_{prior}}\| \tag{13}$$

where $f^g_{pos}(\cdot)$ and $f^g_{prior}(\cdot)$ are both linear transformations. κ^g_{pos} and κ^g_{prior} in both distributions are the constants with equal values.

Prior/Posterior Distribution Without $vMF(\cdot, 0)$ as the prior, we require to recalculate the KL term in generator as:

$$\begin{aligned} &KL(vMF(\mu^g_{pos}, \kappa^g_{pos})\|vMF(\mu^g_{prior}, \kappa^g_{prior})) = \\ &(d/2 - 1)\log\frac{\kappa^g_{pos}}{\kappa^g_{prior}} + \log\frac{I_{d/2-1}(\kappa^g_{prior})}{I_{d/2-1}(\kappa^g_{pos})} \\ &- \kappa^g_{prior}\mu^g_{prior}{\mu^g_{pos}}^{-1}\frac{I_{d/2}(\kappa^g_{pos})}{I_{d/2-1}(\kappa^g_{pos})} \\ &+ \kappa^g_{pos}\frac{I_{d/2}(\kappa^g_{pos})}{I_{d/2-1}(\kappa^g_{pos})} \end{aligned} \tag{14}$$

Response Decoder We employ a RNN decoder similar to the one in extractor, extending it to condition on a personal feature z_r by concatenating z_r to the input of the decoder at each time step. The concrete generative process is as follows:

$$s^R_t = \sigma(s^R_{t-1}, [e_{w_{y,t-1}}, z_r]) \tag{15}$$

$$p_{vocab} = softmax(Vs^R_t + b) \tag{16}$$

[1] The z_r in $q_{\phi_g}(z|\boldsymbol{x}, \boldsymbol{y}, z_r)$ and $p_{\theta_g}(z|\boldsymbol{x}, z_r)$ is introduced formally to satisfy the consistency of the formula derivation, where, in practice, both distributions are equivalent to $q_{\phi_g}(z|\boldsymbol{x}, \boldsymbol{y})$ and $p_{\theta_g}(z|\boldsymbol{x})$ respectively.

where σ is the sigmoid function; $e_{w_{y,i}}$ is the word embedding of the i-th word in response $\boldsymbol{y}$; s_t^R denotes the hidden state at the time step t; V and b are learnable parameters; p_{vocab} stands for the probability distribution over the vocabulary. Then the objective function of the decoder is given by:

$$p_{\theta_g}(\boldsymbol{y}|\boldsymbol{x}, z, z_r) = \prod_{i=1}^{N_y} p_{vocab}(w_{y,i}) \tag{17}$$

where $p_{vocab}(w_{y,i})$ is the probability to generate the word $w_{y,i}$; N_y is the length of the response $\boldsymbol{y}$.

3.5 Training Objective

The entire SSVN model integrates two modules in Figure 1, i.e., the unsupervised extractor and the supervised generator, which can be optimized simultaneously in one framework. Thus, the overall objective function of SSVN is to maximize:

$$\mathcal{L} = \lambda \mathcal{L}_G + (1 - \lambda)\mathcal{L}_E \tag{18}$$

where we have a hyperparameter λ to control the balance between response generation (generator) and personality reconstruction (extractor).

3.6 Sampling Techique for vMF

Similar to Xu and Durrett (2018), we utilize the rejection sampling scheme of Wood (1994) to sample a value $w \in [-1, 1]$, then derive a random unit vector tangent υ on the hypersphere at the mean vector μ. Based on these, our latent variable z can be given by $z = w\mu + \upsilon\sqrt{1 - w^2}$.

4 Experiments

4.1 Datasets

The proposed model is evaluated on two datasets. The first corpus is *Cornell Movie Dialogs Corpus*[2] (Danescu-Niculescu-Mizil and Lee, 2011) that contains more than 80,000 imagined movie conversations. To normalize the length (turns) of the dialogs, we divide the original conversations into consecutive 3-10 utterances. Our second dataset is *Ubuntu Dialogue Corpus*[3] (Lowe et al., 2015). It contains about 500,000 multi-turn dialogues collected from the Ubuntu Internet Relayed Chat channel, each of which starts with a Ubuntu-related technical problem and follows by the corresponding responses about solutions.

In the above two datasets, the last utterance in a conversation is regarded as the response and the remaining ones are the input context. The detailed statistical information is shown in Table 1.

4.2 Baselines

We compare SSVN with the following models:

S2SA: the standard Seq2Seq model with the attention mechanism (Vinyals and V. Le, 2015).

HRED: a hierarchical encoder framework to model multi-turn dialogs (Serban et al., 2016).

VHRED: a hierarchical encoder-decoder with latent stochastic variable (Serban et al., 2017).

HVMN: an encoder-decoder network containing the hierarchical structure and the variational memory (Chen et al., 2018).

4.3 Experimental Details

Our model is implemented using the Tensorflow framework (Abadi et al., 2016) with the following parameter settings: We set word embeddings to size of 200 and initialize them randomly. The shared utterance encoder is a 2-layer bidirectional GRU structure with 600 hidden neurons for each layer, while the both context encoders and the both decoders are the unidirectional ones with hidden size of 600. The dimensions of the latent variable z and z_r are both set to 50. We use the Adam algorithm (Kingma and Ba, 2014) to update the parameters with an initial learning rate of 0.001. In the training, we employ the early-stop strategy (Caruana et al., 2000) to select the best models using the variational lower-bound on the validation set.

4.4 Evaluation Metrics

We use both automatic and human evaluations to analyze the model's performance.

Automatic Evaluation Metrics In our experiment, three embedding-based metrics (**Average**, **Greedy**, **Extreme**) [4] (Liu et al., 2016) are employed to measure the semantic relevance between generated responses and ground truths. Besides, we also adopt **Distinct-1** and **Distinct-2** (Li et al., 2016) to evaluate the diversity of responses.

Human Evaluation In order to assess how well the models can maintain the replier's consistency, we conduct a human evaluation. Specifically, we randomly sample 300 context from the test set and apply 5 models to generate responses for each context. For each response, three annotators are re-

[2] The dataset is available at https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html.

[3] We use the same train-validation-test split as in Chen et al. (2018).

[4] We use the embeddings trained on Google News Corpus: https://code.google.com/archive/p/word2vec/.

Corpus	Train	Valid	Test	Avg. Utterances	Avg. Words	Vocab	Coverage
Cornell	91271	871	702	5.04	16.91	10000	98.26%
Ubuntu	448833	19584	18920	4.94	23.67	20000	99.12%

Table 1: Statistical information including number of dialogs in training, validation and test sets, average number of utterances (turns) per dialog, words per utterance, vocabulary size and its proportion in corpus.

Model	Cornell Movie Dialogs Corpus					Ubuntu Dialogue Corpus				
	Average	Greedy	Extreme	Distinct-1	Distinct-2	Average	Greedy	Extreme	Distinct-1	Distinct-2
S2SA	0.1979	0.1537	0.1373	19/.0013	37/.0092	0.2156	0.1688	0.1265	1543/.0047	5342/.0260
HRED	0.4964	0.3824	0.3317	27/.0068	56/.0171	0.5415	0.4117	0.3193	1924/.0084	8549/.0409
VHRED	0.5148	0.3954	0.3446	126/.0216	329/.0642	0.5341	0.4027	0.3062	2928/.0151	13468/.0772
HVMN	0.5347	0.3876	0.3462	199/.0399	474/.1107	0.5584	0.4229	0.3220	6334/.0193	25136/.1005
SSVN	**0.6417**	**0.4582**	**0.3732**	**591/.0535**	**2535/.2450**	**0.6089**	**0.4433**	**0.3312**	**9562/.0259**	**62539/.1955**

Table 2: Automatic evaluation results on two dialogue datasets. The Distinct-* contains the number of distinct n-grams and its ratio over all generated responses. The embedding-based metrics results of baselines on Ubuntu are the same as Chen et al. (2018).

cruited to give a 4-graded judgement with the following criteria: **1**: the response is ungrammatical or semantically irrelevant; or inconsistent with replier's features (e.g., linguistic characteristics, sentiments and personalities); or has wrong logic; **2**: the response is semantically weak related, but it is too trivial (e.g., "I don't know"); **3**: the response is semantically relevant and informative, but has no obvious consistency about the replier's personal features; **4**: the response is not only semantically related and informative, but also consistent with the individual features of replier.

4.5 Evaluation Results

Automatic Evaluation The metric-based evaluation results are shown in Table 2. From the results, we can observe that:

(1) HRED performs better than S2SA, indicating that the hierarchical structure is benefical.

(2) VHRED outperforms HRED on all metrics on Cornell, which demonstrates that the latent variables are the useful global guidance information. Inversely, VHRED has a worse performance than HRED in terms of three embedding-based metrics on Ubuntu, which is consistent with Chen et al. (2018) due to the domain specific dataset.

(3) On the top of VHRED, HVMN introduces the memory network to enhance the long-term memory, and obtains the best performance among the baseline models.

(4) Compared with all the baselines, our SSVN model achieves the highest scores in terms of all metrics on two datasets, indicating that SSVN can best fit the ground truth semantically and generate more informative responses. Meanwhile, the sign tests show that the improvements of SSVN are statistically significant (p-value$<$0.01).

(5) Noticeably, the models trained on Ubuntu consistently have more distinct n-grams than the same models trained on Cornell, while the distinct ratios do not differ much. The reason is that Ubuntu dataset has more words averagely per utterance than Cornell data (as the statistical details shown in Table 1), which forces the models to produce longer responses.

Human Evaluation The human evaluation results on Cornell data are shown in Table 4, in which the score distribution values represent the percentages of responses belonging to each grade, and Fleiss' kappa (Fleiss and Cohen, 1973) is employed to evaluate the inter-annotator agreement. From the results, we have the following observations:

(1) The percentage of replier-specific responses (i.e., the grade '4') of SSVN model is 22.69%, which is much higher than that of baselines, indicating that the personal feature extractor can effectively capture the personal feature of replier.

(2) SSVN model generates much more informative responses (i.e., 71.56% labeled as '3+4') and much less generic responses (i.e., 20.28% labeled as '2') than all the baselines. The results are in line with the above results of metric-based evaluation.

(3) Kappa scores of the models are all higher than 0.4, demonstrating that the annotators come to a fair agreement. Meanwhile, the sign tests also show that the human evaluation improvements of SSVN to baselines are significant on Cornell dataset (p-value$<$0.01).

Model	Extractor	Generator	Average	Greedy	Extreme	Distinct-1	Distinct-2
SVN	–	vMF	0.6108	0.4334	0.3456	403/.0374	1376/.1367
SSVN$_{Gau}$	Gau	Gau	0.5563	0.4164	0.3406	470/.0703	1859/.3109
SSVN$_{Gau-E}$	Gau	vMF	0.6164	0.4371	0.3648	472/.0439	2100/.2091
SSVN$_{Gau-G}$	vMF	Gau	0.5605	0.4195	0.3467	525/.**0764**	2097/.**3403**
SSVN	vMF	vMF	**0.6417**	**0.4582**	**0.3732**	**591**/.0535	**2535**/.2450

Table 3: Performances of model ablation on Cornell dataset. The 'Gau' and 'vMF' denote the distributions of the latent spaces of extractor or generator, and '–' means no extractor.

Model	score distribution (%)					Kappa
	1	2	3	4	3+4	
S2SA	29.38	45.11	20.47	5.04	25.51	0.41
HRED	26.31	39.38	26.2	8.11	34.31	0.44
VHRED	17.58	36.48	33.55	12.39	45.94	0.46
HVMN	14.03	29.79	39.47	16.71	56.18	0.50
SSVN	**8.16**	**20.28**	**48.87**	**22.69**	**71.56**	0.45

Table 4: Human evaluation results on Cornell dataset. The percentages are calculated by combining the judgements from three annotators together. '3+4' stands for the sum of percentages of the grade '3' and '4' (i.e., the ratio of informative responses).

4.6 Discussions

Model Ablation To investigate the effect of different parts, we conduct a set of experiments on Cornell by removing the extractor or modifying the distribution of extractor and generator. From the results listed in Table 3, we can observe that:

(1) Removing the extractor (denoted as SVN) makes the distinct ratios and numbers drop dramatically, while the embedding-based metric scores are only slightly lower than that of SSVN. This indicates the personal features learned by the extractor not only maintain the replier-consistency, but also improve the diversity of responses. In addition, SSVN$_{Gau-E}$ (replacing the vMF distribution with a Gaussian in extractor) has a better performance than SVN, but a worse one than SSVN, demonstrating the vMF-Extractor is more effective than Gau-Extractor.

(2) As for the generator, when setting the Gaussian as the latent space (denoted as SSVN$_{Gau-G}$), the embedding-based performance deteriorates dramatically whereas the distinct numbers decrease slightly, indicating that the vMF-Generator is more capable of facilitating the generated responses semantically close to the ground truth than Gau-Generator. Notably, the distinct ratios in SSVN$_{Gau-G}$ rise unexpectedly, which will be investigated in (3).

(3) To figure out this special phenomenon, we conduct an experiment on SSVN$_{Gau}$ composed by Gau-Extractor and Gau-Generator. We can see that the SSVN$_{Gau-G}$ and SSVN$_{Gau}$ obtain the best distinct ratios among the ablation models, but their distinct numbers are not the highest. The results indicate that, whatever the latent space of extractor follows, the Gau-Generator always tends to produce informative but very short responses. Meanwhile, their worst embedding-based scores show that the responses generated by Gau-Generator semantically deviate from the ground truth significantly.

The Effect of vMF on KL Besides the metric-based performance, we also evaluate the effectiveness of different settings in sloving the latent space futility problem. Figure 2 visualizes the evolution of the KL loss in both extractor and generator parts. We can see that:

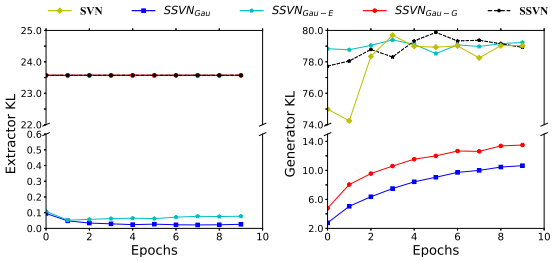

Figure 2: KL cost of different parts in ablation models as training progress.

(1) In Extractor KL, Gau-Extractors (i.e., SSVN$_{Gau}$ and SSVN$_{Gau-E}$) have a KL cost close to 0 at the begining and never recover, while the vMF-Extractors (i.e., SSVN$_{Gau-G}$ and SSVN) can keep a constant KL value as evidenced by Eq.(6). The results indicate that the vMF in extractor can mitigate KL-vanishing and capture the more meaningful personal features.

(2) Surprisingly, in Generator KL, the KL loss presents an upward trend in Gau-Generators (i.e., SSVN$_{Gau}$ and SSVN$_{Gau-G}$). The reason is that, the personal features from the extractor can effectively strengthen the expressiveness of the latent space in the generator, thus the response decoder

is encouraged to exploit the latent variables and the latent space futility problem is alleviated.

(3) Compared with the Gau-Generators in Generator KL, the vMF-Generators (i.e., SVN, SSVN_{Gau-E} and SSVN) have the much higher KL values, indicating that the vMF is a better selection than Gaussian to solve the KL-vanishing problem. Meanwhile, the KL values are relatively stable, which experimentally demonstrates the KL cost mainly depends on the last term in Eq.(14) and the variable term has little effect on it. Last but not least, KL cost can explictly be changed by setting different kappa values.

Impact of the Coefficient λ Recall that Eq.(18) shows the capacity of SSVN in balancing response generation and personality reconstruction. Here we analyze the effects of different coefficient λ on the quality of responses. Figure 3 shows the performances given varying λ. Notably, the performances of embedding-based metrics are changing in a similar trend, as the same case in Distinct-1 and Distinct-2, thus we only consider Average and Distinct-1 as the major analysis items.

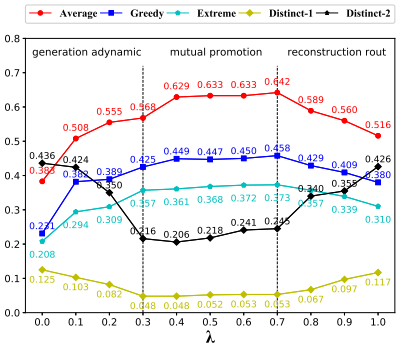

Figure 3: Influences of different λ on Average, Greedy, Extreme, Distinct-1 and Distinct-2.

As we can see, the evolutions of Average and Distinct-1 in Figure 3 can be broadly into three stages: generation adynamic stage, mutual promotion stage and reconstruction rout stage.

(1) The first stage shows that as λ increases, the Average monotonically increases while diversity decreases. This is because the lower λ gives the model less incentive to optimize the generator, which makes the response decoder incapable of utilizing the higher-quality personal features, resulting in the diverse but semantically inappropriate responses.

(2) When λ moves to the second stage, the performances of the Average and diversity improve simultaneously, implying that the response generation and personality reconstruction achieve en expected balance.

(3) For the reconstruction rout stage, although the model focuses on response generation, the larger λ does not bring an improvement of Average, but instead increases diversity. The result indicates that the unprofitable personal features in thwarted personality reconstruction part, as a random disturbance, can increase the diversity of the responses, but severely bias the generation of response decoder semantically.

Observed from the curves of all metrics, the best performance of embedding-based metrics is achieved at $\lambda = 0.7$, while the diversity reaches the peak in the mutual promotion stage. Thus, we set λ to 0.7 in all previous experiments.

Case Study Besides the quantitative analysis, we also organize some examples (seen in Table 5) from different models to analyze the performances of the methods qualitatively. They are chosen randomly from the responses produced by the proposed model, and showed together with the corresponding contexts and the outputs of the baselines. From the case 1, we can observe that the SSVN can extract the personal feature of the replier that the speaker A prefers to acquire further information from others, which guides the generator to produce an interrogative response to promote the replier-consistency. Meanwhile, the SSVN can also extract the firm attitude of the replier in case 2 and the pleading tone of the replier in case 3. By contrast, the baselines favor to produce the bad responses, such as containing more 'unk'.

Error Analysis To improve the performance of SSVN in the future, we take the worse cases (i.e., the grade '1' and '2') in human judgement as an example to analyze our errors. Specifically, we divide the errors with the grade '1' into grammatical error, replier-nonconsistency and logic contradiction, which occupy 19.79%, 31.49% and 48.72%, respectively. We can find that 1) logic condradiction scenes make up most of the errors as SSVN pays little attention to this issue. 2) although considering the personal features from replier, there still exists 31.49% replier-inconsistent cases, incidating that only strengthening the VAE with the vMF distribution may not be a perfect approach for personal feature extraction. As for grade '2', the consistency of replier's features improve the of response diversity significantly, but the model still has "safe response" problem as the baseline.

ID	Context	Response
1	A: Where are we going? B: We're getting out of this place. We've done enough. ground truth: Wait a minute, whoa. We can't go now. What's ⟨unk⟩ got planned with that nerve agent?	SSVN: When we're going out? S2SA: We know he is on you. HRED: I'll ⟨unk⟩ there's no ⟨unk⟩, or there's no this case. ⟨unk⟩ ⟨unk⟩ well, because he is the ⟨unk⟩. VHRED: It motel yet . HVMN: What kind of ⟨unk⟩ were you?
2	A: There. B: I beg your pardon? A: There wolf! There castle! B: Why are you talking like that? ground truth: I thought you wanted to.	SSVN: Well, I want you to get out of here! S2SA: Me ⟨unk⟩., HRED: So how do you believe that? VHRED: Fourth ⟨unk⟩ she wasn't ⟨unk⟩! HVMN: Old ⟨unk⟩. These ⟨unk⟩ is ⟨unk⟩ – ⟨unk⟩ ⟨unk⟩. ⟨unk⟩ years ...
3	A: James ... B: I have to go. A: Then take me with you. B: No. You'll be safe here. ground truth: I don't want to be safe!	SSVN: Please, I have to go. I can't take a walk. S2SA: Tell you now ⟨unk⟩ something of ⟨unk⟩, doesn't ⟨unk⟩! HRED: ⟨unk⟩ just actually, the ⟨unk⟩ ⟨unk⟩ of making the ⟨unk⟩ ⟨unk⟩. She's dead. VHRED: Gross ... HVMN: No.

Table 5: Generated responses of all the models on Cornell Movie Dialogs Corpus.

The above analysis sheds light on our future directions: 1) modeling the logic consistency between the context and response; 2) exploring advanced methods for extracting personal features; 3) improving the response diversity.

5 Conclusion and Future Work

In this work, we propose a semi-supervised stable variational network for addressing the latent space futility and replier-consistency decay issues. Different from the traditional variational dialog models, the proposed model selects the vMF as the prior and posterior to resolve the latent space futility issue, and then integrates a unsupervised extractor to obtain the replier-tailored personal features to ensure the replier-consistency. Experimental results on two dialog datasets demonstrate the effectiveness of our model, especially on replier-consistency in terms of human evaluation. However, the error analysis shows that there are still challenges in dialogue generation, which we would like to explore in the future.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. Our work is supported by National Natural Science Foundation of China (61976154), Tianjin Natural Science Foundation (18JCYBJC15500), National Natural Science Foundation of China (61771333), Tianjin Municipal Science and Technology Project (18ZXZNGX00330), and the Foundation of State Key Laboratory of Cognitive Intelligence, iFLYTEK(CIOS-20190001).

References

Martn Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, and Xiaoqiang Zheng. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467[cs]*, arXiv:1603.04467.

Rich Caruana, Steve Lawrence, and C. Lee Giles. 2000. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in neural information processing systems 13 (NIPS)*, pages 402–408.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017a. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35.

Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical variational memory network for dialogue generation. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*, pages 1653–1662.

Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2017. Variational lossy autoencoder. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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 2nd Workshop on Cognitive Modeling and Computational Linguistics (CMCL '11)*, pages 76–87.

Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. 2018. Hyperspherical variational auto-encoders. *arXiv:1804.00891[cs]*, arXiv:1804.00891.

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.

Barbara J. Grosz. 2016. Ai100 report. https://ai100.stanford. edu/2016-report.

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics (TACL)*, 6:437–450.

D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 3581–3589.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR)*.

Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, and Xi Chen. 2016. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 4743–4751.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

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. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2122–2132.

Ryan Lowe, Nissan Pow, Iulian V. Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 285–294.

Danilo Jimenez Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1530–1538.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31nd International Conference on Machine Learning (ICML)*, pages 1278–1286.

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

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. 2017. A hybrid convolutional variational autoencoder for text generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–637.

Iulian Serban, Alessandro Sordoni, Ryan Joseph Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 3295–3301.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 3776–3784.

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

Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018. Improving variational encoder-decoders in dialogue generation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5456–5463.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 3483–3491.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 3104–3112.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*.

Andrew Wood. 1994. Simulation of the von mises fisher distribution. *Communications in Statistics Simulation and Computation*, 23(1):157–164.

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Levy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. 2017. Data noising as smoothing in neural network language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jiacheng Xu and Greg Durrett. 2018. Spherical latent spaces for stable variational autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4503–4513.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kir. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3881–3890.

Kaisheng Yao, Geoffrey Zweig, and Baolin Peng. 2015. Attention with intention for a neural network conversation model. *ArXiv*, abs/1510.08565.

Saizheng Zhang, Emily Dinanz, Jack Urbanekz, Arthur Szlamz, Douwe Kielaz, 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 (ACL)*, pages 1–10.

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 (ACL)*, pages 654–664.