Exploiting Explicit and Inferred Implicit Personas for Multi-turn Dialogue Generation

Ruifang Wang[1,2], Ruifang He[1,2(✉)], Longbiao Wang[2(✉)], Yuke Si[2], Huanyu Liu[2], Haocheng Wang[2], and Jianwu Dang[2,3]

[1] State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing, China

[2] Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
{ruifang_wang,rfhe,longbiao_wang,siyuke,huanyuliu,haochengwang}@tju.edu.cn

[3] Japan Advanced Institute of Science and Technology, Ishikawa, Japan
jdang@jaist.ac.jp

Abstract. Learning and utilizing personas in open-domain dialogue have become a hotspot in recent years. The existing methods that only use predefined explicit personas enhance the personality to some extent, however, they cannot easily avoid persona inconsistency and weak diversity responses. To address these problems, this paper proposes an effective model called Exploiting **E**xplicit and Inferred **I**mplicit **P**ersonas for Multi-turn **D**ialogue Generation (**EIPD**). Specifically, 1) an explicit persona extractor is designed to improve persona consistency; 2) Taking advantage of the von Mises-Fisher (**vMF**) distribution in modeling directional data (e.g., the different persona state), we introduce the implicit persona inference to increase diversity; 3) during the generation, the persona response generator fuses the explicit and implicit personas in the response. The experimental results on the ConvAI2 persona-chat dataset demonstrate that our model performs better than commonly used baselines. Further analysis of the ablation experiments shows that EIPD can generate more persona-consistent and diverse responses.

Keywords: Persona-based dialogue generation · Implicit personas · vMF

1 Introduction

With the development of open-domain dialogue system, great progress has been achieved in many fields, such as intelligent assistants, customer service and chatbots [3]. The end-to-end network [14] has been proven effective for generative dialogue systems. However, it is still difficult to build more engaging and realistic conversations owing to the lack of interlocutor personas.

Several efforts have been made to explore the abilities of personas for facilitating response generation [7]. [20] introduced a novel dataset, PERSONA-CHAT,

L. Wang et al. (Eds.): NLPCC 2021, LNAI 13028, pp. 493–504, 2021.
https://doi.org/10.1007/978-3-030-88480-2_39

Personas	
1. *My skin is olive colored.*	2. My purse has a picture of a skunk on it.
3. *My eyes are green.*	4. *I wear glasses that are cateye.*
5. *I want to be a librarian.*	

Dialogue1	Dialogue2
Context1: Hi, how are you? please tell me about yourself !	**Context1**: You sound very pretty ! what else do you like ?
Context2: Hello, blonde hair, blue eyes. and yourself ?	**Response**: Books ! which probably explains why I'm studying to become a librarian. you ?
Response: Yes , I've olive skin color with green eyes and cateye glasses.	

Fig. 1. The two dialogues from ConvAI2 persona-chat, where the same colors of sentences imply that the sentences are related to each other in the conversation.

where each dialogue is assigned a character description using 5 sentences as a persona profile. We define these persona profiles as **explicit personas**. Then, [12,13] generated responses with this kind of personas. However, in real conversations, sometimes the repliers answer with explicit personas directly, and sometimes they answer with some useful information that can be inferred from the explicit personas and context, which we define as **implicit personas**. Specifically, as shown in Fig. 1, the two dialogues are associated with the same explicit personas. The response in Dialogue1 is directly associated with explicit personas, *'My skin is olive colored. My eyes are green. I wear glasses that are cateye.'*. It describes the image of the speaker which are consistent with the context. Therefore, how to capture context-relevant personas is essential in persona-based dialogue. However, in Dialogue2, the response not only mentions the persona *'I want to be a librarian.'* but also explain the reason why the speaker wants to be a librarian. This kind of information does not appear in the context and explicit personas, but it can be inferred from persona-based context. This indicates it is possible to use implicit personas in some responses. Although some persona-based dialogue methods have been proposed, the following challenges still exist: 1) In multi-turn dialogue, as shown in Fig. 1, the response is related to some contextual personas, and previous methods cannot effectively capture the key explicit personas, which is not conducive to persona consistency. 2) In the persona-based dialogue, the attractive responses are not only persona-consistent but also diverse, while the existing methods mainly focus on persona consistency. 3) Previous methods usually take explicit personas into consideration, but neglect that both explicit and implicit personas mentioned above can interact with each other in one model at the same time.

To tackle these challenges, we propose a model called Exploiting Explicit and Inferred Implicit Personas for Multi-turn Dialogue Generation (EIPD), which consists of three components. Specifically, the explicit persona extractor mainly adopts a transformer encoder to acquire some explicit personas relevant to the context. Second, the implicit persona inference module employs the von Mises-Fisher (vMF) distribution, which is suitable for modeling directional data to

reason the implicit personas and improve the response diversity. Third, the persona-response generator is designed to guide the implicit personas and fuse the two kinds of personas to generate the response. Finally, the ConvAI2 persona-chat dataset is used to evaluate the effectiveness of proposed model. We summarize the contributions of this work as follows:

- It is the first time that an effective framework for multi-turn dialogue generation takes two kinds of personas into consideration simultaneously.
- An implicit personas inference module with an vMF distribution is devised to reason the implicit personas.
- The persona generator is used to supervise the generation of implicit personas.
- The experimental results demonstrate that our model can generate responses with more diversity and persona consistency compared with baseline results.

2 Related Work

2.1 Persona-Based Dialogue Model

In open-domain dialogue generation, the persona-based dialogue model has attracted an increasing number of researchers' attention. Recent works focus on improve the persona-based dialog generation performance as well as persona consistency. [11] assigned a desired identity to chatbot which can generate coherent response. [20] constructed a persona-chat dataset with different speaker profiles. Based on this dataset, [13] proposed an Reinforcement Learning framework to improve persona consistency of response. Besides these works using speaker profiles, other works using implicit information to achieve it. [7] used pretrained speaker embeddings and dialogue context to boost informative and diverse response. [10] proposed a multi-task learning approach that incorporated speaker characteristics to train the neural conversation models. Despite the success of using implicit persona in conversation, they are still difficult to learn implicit personas displayed by the speakers automatically.

2.2 von Mises-Fisher Distribution

The von Mises-Fisher(vMF) distribution represents a latent hyperspherical space which can model directional data better. Considering this characteristic, the vMF distribution is introduced into some NLP works. Both [1] and [9] integrated vMF into a topic model to explore the semantic consistency and to improve the performance. [18] replaced Gaussian distribution with vMF distribution in CVAE and discovered that the 'collapse' problem can also be alleviated. [5] used vMF distribution to draw the context word vectors to improve the embedding models. Different from these works, we apply vMF distribution in the Conditional Variational Autoencoder(CVAE) framework to infer the implicit personas.

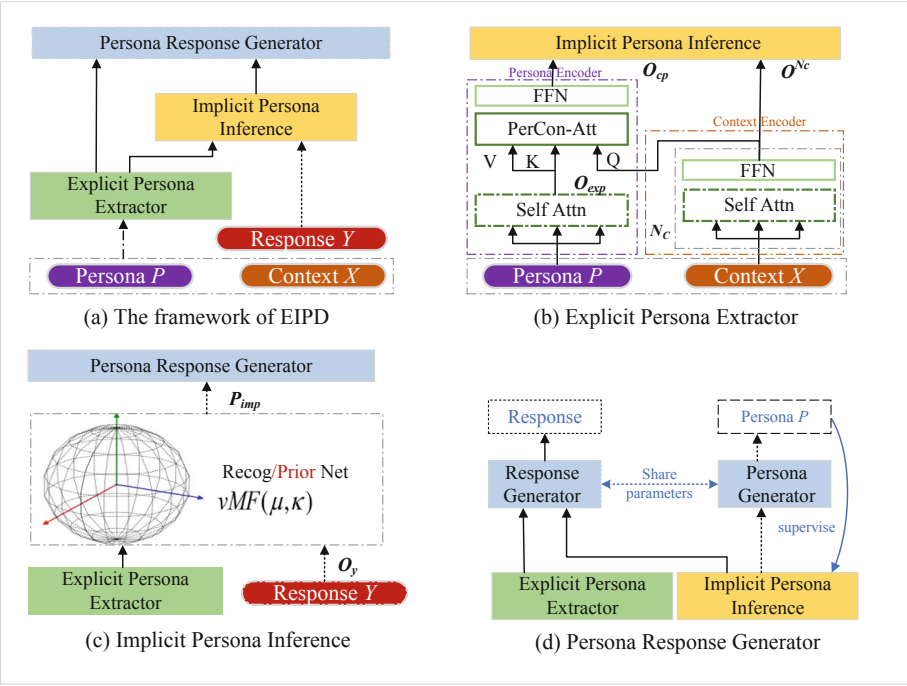

Fig. 2. The framework of the EIPD model, including explicit persona extractor, implicit persona inference and persona response generator. The process represented by the dotted line only occurs during the training.

3 The Proposed Model

A persona-based dialogue system generates responses with context and personas. Our problem is formulated as follows: context $X = \{x_1, x_2, ..., x_m\}$, each utterance $x_i = (w^x_{i,1}, w^x_{i,2}, ..., w^x_{i,M_i})$, a set of explicit personas $P_{exp} = \{p_1, p_2, ..., p_n\}$, each persona $p_i = (w^p_{i,1}, w^p_{i,2}, ..., w^p_{i,N_i})$, and response $Y = \{w^y_1, w^y_2, ..., w^y_k\}$. Given X, the implicit personas P_{imp} are explored by the implicit persona inference module with the supervision of explicit personas. By leveraging the context, explicit personas, and implicit personas, the goal is to generate a diverse and persona-consistent response Y. We drop the subscript of P_{exp} for simplicity.

As shown in Fig. 2(a), the whole framework can be divided into three modules: (1) Explicit Persona Extractor, (2) Implicit Persona Inference, and (3) Persona Response Generator.

3.1 Explicit Persona Extractor

Following Transformer [15], this component (Fig. 2(b)), which includes a context encoder and a persona encoder, takes context and explicit personas as the input and extracts the most relevant explicit personas to improve persona consistency.

Context Encoder: We use the transformer encoder to encode the context X. The multi-head self-attention is defined as MultiHead(Q, K, V), where Q, K, and V represent query, key, and value, respectively. The encoder is composed of N_c layers. The encoding of context is as follows:

$$H_c^n = \text{MultiHead}(O_c^{n-1}, O_c^{n-1}, O_c^{n-1}) \tag{1}$$

$$O_c^n = \text{FFN}(H_c^n) \tag{2}$$

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{3}$$

where n $\in$ (2,N_c). H_c^n and O_c^n are the *n-th* layer output of the multi-head self-attention and feed-forward network, respectively. In the first layer, O_c^1 represents the word embedding and positional embedding of the input. Following [15], we also add layer normalization to the sub layers, and we can finally obtain the context representation O^{N_c} after N_c layers.

Persona Encoder: According to the examples in Fig. 1, we observe the following: 1) The response Y is often related to some personas p_i and contexts X_j. 2) The relevance between X and p_i is beneficial to generate an informative and consistent response. Therefore, we want to consider them. Specifically, we use another multi-head self-attention to encode the explicit personas. O_{exp} represents the output of this attention mechanism. We then use PerCon-Attention which takes O^{N_c} as query, O_{exp} as key and value to compute the contextual explicit persona hidden vector O_{cp} based on the following equations:

$$H_{cp} = \text{PerConAtt}(O^{N_c}, O_{exp}, O_{exp}) \tag{4}$$

$$O_{cp} = \text{FFN}(H_{cp}) \tag{5}$$

3.2 Implicit Persona Inference

According to Fig. 1, we can see that the personas shown in the response are not entirely extracted from the given explicit personas. We therefore employ an inference module using vMF distribution to reason the implicit personas (Fig. 2(c)) for the personalized and diverse responses.

Since different speakers express different implicit personas, this information can be represented in different directions in the semantic space. The vMF distribution [18] can model directional data better, therefore, we introduce it into the CVAE framework. Specifically, in the CVAE framework, the prior network $p_\theta(z|X, P)$ and the recognition network $q_\varphi(z|X, P, Y)$ are used to sample the latent variable z, namely, implicit personas, and can be written as p_{imp}. In our settings, p_{imp} follows the vMF distribution, specifically the prior network $p_\theta(p_{imp}|X, P) \sim vMF(\mu_{prior}, \kappa_{prior})$ and the posterior network $q_\varphi(p_{imp}|X, P, Y) \sim vMF(\mu_{pos}, \kappa_{pos})$.

VMF Distribution: The von Mises-Fisher distribution is defined over a hypersphere of unit norm, depending on the direction vector $\mu \in R^m$ with $||\mu|| = 1$ and a concentration parameter $\kappa \in R_{\geq 0}$, where m denotes the dimension of the word vectors. The Probability Density Function of the vMF distribution for a random unit vector $z \in R^m$ is defined as:

$$f_m(p_{imp}; \mu, \kappa) = C_m(\kappa) \exp(\kappa \mu^{\mathrm{T}} p_{imp}) \tag{6}$$

$$C_m(\kappa) = \frac{\kappa^{m/2-1}}{(2\pi)^{m/2} I_{m/2-1}(\kappa)} \tag{7}$$

where $C_m(\kappa)$ is the normalization constant and $I_{m/2-1}$ stands for the modified Bessel function of the first kind at order v. Inspired by NVSRN [2], we encode Y into representations O_y, set κ_{prior} and κ_{pos} as constants and compute μ_{prior}, μ_{pos} as:

$$\mu_{pos}^{\sim} = f_{pos}([O^{N_c}, O_{cp}, O_y]) \tag{8}$$

$$\mu_{pos} = \mu_{pos}^{\sim} / ||\mu_{pos}^{\sim}|| \tag{9}$$

$$\mu_{prior}^{\sim} = f_{prior}([O^{N_c}, O_{cp}]) \tag{10}$$

$$\mu_{prior} = \mu_{prior}^{\sim} / ||\mu_{prior}^{\sim}|| \tag{11}$$

where f_{prior} and f_{pos} are two transformations and $||\cdot||$ denotes the 2-norm used to ensure the normalization. Since the prior $p_\theta(p_{imp}|X, P)$ follows the $vMF(\mu_{prior}, \kappa_{prior})$ rather than $vMF(\cdot, 0)$, the KL divergence will be computed as:

$$\begin{aligned} \mathcal{L}_{KL} &= KL(q_\varphi(p_{imp}|X, Y, P)||p_\theta(p_{imp}|X, P)) \\ &= (m/2-1)\log\frac{\kappa_{pos}}{\kappa_{prior}} + \log\frac{I_{m/2-1}(\kappa_{prior})}{I_{m/2-1}(\kappa_{pos})} \\ &- \kappa_{prior}\mu_{prior}\mu_{pos}^{-1}\frac{I_{m/2}(\kappa_{pos})}{I_{m/2-1}(\kappa_{prior})} + \kappa_{pos}\frac{I_{m/2}(\kappa_{pos})}{I_{m/2-1}(\kappa_{prior})} \end{aligned} \tag{12}$$

Sampling Technique for vMF: Following the implementation of [4], we use the rejection sampling scheme to sample $w \in [-1, 1]$, and then the latent variable p_{imp} is derived from $p_{imp} = w\mu + v\sqrt{1-w^2}$, where v is a randomly sampled unit vector tangent to the hypersphere at μ.

3.3 Persona Response Generator

This component comprises a response generator and a persona generator (Fig. 2(d)). Considering the interaction between the two kinds of personas, we use the two generators to further enhance the modeling of directional data and better fuse the implicit and explicit personas.

Persona Generator: To strengthen the supervision for implicit personas, during this process, we employ an RNN decoder that receives implicit persona p_{imp} as the initial hidden state and then generates tokens sequentially under the probability distributions:

$$p_{\theta_p}(P|p_{imp}) = \prod_{i}^{n} \prod_{j=1}^{N_i} p(w_{i,j}|P_{<i}, w_{i<j}) \tag{13}$$

where n is the number of turns of explicit personas; N_i is the length of the i-th utterance p_i. During this process, the loss function is:

$$\mathcal{L}_p = \mathbf{E}_{q_\varphi(p_{imp}|X,P,Y)}[\log p_\theta(P|p_{imp})] \tag{14}$$

Response Generator: Finally, conditioned based on explicit personas, implicit personas and context, we employ a response decoder to generate the response Y:

$$p_{\theta_g}(Y|X, P, p_{imp}) = \prod_{i=1}^{k} p_{vocab}(w_{y,i}) \tag{15}$$

where p_{vocab} is the vocabulary's probability distribution; $p_{vocab}(w_{y,i})$ is the probability of the word $w_{y,i}$; k is the length of the response Y. In general, the ELBO in the decoder can be rewritten as:

$$\mathcal{L}_r = \mathbf{E}_{q_\varphi(p_{imp}|X,Y,P)}[\log p_\theta(Y|p_{imp}, X, P)] - \mathcal{L}_{KL} \tag{16}$$

3.4 Training Objective

In the EIPD model, the overall objective is:

$$\mathcal{L} = \lambda \mathcal{L}_p + (1 - \lambda)\mathcal{L}_r \tag{17}$$

where the hyperparameter λ is used to control the balance between response generator and persona generator.

4 Experiments

4.1 Experimental Settings

Dataset: We use the released ConvAI2 persona-chat dataset, which is an extended version of PERSONA-CHAT [20][1], to verify our proposed method. The dataset consists of 164,356 utterances in 10,981 dialogues, and each speaker has at least 4 persona profiles. We randomly split the data into the training, validation, and test sets, which respectively contain 67112, 8395, and 4478 dialogues.

[1] http://convai.io/.

Baselines: We compared the proposed EIPD model with five commonly used baseline models. **S2SAP**: the Seq2Seq model, which integrates context and persona as the input [20]. **CVAE**[2]: an RNN-based model that exploits latent variables to improve the diversity of the response [21]. **Trans**[3]: the transformer model [15] that concatenates personas and context as the input. **PerCVAE**[4]: a memory augmented CVAE model that uses multi-hop attention to exploit the persona information to improve the response quality [12]. **TransferTransfo**[5]: a finetuned GPT2 that takes personas and dialogue context as the input [16] (Table 1).

Table 1. Objective (on the left) and subjective evaluation (on the right) results with respect to the ConvAI2 persona-chat dataset. Results in bold represent the best scores. In the subjective evaluations, the percentages of each kind of response are calculated by combining the evaluations from three annotators together. The Kappa scores of all models are higher than 0.4, which indicates that the three annotators reach a fair agreement.

Model	Dist-1	BLEU-1	BLEU-2	F1	G1	G2	G3	G4	G3&4
S2SAP	0.0151	0.1467	0.1439	0.2309	38.25	29.67	27.35	4.73	32.08
CVAE	0.0165	0.1356	0.1502	0.1903	37.25	25.00	26.00	11.75	37.75
Trans	0.0267	0.1531	0.1621	0.1921	32.25	25.50	29.00	13.25	44.25
PerCVAE	0.0374	0.2047	0.1858	0.2404	21.43	18.73	39.25	20.59	59.84
TransferTransfo	0.0332	**0.2532**	0.2249	0.1973	20.34	17.14	42.73	19.79	62.52
EIPD	**0.0388**	0.2263	**0.2323**	**0.2452**	**18.36**	**14.75**	**44.75**	**22.14**	**66.89**

Parameters: For the RNN-based models, we set word embeddings to the size of 300. The encoder is a 2-layer GRU structure with a hidden size of 600. For the Transformer, the size of word embedding is set to 512, and the numbers of layers of encoder and decoder are set to 3 and 1. Besides, the number of heads in multi-head attention is 8, and the inner-layer size of the feed-forward network is 2048. In our model, the parameters of the explicit persona extractor are the same as those of Transformer. The dimension of the latent variable is set to 180. We use the Adam algorithm to update the parameters with a learning rate of 0.0001. The batch size is set to 32. An early-stop strategy is used to obtain the best model. Our model is implemented using the Tensorflow framework. We conduct all experiments on a GPU.

Evaluations: In our experiments, we use Dist-1, BLEU-1/2 and F1 to evaluate our method. In addition to the automatic metrics, we recruit three human annotators familiar with the NLP tasks to judge the quality of the generated

[2] https://github.com/snakeztc/NeuralDialog-CVAE.
[3] http://github.com/atselousov/transformerchatbot.
[4] https://github.com/vsharecodes/percvae.
[5] http://github.com/huggingface/transfer-learning-conv-ai.

responses. We sampled 200 context-response-persona triples from the above models. They are required to provide 4-graded judgements according to the following criteria: **G1**: The generated response is not grammatically correct, is irrelevant to the semantics of context or is inconsistent with the given personas. **G2**: The generated response is fluent and weakly related to the context, such as some generic responses. **G3**: The generated response is fluent and relevant to the context semantics and slightly consistent with the personas. **G4**: The generated response is not only fluent and semantically relevant but also consistent with the given personas.

Table 2. Performances of model ablation. EIPD is significantly better than the ablation approaches.

Model	Dist-1	BLEU-1	BLEU-2	F1
D	0.0007	0.1248	0.1365	0.2038
IPD	0.0301	0.1465	0.1526	0.2186
EPD	0.0354	0.2142	0.2053	0.2439
EIPD_{Gau}	0.0345	0.2171	0.2064	0.2348
EIPD_{pd}	0.0363	0.2121	0.2105	0.2208
EIPD	**0.0388**	**0.2263**	**0.2323**	**0.2452**

4.2 Experimental Results

Objective and Subjective Evaluations: For objective evaluation, (1) Dist-1 is the ratios of distinct unigrams which can reflect the diversity of the generated response. It can be found that the performance of S2SAP is the worst because it only roughly combines the explicit personas. PerCVAE surpassed other baselines due to the exploitation of explicit personas. Compared with the baselines, EIPD outperforms them, which indicates that the proposed model can generate diverse responses. (2) BLEU-1/2 evaluates how many n-grams (n = 1,2) in the generated responses overlap with them in the ground truth. EIPD performs better than baselines except for TransferTransfo in BLEU-1, and we speculate that the reason may be that the pretrained language model contains semantic information. (3) For F1, the score of EIPD is higher than others, demonstrating that the model can generate more accurate information.

For subjective evaluation, the responses generated by EIPD are more engaging as compared to the responses from all baselines. It can be determined that the percentage of diverse and persona-consistent responses (the grade 'G3&4') is 66.89%, obviously higher than others, which indicates that EIPD can generate persona-consistent responses. Additionally, the percentage of 'G2' is declining, while, the percentage of 'G3' is rising. This proves that EIPD has the ability to generate context-relevant responses, and alleviate the problem of generic responses at the same time. Among the baselines, the results of S2SA perform poorly since it the model does not take any kind of personas into consideraiton.

By adding explicit personas or global information, the performance of these models improve gradually, yet still worse than our model.

Ablation Analysis: To investigate the effects of specific modules in EIPD, we ablated our model through several different approaches: **D**: A generative dialog model without explicit and implicit personas. **EPD**: It removes the implicit persona inference, that is, the model does not use implicit personas. **IPD**: It replaces the explicit persona extractor with the RNN to represent the explicit personas. **EIPD$_{\mathbf{Gau}}$**: This model replaces the vMF distribution with the Gaussian distribution. **EIPD$_{\mathbf{pd}}$**: This approach deletes the persona generator, so the generation of implicit personas loses the supervision of the explicit personas.

As shown in Table 2, from D, IPD, EPD, EIPD$_{\text{Gau}}$, and EIPD$_{\text{pd}}$ to EIPD, every step yields an observed improvement on the automatic metrics. EIPD achieves the best performance among all the methods. Specifically, compared with D, the improvements of EPD and IPD on all metrics imply that the explicit persona extractor can capture the explicit personas related to some context, and the implicit persona inference module can obtain the implicit personas inferred from the given context and explicit personas. Furthermore, we note that EIPD performs better than EIPD$_{\text{Gau}}$ on all metrics, which proves that the vMF distribution is more useful than the Gaussian distribution in this framework. Specifically, the implicit persona inference module can reason the more rational implicit personas with vMF distribution, and this phenomenon is consistent with the characteristics of vMF, which is good at modeling directional data, such as the personalities of different speakers. In addition, the performance of EIPD$_{\text{pd}}$ is inferior to EIPD, which verifies that the persona generator can facilitate the generation of persona-consistent and diverse responses.

Table 3. An example of dialogue with the personas 'Black coffee is my addiction. My favorite hobby is gardening. My family gets together every Saturday. My husband died last year.' in ConvAI2 persona-chat dataset.

Context
A:Hello how are you doing?
B:I'm good how are you?
A:Good thanks. So what is life like for you?
SASP:I'm a student and I work a lot and a lot.
CVAE:I am, as little since I am super excited about me, school.
Trans:I like to play games with friends what about you?
PerCVAE:I would like to talk about you. Would you like a happy person?
TransferTransfo:That's so good, I prefer to read.
Golden:I am currently struggling in school.
EIPD-1:It is ok. I do not get much done unless I work on my garden.
EIPD-2:It is ok now. I like to walk outside and explore the outdoors.

Case Study: According to Table 3, we can determine that the baseline models often generate some fluent but irrelevant and weak personalized responses. For comparison, we use the EIPD to generate different responses through implicit persona inference, and we find that the responses are related to the personas *'My favorite hobby is gardening'*. The first response directly answers the speaker's attitude about *gardening*, and the second response expands the information about the given personas.

5 Conclusion and Future Work

In this paper, we propose an effective EIPD for multi-turn persona-based dialogue. To the best of our knowledge, we are the first to fuse the explicit personas and implicit personas to generate more realistic responses. It uses an explicit persona extractor to improve the persona consistency, and employs an implicit persona inference module with vMF distribution to improve the diversity. Finally, the persona response generator is used to fuse personas and generate the response. Experimental results on ConvAI2 persona-chat dataset demonstrate the effectiveness of our model and verify the importance of implicit personas. In the future, we would like to use knowledge graphs and pretrained language model to strengthen the inference of implicit personas.

Acknowledgement. This work was supported by the National Key RD Program of China under Grant 2018YFB1305200, the National Natural Science Foundation of China under Grant (61771333, 61976154), the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330, and the State Key Laboratory of Communication Content Cognition, People's Daily Online (No. A32003).

References

1. Batmanghelich, K., Saeedi, A., Narasimhan, K., Gershman. S.: Nonparametric spherical topic modeling with word embeddings. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 537–542 (2016)
2. Chang, J., et al.: NVSRN: a neural variational scaling reasoning network for initiative response generation. In: 2019 IEEE International Conference on Data Mining, pp. 51–60 (2019)
3. Chen, H., Liu, X., Yin, D., Tang, J.: A survey on dialogue systems: recent advances and new frontiers. ACM SIGKDD Explor. Newslett. **19**(2), 25–35 (2017)
4. Guu, K., Hashimoto, T.B., Oren, Y., Liang, P.: Generating sentences by editing prototypes. Trans. Assoc. Comput. Linguist. **6**, 437–450 (2018)
5. Jameel, S., Schockaert, S.: Word and document embedding with vMF-mixture priors on context word vectors. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3319–3328 (2019)
6. Kingma, D., Mohamed, S., Jimenez, D., Welling, M.: Semi-supervised learning with deep generative models. In: Advances in Neural Information Processing Systems, pp. 3581–3589 (2014)

7. Kottur, S., Wang, X., Carvalho, V.: Exploring personalized neural conversational models. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-2017, pp. 3728–3734 (2017)
8. Li, J., Galley, M., Brockett, C., Spithourakis, G.P., Gao, J., Dolan, B.: A persona based neural conversation model. arXiv preprint arXiv:1603.06155 (2016)
9. Li, X., Chi, J., Li, C., Ouyang, J., Fu, B.: Integrating topic modeling with word embeddings by mixtures of vMFs. In: COLING 2016, 26th International Conference on Computational Linguistics, pp. 151–160 (2016)
10. Luan, Y., Brockett, C., Dolan, B., Gao, J., Galley, M.: Multi-task learning for speaker-role adaptation in neural conversation models. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing, pp. 605–614 (2017)
11. Qian, Q., Huang, M., Zhao, H., Xu, J., Zhu, X.: Assigning personality identity to a chatting machine for coherent conversation generation. arXiv preprint arXiv:1706.02861 (2017)
12. Song, H., Zhang, W., Cui, Y., Wang, D., Liu, T.: Exploiting persona information for diverse generation of conversational responses. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-2019, pp. 5190–5196 (2019)
13. Song, H., Zhang, Hu, J., Liu, T.: Generating persona consistent dialogues by exploiting natural language inference. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, pp. 8878–8885 (2020)
14. Sutskever, I., Vinyals, O., Le, Q.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
15. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
16. Wolf, T., Sanh, V., Chaumond, J., Delangue, C.: Transfertransfo: a transfer-learning approach for neural network based conversational agents. arXiv preprint arXiv:1901.08149 (2019)
17. Wu, B., et al.: Guiding variational response generator to exploit persona. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 53–65 (2020)
18. Xu, J., Durrett, G.: Spherical latent spaces for stable variational autoencoders. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4503–4513 (2018)
19. Yao, K., Zweig, G., Peng, B.: Attention with intention for a neural network conversation model. arXiv preprint arXiv:1510.08565v3 (2015)
20. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: I have a dog, do you have pets too? arXiv preprint arXiv:1801.07243 (2018)
21. Zhao, T., Zhao, R., Eskenazi, M.: 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, pp. 654–664 (2017)