Improving Personality Consistency in Conversation by Persona Extending

Yifan Liu
CCIIP Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology
Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL)
China
yifaan@hust.edu.cn

Wei Wei*
CCIIP Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology
Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL)
China
weiw@hust.edu.cn

Jiayi Liu
Alibaba Group
China
ljy269999@alibaba-inc.com

Xianling Mao
Beijing Institute of Technology
China
maoxl@bit.edu.cn

Rui Fang
Ping An Property & Casualty Insurance company of China, Ltd
China
fangrui051@pingan.com.cn

Dangyang Chen
Ping An Property & Casualty Insurance company of China, Ltd
China
chendangyang273@pingan.com.cn

ABSTRACT

Endowing chatbots with a consistent personality plays a vital role for agents to deliver human-like interactions. However, existing personalized approaches commonly generate responses in light of static predefined personas depicted with textual description, which may severely restrict the interactivity of human and the chatbot, especially when the agent needs to answer the query excluded in the predefined personas, which is so-called out-of-predefined persona problem (named **OOP** for simplicity). To alleviate the problem, in this paper we propose a novel *retrieval-to-prediction* paradigm consisting of two subcomponents, namely, (1) ***Persona Retrieval Model (PRM)***, it retrieves a persona from a global collection based on a Natural Language Inference (NLI) model, the inferred persona is consistent with the predefined personas; and (2) ***Posterior-scored Transformer (PS-Transformer)***, it adopts a persona posterior distribution that further considers the actual personas used in the ground response, maximally mitigating the gap between training and inferring. Furthermore, we present a dataset called IT-ConvAI2 that first highlights the OOP problem in personalized dialogue. Extensive experiments on both IT-ConvAI2 and ConvAI2 demonstrate that our proposed model yields considerable improvements in both automatic metrics and human evaluations. All the data and codes are publicly available at https://github.com/CCIIPLab/Persona_Extend/.

*Corresponding author.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA.

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557359

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; *Natural language processing*; *Discourse, dialogue and pragmatics.*

KEYWORDS

dialogue generation, personality consistency, persona expanding

ACM Reference Format:
Yifan Liu, Wei Wei, Jiayi Liu, Xianling Mao, Rui Fang, and Dangyang Chen. 2022. Improving Personality Consistency in Conversation by Persona Extending. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557359

1 INTRODUCTION

Developing a more human-like dialogue system has been an important topic in artificial intelligence, where one of the major challenges is to maintain a consistent persona [15, 19, 21, 25, 28–30, 38]. Key-value lists are first used to construct structured profile explicitly, including name, gender, age, location, etc [25, 46, 47]. More recently, Zhang et al. [42] define the personality as several textual persona sentences as stated in Figure 1. As the unstructured personas are natural, vivid, and facilitate the description of complicated personalities, it sparks a wide range of interest in developing generators of personality-consistent responses [6]. To enhance the understanding of predefined textual personas: Wolf et al. [38] first employ pretrained model that leveraged the general dialogue corpus to understand textual personas better. Song et al. [28] pretrain an encoder with non-dialogue inference data to strengthen consistency understanding. Xu et al. [39] propose enriching predefined persona by searching related topics, and Majumder et al. [20] generalize predefined personas by leveraging commonsense to guess the underlying personas.

However, there are several limitations for existing methods on generating responses based on textual persona sentences. **First**,

Persona
P1: I have two cats.
P2: I live in a cabin by the lake.
P3: My favorite color is blue.
P4: I am an avid kayaker.
Conversations
Q1: Do you like animals?
R1: I have **two kitties**!
Q2: Wow, what is your family like?
R2: I am **the grandfather of four children**.
Q3: How about your family?
R3: Oh god, I might **be single** all my life.
Q4: Well, what do you usually do on weekends?
R4: I am a **cat lover**, and I take care of my **two cats**.

Figure 1: Examples generated by *Transformer* to illustrate following problems: (1) Response Conflict: R2 & R3 fabricate different personas, which are conflict with each other. (2) Inappropriate Persona: P4 is more appropriate for answering Q4 than P1.

most current methods adopt only the predefined personas for response genreation, and thus easily fail in generating the reasonable response if confronting the OOP problem. As shown in Figure 1, Q2 and Q3 are two classical OOP examples, which cannot directly answer the query like "farmily sistutation" with the given personas. However, without more external knowledge, the agent may fabricate several inappropriate personas (e.g., "have four children") that may be inconsistent with the prior persons, which is so-called personality inconsistent generation problem. **Second**, although there exists several works on expanding predefined personas for generation, they merely focus on paraphrasing a specific predefined, without considering the consistency of the expanded persona with the given query and other predefined personas [20, 39]. It may extend a persona that is not suitable for response, even inconsistent with the rest of personas, and lead to contradiction problems. **Third**, some methods [46, 47] simply fuse all personas into the generation process cursorily, which may lead to the output of an inappropriate response with an inconsistent persona. As shown in Figure 1, such as "I am kayaker" may be more relevant to Q4, however the agent still graft "I have two cats", as "I have pets" is a more general persona in the whole dataset, as compared to "I am kayaker", which is so-called long-tail bias problem. Under such circumstance, it is non-trivial to directly solve the OOP problem.

In this paper, we argue the importance of addressing the OOP problem, which may significantly improve the consistency of existing personalized dialogue systems. Recall the examples shown in Figure 1, for the OOP queries (e.g., Q2 and Q3), an reasonable solution is to obtain an appropriate persona from an external knowledge based on the per-defined personas. However, the generator may overlook the appropriate persona we expand (e.g., R4), so we must filter the existing textual personas before generating responses. Therefore, we design a pipeline that retrieves persona and selects persona for addressing the OOP problem. Inspired by this, our research starts by asking: **What is the principle of retrieving persona for OOP query?** Here, the first important issue is whether the retrieved personas are semantically consistent with the predefined ones. For example, "I don't like pets" obviously implies "I don't have a dog or a cat". Therefore, the retrieved personas should be compared with the predefined ones for semantic conflict checking. The second question is: **How to ensure that endowing the chatbot (e.g., the generated response) with the retrieved persona?** Generally, the existing generative models trend to select commonly appeared personas for generation. With the target of avoiding the general response generation, the generation model cannot use all of retrieved personas. Instead, we encourage the model to select the most query-relevant persona before generation, significantly improving the relevance of the generated response to the context.

Therefore, this paper proposes a novel retrieval-to-prediction pipeline consisting of *PRM* and *PS-Transformer*. Specifically, *PRM* is designed as a ranking module that extends personas by retrieving from a global persona set[1]. In particular, we leverage Natural Language Inference (NLI) to select personas that do not conflict with predefined personas. *PS-Transformer* adopts *Target-Guided Persona Scorer* to predict the availabilities of each persona to the query by posterior information. Incorporated with such a persona distribution, our proposed model is able to select the most suitable persona to generate responses. We build a challenging set named Inadequate-Tiny-ConvAI2 (IT-ConvAI2) by removing those query-related personas from the original ConvAI2 dataset. In this way, we verify that the *PRM* could steadily extend a suitable new persona to tackle the OOP problem and facilitate *PS-Transformer* to generate personality-consistent responses. On both IT-ConvAI2 and ConvAI2, we demonstrate that our method directly improves the coherence of generation at the personality level.

The main contributions of this research are summarized:

First, we propose a novel framework solving the OOP problem in dialogue generation. This framework involves two processes, i.e., conflict-detecting persona retrieving and dialogue generation with selected personas.

Second, we are the first to leverage NLI to estimate the coherence from persona candidates to predefined personas. Extensive experiments demonstrate that our proposed *PRM* can gather better personas than others.

Third, we propose a novel *PS-Transformer* introducing the *Target-Guided Persona Scorer* to predict persona distributions instead of fusing them roughly. The *PS-Transformer* yields the best results on both IT-ConvAI2 and ConvAI2.

2 RELATED WORK

2.1 Personalized Dialogue

Although neural response generation models have achieved promising results [13, 18, 27, 34, 35, 43, 45], they are still unsatisfactory. Previous work [5] investigated that topic changing will significantly satisfy conversational participants. Furthermore, Mitsuda et al. [22] proposed that 78.5% of the perceived information during chit-chat is directly related to personal information. Li et al. [15]

[1]In this paper we simply take all personas from the test set of ConvAI2 as our global personas.

first proposed a personalized dialogue system to introduce personal information into dialogue generation. After this, Qian et al. [25] proposed WD Profile Dataset, and Zhang et al. [42] proposed ConvAI2. Such personalized dialogue contributed to the development of both retrieval-based and generative-based personalized dialogue models.

In the line of retrieval-based methods [10, 11, 42], Gu et al. [11] found it is helpful to utilize personas in response selection. Although our proposed *PRM* retrieves personas, our pipeline method does not belong to the retrieval-based methods. Because our method does not directly take the retrieval results as responses, but uses them as the basis for generating, which facilitates the generation of informative and consistent responses.

In the line of generative-based methods, Li et al. [15] first took user embedding as an implicit persona in multi-turn dialogues. However, it relied on expensive speak-tagged dialogue data. Recent works incorporated explicit persona into the generation in two ways: (1) Qian et al. [25] and Zheng et al. [47] defined personality as structured key-value profiles consisting of some basic personal information such as name, age, and location. (2) Zhang et al. [42] contributed a chat-oriented dataset, taking personality as a predefined collection of textually described persona sentences. Most of the persona dialogue methods[28–30, 38, 41, 42] focused on how to understand personas better in the latter high-quality corpus. Specifically, Zhang et al. [42] employed basic Seq2Seq splicing personas with the query without distinguishing them. Wolf et al. [38] first introduced transfer learning by fine-tuning pretrained model to improve the quality of generation. However, all methods above take the agent's personality as a predefined closed set. Once the query goes beyond predefined personas (OOP problem), the agent tends to fabricate a new persona, resulting in a risk of inconsistent personality. To tackle the problem, we propose our retrieval-to-prediction pipeline that extends persona before generation.

2.2 Natural Language Inference

The task of Natural Language Inference (NLI) is to learn a function $f_{\text{NLI}}(p, h) = \{\text{E}, \text{N}, \text{C}\}$, where p and h denote premise and hypothesis respectively. The outputs E, N and C represent the conjunction, neural and contradiction relations between premises and hypotheses. Since the release of the large-scale corpus SNLI [1], deep neural network approaches have made promising progress[2, 8, 14]. Welleck et al. [36] modeled the detection of conversational consistency as an NLI task and proposed the Dialogue NLI dataset. And Song et al. [30] adopted the RL framework to leverage NLI knowledge as a reward. Song et al. [28] further pretrained on NLI task to ensure generating responses that entail predefined personas.

Motivated by this, we argue that NLI is crucial for personal retrieval to identify the relevances between persona candidates and predefined personas. So we consider the entail and conflict with the predefined personas in the NLI perspective when *PRM* retrieves the persona, thus providing suitable persona for the generative model.

2.3 Knowledge Enhanced Dialogue

The incorporation of knowledge has been shown to be an effective way to improve the performance of dialogue generation. There is a trend to leverage many domain-specific knowledge bases to ground

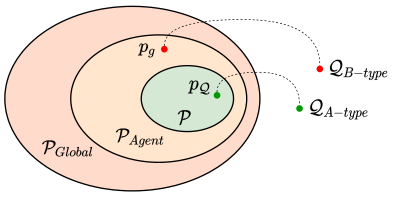

Figure 2: Relations between personas and queries.

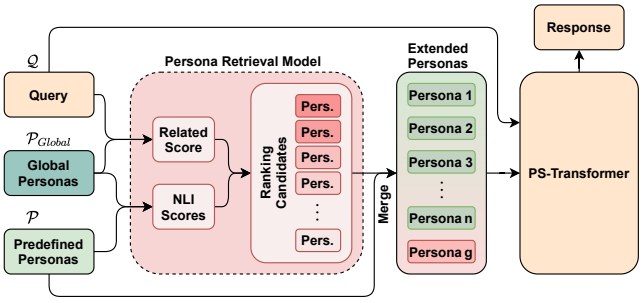

Figure 3: Overview of the retrieval-to-prediction paradigm.

neural models [9, 23, 40, 44, 48, 49], in which the textual persona sentences are one of the most frequently considered knowledge [16]. Recently, Lian et al. [16] propose that compared with the knowledge posterior distribution that further considers the actual knowledge used in real responses, the prior distribution has a large variance, and therefore, it is difficult for existing models to simply select the appropriate knowledge based on the prior distribution during training. On this basis, Song et al. [29] and Gu et al. [12] use posterior distributions effectively to ensures that knowledge is better utilized in generating responses.

We borrow the idea that leverages posterior distribution to select the appropriate knowledge with several differences in motivation and methodology: (1) We use the posterior distribution to select the actual personas rather than traditional knowledge in the grounded response. (2) Compared to fusing all personas into one representation [12], we consider the modeling of persona selection distribution.

3 METHODOLOGY

3.1 Task Definition

In this paper, our personalized dialogue generation problem aims at endowing a dialogue system with a consistent personality for building a human-like conversation sysmte, which can be formally defined as follows, given a query $\mathcal{Q} = \{q_i\}_{i=1}^{m}$ and a set of predefined personas $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$, where each persona depicted with a sentence $p_i = \{w_j\}_{j=1}^{m} (i \in \{1, 2, \ldots, n\})$, the task aims to generate a response $\mathcal{R} = \{r_i\}_{i=1}^{m}$ coherent to both the query and agent's personas.

As stated in Figure 2, assuming a global persona collection $\mathcal{P}_{Global}$ for all agents, personas belonging to a specific agent could be declared as $\mathcal{P}_{Agent}$ ($\mathcal{P}_{Agent} \subsetneqq \mathcal{P}_{Global}$). Usually, we handly predefine a persona set $\mathcal{P}$ ($\mathcal{P} \subsetneqq \mathcal{P}_{Agent}$) for the agent. On this assumption, we divide queries into A-type and B-type. A-type queries

can be answered based on predefined $p_{\mathcal{Q}}$ ($p_{\mathcal{Q}} \in \mathcal{P}$), while B-type queries need us to detect a new persona p_g ($p_g \in \mathcal{P}_{Agent},\ p_g \notin \mathcal{P}$) to tackle.

To simplify the problem, we assume that global persona set P_{Global} must contain at least one persona related to the query. So this paper focuses on retrieving a suitable persona p_g from P_{Global} and generating responses coherent with extended personas. The personalized dialogue generation can be briefly stated below:

$$\begin{aligned}&\mathbf{P}(\mathcal{R}|\mathcal{Q},\mathcal{P},\mathcal{P}_{Global})\\ =&\mathbf{P}(\mathcal{R}|\mathcal{Q},\mathcal{P}\cup\{p_g\})\cdot\mathbf{P}(p_g|\mathcal{Q},\mathcal{P},\mathcal{P}_{Global}),\end{aligned} \tag{1}$$

where $\mathbf{P}(p_g|\mathcal{Q},\mathcal{P},\mathcal{P}_{Global})$ denotes detecting a new persona p_g from $\mathcal{P}_{Global}$ considering current query $\mathcal{Q}$ and predefined personas $\mathcal{P}$. And $\mathbf{P}(\mathcal{R}|\mathcal{Q},\mathcal{P}\cup\{p_g\}) = \prod_{t=1}^{n_r}\mathbf{P}(r_t|\mathcal{Q},\mathcal{P}\cup\{p_g\},r_{<t})$ represents the response generation based on both the context query $\mathcal{Q}$ and extended personas $\mathcal{P}\cup\{p_g\}$.

3.2 Overview

As stated in Figure 3, we designed a retrieval-to-prediction pipeline that combines persona extending and response generation. The pipeline consists of two stages: *PRM* retrieves a persona from the global persona set based on the predefined personas and the context query. Then *PS-Transformer* generates a response with the query and the extended personas. The details of our method will be explained in Section 3.3 and Section 3.4, respectively.

3.3 Persona Retrieval Model

The *Persona Retrieval Model (PRM)* is responsible for addressing $\boldsymbol{P}(p_g|\mathcal{Q},\mathcal{P},\mathcal{P}_{Global})$, i.e. ranking all the candidate personas and picking the mostly one for the agent. Firstly we handly prepare a collection of persona candidates $\mathcal{P}_{Global}$[2] from the ConvAI2 dataset. As stated in Figure 3, the *PRM* ranks all the candidates based on both its relanvences to query and predefined personas.

We employ a sentence-pair matching model to estimate the logical association between query and persona candidate. The score predicted by the binary classification model is the query-persona relevance, as stated in Equation 2.

$$r = \text{Related}(\mathcal{Q}, p_g) \tag{2}$$

Song et al. [30] finds that NLI models could be used to calculate the coherence between response and query. Inspired by their work, We empirically adopt a standard pretrained NLI model [7] to check textual entailment and conflict between the persona candidate and predefined personas. We apply the maximum algorithm to encourage the persona candidate p_g closing to predefined personas $\mathcal{P}$, as shown in Equation 3.

$$\begin{aligned}e &= \text{Entail}(\mathcal{P}, p_g)\\ &= \max_{p_i\in\mathcal{P}}\{\text{Entail}(p_i, p_g)\}\end{aligned} \tag{3}$$

where $\text{Entail}(\cdot,\cdot)$ is the entailment score predicted by our model. The persona candidate p_g should be punished if it conflicts with any predefined persona of $\mathcal{P}$, so we also apply the maximum to calculate the conflict score for persona candidate p_g in Equation 4.

$$\begin{aligned}c &= \text{Conflict}(\mathcal{P}, p_g)\\ &= \max_{p_i\in\mathcal{P}}\{\text{Conflict}(p_i, p_g)\}\end{aligned} \tag{4}$$

where $\text{Conflict}(\cdot,\cdot)$ is the conflict score given by our model.

We propose two approaches to combine scores r, e, c as our ranking methods:

(1) **Heuristic Rules** (NLI_{HR}): We first retrieve top-10 candidates from $\mathcal{P}_{Global}$ with the highest r score, and these persona candidates should be most relevant to the query $\mathcal{Q}$. Then, we take the persona with the highest e score from the top-3 lowest c scores to encourage both its low conflict and strong entailment to predefined personas.
(2) **Weight Combination** (NLI_{WC}): We adopt three regulator α, β, and γ to construct a combined score $S = \alpha\cdot r+\beta\cdot(1-c)+\gamma\cdot e$. Then we sort the candidates with S scores in descending order and take the first one as a result. In this paper we set $\alpha = 0.75, \beta = 0.25, \gamma = 0.10$.

3.4 Posterior-scored Transformer

The dialogue generator we proposed is a transformer-based model stated in Figure 4. Following the champion model in the ConvAI2 competition [3], we adopt OpenAI GPT [26] as our weight-shared encoder $\text{Encoder}_{\text{GPT}}(\cdot)$ and decoder $\text{Decoder}_{\text{GPT}}(\cdot)$.

3.4.1 ***Target-Guided Persona Scorer.*** Let $\mathcal{Q}, p_i, \mathcal{G}$ denote query, i^{th}-persona, and ground truth (also known as target response), respectively. As stated in Equation 5, we first adopt $\text{Encoder}_{\text{GPT}}(\cdot)$ to turn the token-level embeddings into fixed-length representations at timestamp t.

$$\begin{aligned}\mathbf{E}_{\mathcal{Q}}, \mathbf{H}^t_{\mathcal{Q}} &= \text{Encoder}_{\text{GPT}}(\mathcal{Q})\\ \mathbf{E}_{\mathcal{G}}, \mathbf{H}^t_{\mathcal{G}} &= \text{Encoder}_{\text{GPT}}(\mathcal{G})\\ \mathbf{E}_{p_i}, \mathbf{H}^t_{p_i} &= \text{Encoder}_{\text{GPT}}(p_i), i = 1, 2, ..., n_p\end{aligned} \tag{5}$$

where $\mathbf{E}_*$ represents time-independent sentence embeddings of each input after self-attention only. $\mathbf{H}^t_*$ denotes the hidden states of each input after interacting with generated $\mathcal{G}_{<t}$.

The multi-head self-attention (denoted as MHA, [32]) is used to compute the importance from i^{th}-persona to either query $\mathcal{Q}$ or the ground truth $\mathcal{G}$. For each persona p_i we calculate the attention $\mathbf{A}^*_i$ in Equation 6.

$$\begin{aligned}\mathbf{A}^{pri}_i &= \text{MHA}_{pri}(\mathbf{Q}=\mathbf{E}_{p_i}, \mathbf{K}=\mathbf{E}_{\mathcal{Q}}, \mathbf{V}=\mathbf{E}_{\mathcal{Q}})\\ \mathbf{A}^{post}_i &= \text{MHA}_{post}(\mathbf{Q}=\mathbf{E}_{p_i}, \mathbf{K}=\mathbf{E}_{\mathcal{G}}, \mathbf{V}=\mathbf{E}_{\mathcal{G}})\end{aligned} \tag{6}$$

Since attention $\mathbf{A}^*_i$ denotes the importance of each persona to the response. A two-layer multilayer feedforward perceptron (MLP) with a sigmoid activation is used to turn them into a comprehensed weight as stated in Equation 7.

$$w^*_i = \sigma(\text{MLP}(\mathbf{A}^*_i)),\ (* = post \text{ or } pri) \tag{7}$$

The binary cross entropy (BCE) loss is adopted to optimize the capture of weight w^{post}_i:

$$\mathcal{L}_1 = -[w_i \log w^{post}_i + (1-w_i)\log(1-w^{post}_i)] \tag{8}$$

[2]To avoid the label leaking, we make sure that all candidates did not be used for training our sentence-pair model and NLI model.

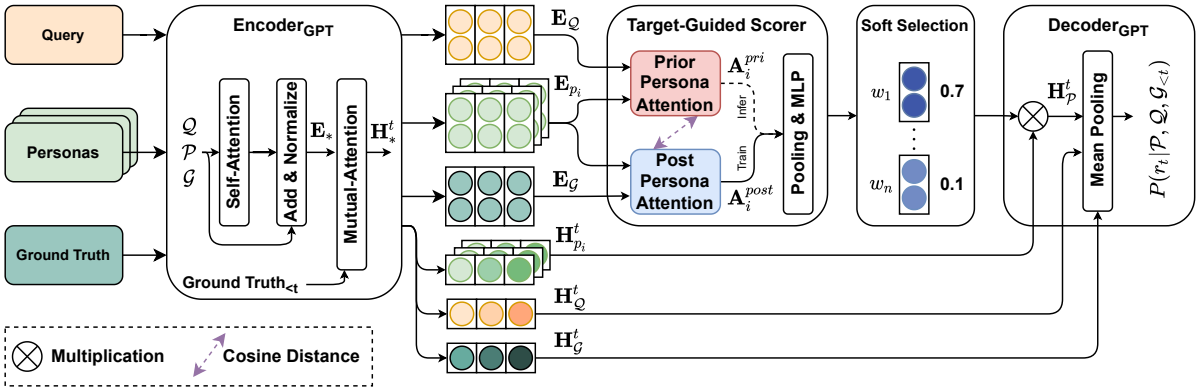

Figure 4: The neural network architecture of the proposed *PS-Transformer*.

Besides, the cosine embedding loss is used to gain both attentions from prior and posterior network as stated in Equation 9.

$$\mathcal{L}_2 = 1 - \cos(\mathbf{A}_i^{post}, \mathbf{A}_i^{pri}) \tag{9}$$

3.4.2 ***Decoder for Weighted-sum Attentions.*** Firstly, the representation $\mathbf{H}_{\mathcal{P}}$ for the predefined persona set $\mathcal{P}$ could be incorporated from $\mathbf{H}_{p_i}$ in Equation 5 based on w_i^{post} given by Equation 7.

$$\mathbf{H}_{\mathcal{P}}^t = \sum_{i=1}^{n_p} w_i^{post} \cdot \mathbf{H}_{p_i}^t \tag{10}$$

To give consideration to both query and the past generated words, In each timestamp t of decoding, representations of query, personas and past generated words are treated equally. The prediction of word r_t is stated in Equation 11.

$$\begin{aligned} \mathbf{H}_{dec}^t &= \text{mean}(\mathbf{H}_{\mathcal{Q}}^t, \mathbf{H}_{\mathcal{P}}^t, \mathbf{H}_{\mathcal{G}}^t) \\ r_t &= \text{Decoder}_{\text{GPT}}(\mathbf{H}_{dec}^t) \end{aligned} \tag{11}$$

where mean$(\cdot)$ denotes averaging given matrices by element.

In essence, the *PS-Transformer* read the persona set $\mathcal{P}$ and the query $\mathcal{Q}$ to predict the target response $\mathcal{G}$. So we apply the negative log-lokeihood loss during training.

$$\begin{aligned} \mathcal{L}_3 &= -\log\left(\boldsymbol{P}(\mathcal{G}|\mathcal{P}, \mathcal{Q})\right) \\ &= -\sum_{t=1}^{|\mathcal{G}|} \log(\boldsymbol{P}(r_t|\mathcal{P}, \mathcal{Q}, \mathcal{G}_{<t})) \end{aligned} \tag{12}$$

3.4.3 ***Inferrence Network.*** Similar to Equation 10, the predefined personas are soft-selected by weighted summation based on the w_i^{pri} predicted in Equation 7:

$$\mathbf{H}_{\mathcal{P}}^t = \sum_{i=1}^{n_p} w_i^{pri} \cdot \mathbf{H}_{p_i}^t \tag{13}$$

During decoding, the response is generated in a self-recursion way as stated in Equation 14.

$$\begin{aligned} \mathbf{H}_{dec}^t &= \text{mean}(\mathbf{H}_{\mathcal{Q}}^t, \mathbf{H}_{\mathcal{P}}^t, \mathbf{H}_{\mathcal{R}}^t) \\ r_t &= \text{Decoder}_{\text{GPT}}(\mathbf{H}_{dec}^t) \end{aligned} \tag{14}$$

4 EXPERIMENTAL SETUP

4.1 Research Questions

To fully demonstrate the superiority of our method, we conduct experiments to verify the following six research questions (RQ):

- (**RQ1**): Can our proposed pipeline, consisting of *PRM* and *PS-Transformer*, yield good results on automatic metrics in response to OOP queries? (See Section 5.1)
- (**RQ2**): Can our proposed *PRM* actually solve the OOP problem to some extent? Will the quality of the response generated by *PS-Transformer* be better if we solved the OOP problem better? (See Section 5.2)
- (**RQ3**): Can our proposed *PS-Transformer* more accurately select the personality used to generate the response, compared to other baselines? (See Section 5.3)
- (**RQ4**): What is the impact of the key components in the *PS-Transformer* on performance? (See Section 5.4)
- (**RQ5**): Can *PS-Transformer*'s performance on IT-ConvAI2 be generalized to the original ConvAI2? (See Section 5.5)
- (**RQ6**): How does our response method differ from baselines? (See Section 5.6)

4.2 Datasets

ConvAI2[3] It is published for the second Conversational Intelligence Challenge [3], and both speakers of each conversation consist of at least five persona descriptions. The dataset contains 17,878/1,000 multi-turn dialogues conditioned on 1,155/100 personas for train/test.

Inadequate-Tiny-ConvAI2 (IT-ConvAI2) Since ConvAI2 encourages conversation participants to exchange their persona information, speakers tend to express their personas actively without being asked, resulting in fewer OOP queries than in actual practice. To obtain a realistic evaluation of persona-missing conversation, we build IT-ConvAI2 in two steps: (1) We first extract queries asking for persona and responses related to personas, respectively. If the extracted query and response are present in a conversation

[3]ConvAI2 is available at https://github.com/facebookresearch/ParlAI/tree/master/parlai/tasks/convai2.

Table 1: Automatic evaluation on IT-ConvAI2. In this evaluation, we adopt NLI_{WC} in Section 3.3 as the *PRM*.

Model	Pretrained	IT-ConvAI2				IT-ConvAI2 with *PRM*			
		Consist	Quality			Consist	Quality		
		Entail	BLEU	ROUGE	CIDEr	Entail	BLEU	ROUGE	CIDEr
Seq2Seq	✗	0.115	5.62	1.71	8.77	0.178	5.69	1.71	9.06
PerCVAE	✗	0.306	2.26	0.93	4.46	0.380	2.27	0.96	4.22
DialogWAE	✗	0.077	4.13	1.12	5.81	0.103	3.84	1.09	5.27
Transformer	✓	0.539	6.21	1.55	10.56	0.495	6.17	1.52	11.11
TransferTransfo	✓	0.546	5.12	1.34	13.23	0.645	5.18	1.36	12.85
BoB	✓	0.505	5.39	1.43	11.39	0.628	5.35	1.40	10.74
PS-Transformer	✓	**0.560**	**7.12**	**1.71**	**14.43**	**0.670**	**7.35**	**1.73**	**15.88**

triad $(query, response, personas)$, we will collect them into Tiny-ConvAI2. (2) To build IT-ConvAI2, for each conversation in Tiny-ConvAI2, those personas involved in response will be removed. As a result, we manually collect 1,595 conversations as IT-ConvAI2.

4.3 Baseline Methods

We compared our proposed approach with the following strong models:

- Generative Based: **Seq2Seq** [42] is a traditional LSTM-based encoder-decoder model prepending all personas to the query. **PerCVAE** [29] further incorporates personas with contexts by a memory network. **DialogWAE** [12] contains a conditional Wasserstein Auto-Encoder, and we adapt it to personalized dialogue generation by concatenating personas with the query directly.
- Pre-training & Fine-tuning Based: **Transformer** [3] achieves state-of-the-art performance in the manual metrics of the ConvAI2 competition while **TransferTransfo** [38] tops automatic evaluations. **BERT-Over-BERT (BoB)** [28] contains two decoders pretrained on NLI task. It is good at generating responses entailed with personas.

4.4 Evaluation Metrics

4.4.1 Automatic Evaluation. To highlight the quality of generation on both personality and contextual aspects, we evaluate each response with two aspects:

- **Consistency**: Following Dziri et al. [4], we employ ESIM[4] [2] to automatically evaluate the **entailment score** between the generated response $\mathcal{R}$ and the agent's personas $\mathcal{P} = \{p_1, p_2, ..., p_n\}$:

$$\begin{aligned} e' &= \text{Entail}'(\mathcal{P}, \mathcal{R}) \\ &= \max_{p_i \in \mathcal{P}} \{\text{Entail}'(p_i, \mathcal{R})\} \end{aligned} \tag{15}$$

- **Quality**: **BLEU** [24] and **ROUGE** [17] are used to measure the relevance between the ground truth and generated response. We also employ **CIDEr** [33] to capture the overlap of persona information between the machine response and human reference.

[4]We use an NLI model different from the one in *PRM* for a fair evaluation.

4.4.2 Human Evaluation for PRM. Three masters students in the field of dialogue were asked to evaluate per *PRM* according to three metrics:

- **Query-relevance S_q^p (0-1)**: To indicate if the retrieved persona is related to the query based on 1/0 scoring schema.
- **Persona-entailment S_p^p (0-2)**: Scoring how the retrieved persona entails with the query. 0 means conflict, 1 means neutral and 2 means entailment.
- **DCG@3**: We collect the top three retrieved results for each method and calculate the DCG@3 in Equation 16.

$$\text{DCG}_3 = \sum_{i=1}^{3} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{16}$$

where $rel_i = S_p^p$ if the retrieved persona is related to the query, otherwise $rel_i = 0$.

4.4.3 Human Evaluation for PS-Transformer. Three judges are asked to evaluate Query-relevance S_q (1-3), Persona-entailment S_p (1-3) and Response-fluency S_r (1-3) of generated responses:

- For **Query-relevance** S_q, 1 point means that the response is irrelevant with the query. 2 point means that the response is relevant with query, but is the general response. 3 means that the response perfectly answers the query.
- **Persona-entailment** S_p measures whether the response is entailed with predefined personas. 1 means the response doesn't contain any persona. 2 means the response contains persona but not in predefined persona set. 3 means the response contains predefined persona.
- **Response-fluency** S_r is used to evaluate the syntactic and logical fluency of the response. The higher the score, the better the performance. 3 point means that the response is both grammatically and logically correct.

4.5 Implementation Details

- The sentence-pair classifier and the NLI scorer of *PRM* are both BERT-based models. We manually annotate one thousand related $(query, persona)$ pairs for training the sentence

pair classifier. NLI scorer is pretrained on both SNLI[5] and MultiNLI[6][37], then is finetuned on DNLI[7].
- To train the *Target-Guided Persona Scorer*, we follow Song et al. [29] labelling each response with its corresponding persona by inverse document frequency. The response has a tf-idf similarity with each persona, and we label each (*response*, *persona*) pair with 1/0 according to whether the similarity is higher than a threshold.
- We employ OpenAI's GPT [26] to initialize *Transformer*, *TransferTransfo* and our *PS-Transformer*.

5 RESULTS AND ANALYSIS

5.1 Automatic Evaluations (RQ1)

As shown in Table 1, we evaluate all the methods on *IT-ConvAI2*, and *IT-ConvAI2 with PRM*, respectively, and we have the following observations:

Our pipeline method has the best overall performance in response to OOP queries. *PS-Transformer* outperforms all baselines regardless of whether our proposed *PRM* is applied to baselines or not. In particular, for *PS-Transformer*, the personality of *PRM* retrieval brings a significant improvement to the entailment of the response. This result shows that in the view of the generative model, the *PRM* retrieval results are very suitable for responding to OOP queries.

The *Persona Retrieval Model (PRM)* helps almost all methods generate personality-consistent responses. All methods except *Transformer* reach a higher entailment score after being given a new persona by *PRM*. Thus, we can generalize a general conclusion that retrieving a suitable persona using our proposed *PRM* in response to an OOP query can help the vast majority of generators to produce a personality-consistent response. It also helps to reduce the risk of fabricating a random persona and generating personality-conflicting responses. In addition, only *PS-Transformer*, when combined with *PRM*, shows a significant improvement not only in entailment but also in response generation quality, which implies that *PS-Transformer* is better than baselines for selection and utilization of personality information.

5.2 Human Evaluations for *PRM*s (RQ2)

To demonstrate the significance of our retrieval methods in Section 3.3, we prepare some *PRM*s based on other retrieval methods for an ablation-like experiment: (1) **BM25** algorithm is used to retrieve the persona most similar to the query in lexical. (2) **Sentence-pair Classifier (Classify$_{SP}$)** only adopts the r score in Equation 2 to rank persona candidates. We employ judges to evaluate a random sample of 150 items per *PRM* according to metrics mentioned in Section 4.4. We further adopt *PS-Transformer* to generate responses based on extended personas from different *Persona Retrieval Model*s, and we apply automatic evaluations on those responses.

Our proposed *PRM* can effectively solve the OOP problem, and the NLI contributes to improving retrieval performance. As retrieval quality is shown in Table 2, 59% of the personas retrieved by our proposed NLI_{WC} correspond to OOP queries, and the vast majority of these personalities are non-conflicting with predefined personas. Also, NLI_{WC} outperforms other *PRM*s in terms of the Query-relevance S_q^p, the Persona-entailment S_p^p, and the overall ranking performance DCG@3. Specifically, NLI_{WC} significantly outperforms *BM25* in terms of S_q^p score, indicating that it is effective to consider the semantic relevance between query and retrieved persona. In addition, NLI_{WC} significantly outperforms $Classify_{SP}$ in terms of S_p^p score and DCG@3 score, which indicates that natural language inference can effectively reduce the possibility of conflict between retrieved persona and predefined personas, thus improving the final persona retrieval performance.

Table 2: The left shows human evaluation of *PRM*s. The Fleiss´ kappa values of S_q^p, S_p^p, DCG@3 are 0.62, 0.49, and 0.57, respectively, indicating *Substantial*, *Moderate*, and *Moderate agreement*. The maximum value of the S_q^p, S_p^p are 1, 2, respectively. The right shows automatic evaluations for generated responses based on different *PRM*s.

Model	Retrieval Quality			Response Quality		
	S_q^p	S_p^p	DCG@3	Entail	Conflict	BLEU
BM25	0.35	0.86	0.77	0.592	0.237	5.65
Classify$_{SP}$	0.56	0.87	1.01	0.650	0.304	7.16
NLI$_{HR}$	0.55	0.90	1.15	0.643	0.241	7.35
NLI$_{WC}$	**0.59**	**0.97**	**1.33**	**0.670**	**0.214**	**7.35**

The better the OOP Problem is solved, the higher the response quality of our proposed pipeline method. As response quality is shown in Table 2, *BM25* retrieves persona considering text similarity only, which makes the retrieved persona weakly correlated with query and even becomes noise during generation. Therefore, *BM25* produces less improvement in Entail score than other *PRM*s and reduces BLEU score that reflects generative performance. Since $Classify_{SP}$ ignores the relevance between retrieved persona and predefined personas, the retrieved persona may conflict with predefined personas. In such a case, no matter which persona the generative model selects, the response will conflict with the existing persona set, resulting in a higher Conflict score. Compared to $Classify_{SP}$, NLI-based *PRM*s (NLI_{HR} and NLI_{WC}) reduce the risk of personality conflicts by considering the NLI relevance from retrieval candidates to predefined personas, responses generated based on such *PRM*s also perform well with higher BLEU scores than other *PRM*s. The results show that NLI_{WC} outperforms all other *PRM*s in persona retrieval and is most helpful in improving the Entail score and reducing the Conflict score of generated responses.

5.3 Human Evaluations for Generative Models (RQ3)

We randomly sample 150 predictions from column *IT-ConvAI2 with PRM* in Table 1 and invite three graduate students for evaluation. The human evaluation results are shown in Table 3, and Figure 5 shows the detailed compositions of S_q and S_p scores.

The *PS-Transformer* outperforms baselines by generating persona-targeted responses. As the detailed composition of S_p

[5]The SNLI is available at https://nlp.stanford.edu/projects/snli.
[6]The MultiNLI is available at https://cims.nyu.edu/~sbowman/multinli.
[7]The DNLI is available at https://wellecks.github.io/dialogue_nli.

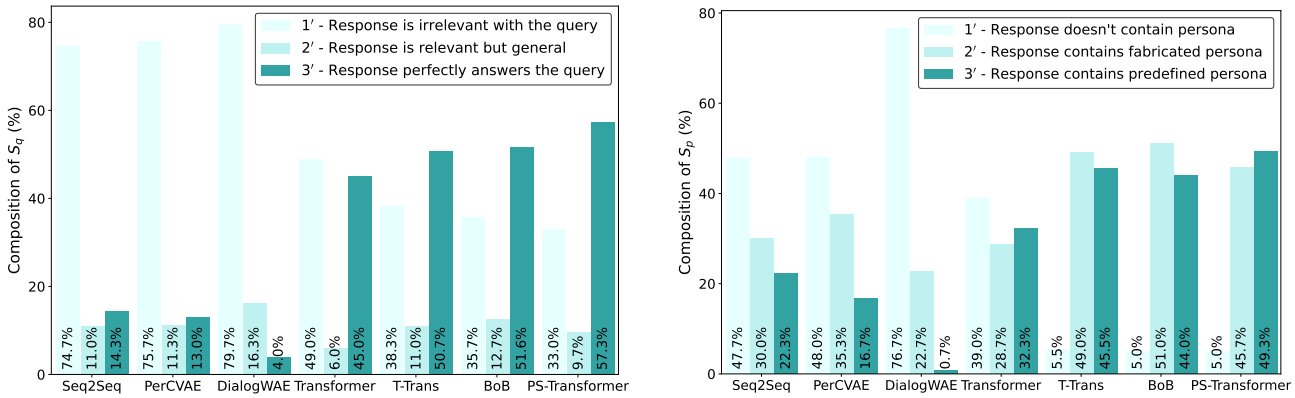

Figure 5: The composition of human evaluation for generated responses (Query-relevance S_q and Persona-entailment S_p). The scoring criteria are shown in the legend. T-Trans is shorthand for TransferTransfo.

Table 3: Human evaluation of all the generative methods. The Fleiss´kappa values of S_q, S_p, S_r are 0.56, 0.70, and 0.42, respectively, indicating *Moderate*, *Substantial* and *Moderate agreement*.

Model	S_q	S_p	S_r
Seq2Seq	1.40	1.75	2.59
PerCVAE	1.37	1.70	2.54
DialogWAE	1.25	1.24	2.32
Transformer	1.96	1.93	2.93
TransferTransfo	2.12	2.40	2.90
BoB	2.16	2.39	2.89
PS-Transformer	**2.24**	**2.44**	**2.95**

is shown in Figure 5, the generated results of *PS-Transformer* have a higher probability of containing predefined personas than all baselines. Compared to *TransferTransfo* and *BoB*, *PS-Transformer* has a lower probability of fabricating persona. This is because *PS-Transformer* determines which personas should be used before generation, so those selected personas are more likely to be reflected in the response. In addition, as the detailed composition of S_q is shown in Figure 5, responses generated by *PS-Transformer* are the most consistent with queries because the personas selected by *PS-Transformer* in advance are strongly correlated with queries. Thus the responses generated based on the selected personas strongly correlate with the context. Not only do the results demonstrate that the *Target-Guided Persona Scorer* plays a vital role in accurately selecting persona to generate context-coherence responses, but they are also consistent with the automatic evaluation result that *PS-Transformer* significantly outperforms other methods in both personality coherence and generating quality.

5.4 Ablation Study (RQ4)

As reported in Table 4, we designed and evaluated two variants of *PS-Transformer*: (1) We first remove the posterior network (Eq. 9) by directly training the model with prior attention $\mathbf{A}_i^{pri}$. It means

Table 4: Ablation study on *IT-ConvAI2 with PRM*.

Settings	Consist	Quality		
	Entail	BLEU	ROUGE	CIDEr
PS-Transformer	**0.670**	**7.35**	**1.73**	**15.88**
- w/o Posterior Network	0.660	6.71	1.64	14.67
- w/o Scorer	0.356	3.54	1.02	10.05

Table 5: Automatic evaluation on original ConvAI2.

Model	Consist	Quality		
	Entail	BLEU	ROUGE	CIDEr
Seq2Seq	0.092	5.12	1.43	9.41
PerCVAE	0.287	2.44	0.91	5.49
DialogWAE	0.047	3.71	1.07	5.68
Transformer	0.544	5.47	1.37	10.87
TransferTransfo	0.508	4.70	1.13	13.74
BoB	0.499	4.33	1.10	9.81
PS-Transformer	**0.546**	**6.58**	**1.49**	**14.57**

we drop the actual personas used in real responses modeled by posterior distribution. It results in deteriorated performance, indicating the importance of the guidance from posterior information. (2) We remove the entire scoring mechanism (Eq. 8) by treating all personas equally while generating. The significant decrements of all metrics indicate that considering the relevance of personas to query and accurately selecting personas plays an important role in generating personality-consistent and high-quality responses.

5.5 Effectiveness of *PS-Transformer* on ConvAI2 (RQ5)

As stated in Table 5, the performance of all the methods on ConvAI2 is consistent with those on IT-ConvAI2 (in Table 1). Compared to IT-ConvAI2, a large part of the conversations in ConvAI2 do not

Table 6: Case Study. Responses in red are consistent with agent's personas, and in blue are in conflict with them.

Query	Hello, I'm a art major in college, what do you do?	I am getting in spring, you married?	What grade are you in?
Predefined Personas	I enjoy poetry.	I was poor growing up.	I play with my friends on the playground.
	I am a huge star wars fan.	Cheeseburgers are my favorite food.	Mickey mouse is my favorite character.
	I played football for a division of a college.	I like watching war documentaries.	I love to go to disney world every year.
Extended Persona	I am studying to be a computer programmer.	I'm married.	I am in the third grade.
Seq2Seq	I am a scientist. I love to write poetry.	No, but i do like watching the flintstones. you?	I am in the first grade. how about you ?
PerCVAE	Who r you on there?	Oh no i live at a beach.	I love to sleep in.
DialogWAE	Do you want to go to the beach?	Wow I love to travel. i love it.	Do have any pets?
Transformer	I am a computer science major at hs.	I am getting married in 6 months.	I am in my last year of high school.
TransferTransfo	I am a programmer. What do you do?	Yes, I am married to my wife. How about you?	My grade is 5th grade and you?
BoB	I'm a computer programmer in college.	No, I am not married. I am a single mom.	I am in 10th grade. I love it.
PS-Transformer	I am a student, but I am hoping to be a computer science major.	Yes, I'm married to my wonderful husband.	I am in third grade.

even need to be responded to using personas, but our proposed *PS-Transformer* still outperforms all other baselines. The *Target-Guided Persona Scorer* not only selects personas related to the query, but also excludes irrelevant personas as noise, avoiding the deliberate use of personas when generating responses.

5.6 Case Study (RQ6)

In this section, we present an in-depth analysis of response generation of our proposed approach at the level of personality consistency. As shown in Table 6, we prepare three cases generated by different models to explain the superiority of our motivations in personalized dialogue generation.

For the first case: The results suggest that the response generated by our approach is more consistent with personas. For instance, the response "I am a student, but I am hoping to be a computer science major." preserves the persona "to be a programmer". At the same time, other methods focus on "programmer" only.

For the second case: The persona retrieved by *PRM* is proper for the query. The responses generated by *TransferTransfo* and *PS-Transformer* are coherent at both personality and semantic levels when other methods give wrong or irrelevant answers. It should be noted that althought we determine agent's personas as "married", it is still possible for agents to fabricate personas about "gender", which is a potential problem for further research.

For the third case: The persona retrieved by *PRM* is related to the query and strongly entails all the predefined personas. Though it is hard to exploit persona with numeric information such as "third grade" accurately, *PS-Transformer* still generates the response leveraging the proper persona when others give wrong answers.

5.7 Limitations

A major limitation of our proposed pipeline is that the global persona set used by *PRM* is constructed in advance, which would make the pipeline still unable to handle OOP queries outside of the entire global persona set. A potential solution is introducing a large-scale commonsense knowledge graph (e.g., ConceptNet [31]) to infer new personas, and the utilization of knowledge graphs leaves another research direction.

6 CONCLUSION

In this paper, we propose to tackle the OOP problem in personalized dialogue generation. To tackle the problem above, we formally define the persona extending task and demonstrate that Natural Language Inference can help *PRM* to retrieve a coherent persona for generating response. Besides, the *PS-Transformer* introduces a posterior network named *Target-Guided Persona Scorer* that help select persona accurately, which help generate personality-consistent responses. For future work, we will explore how the extended persona affects the next extension to generalize the *retrieval-to-prediction* paradigm over multi-turn conversations.

7 ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant No.61602197, Grant No.L1924068, Grant No.61772076, in part by CCF-AFSG Research Fund under Grant No.RF20210005, and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL).

REFERENCES

[1] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*. Association for Computational Linguistics, Lisbon, Portugal, 632–642.

[2] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *ACL*. Association for Computational Linguistics, Vancouver, Canada, 1657–1668.

[3] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020. The Second Conversational Intelligence Challenge (ConvAI2). In *NeurIPS*, Sergio Escalera and Ralf Herbrich (Eds.). Springer International Publishing, Cham, 187–208.

[4] Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating Coherence in Dialogue Systems using Entailment. In *NAACL*. Association for Computational Linguistics, Minneapolis, Minnesota, 3806–3812.

[5] Suzanne Eggins and Diana Slade. 2005. *Analysing casual conversation*. Equinox Publishing.

[6] Sarah Fillwock and David Traum. 2018. Identification of Personal Information Shared in Chat-Oriented Dialogue. In *LREC*. European Language Resources Association (ELRA), Miyazaki, Japan.

[7] Yang Gao, Nicolo Colombo, and Wei Wang. 2021. Adapting by Pruning: A Case Study on BERT. *arXiv:2105.03343* (2021).

[8] Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural Language Inference over Interaction Space. In *ICLR*.

[9] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In *ACL*. Association for Computational Linguistics, Berlin, Germany, 1631–1640.

[10] Jia-Chen Gu, Zhen-Hua Ling, Xiaodan Zhu, and Quan Liu. 2019. Dually Interactive Matching Network for Personalized Response Selection in Retrieval-Based Chatbots. In *EMNLP-IJCNLP*. Association for Computational Linguistics, Hong Kong, China, 1845–1854.

[11] Jia-Chen Gu, Hui Liu, Zhen-Hua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Partner Matters! An Empirical Study on Fusing Personas for Personalized Response Selection in Retrieval-Based Chatbots. In *SIGIR* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 565–574.

[12] Xiaodong Gu, Kyunghyun Cho, Jung Woo Ha, and Sunghun Kim. 2019. Dialogwae: Multimodal response generation with conditional Wasserstein auto-encoder. In *ICLR*.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] Seonhoon Kim, Inho Kang, and Nojun Kwak. 2019. Semantic Sentence Matching with Densely-Connected Recurrent and Co-Attentive Information. In *AAAI*, Vol. 33. 6586–6593.

[15] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. In *ACL*. Association for Computational Linguistics, Berlin, Germany, 994–1003.

[16] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to Select Knowledge for Response Generation in Dialog Systems. In *IJCAI*. International Joint Conferences on Artificial Intelligence Organization, 5081–5087.

[17] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81.

[18] Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, Yuchong Hu, and Shanshan Feng. 2021. Target-Guided Emotion-Aware Chat Machine. *ACM Trans. Inf. Syst.* 39, 4, Article 43 (aug 2021), 24 pages.

[19] Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You Impress Me: Dialogue Generation via Mutual Persona Perception. In *ACL*. Association for Computational Linguistics, Online, 1417–1427.

[20] Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? You probably enjoy nature: Persona-grounded Dialog with Commonsense Expansions. In *EMNLP*. Association for Computational Linguistics, Online, 9194–9206.

[21] Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training Millions of Personalized Dialogue Agents. In *EMNLP*. Association for Computational Linguistics, Brussels, Belgium, 2775–2779.

[22] Koh Mitsuda, Ryuichiro Higashinaka, and Yoshihiro Matsuo. 2019. What Information Should a Dialogue System Understand?: Collection and Analysis of Perceived Information in Chat-Oriented Dialogue. In *IWSDS*, Maxine Eskenazi, Laurence Devillers, and Joseph Mariani (Eds.). Springer International Publishing, Cham, 27–36.

[23] Weiran Pan, Wei Wei, and Xian-Ling Mao. 2021. Context-aware Entity Typing in Knowledge Graphs. In *Findings of EMNLP*. 1–8.

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318.

[25] 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*. International Joint Conferences on Artificial Intelligence Organization, 4279–4285.

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[27] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *IJCNLP*. Association for Computational Linguistics, Beijing, China, 1577–1586.

[28] Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data. In *IJCNLP*. Association for Computational Linguistics, Online, 167–177.

[29] Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting Persona Information for Diverse Generation of Conversational Responses. In *IJCAI*. 5190–5196.

[30] Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. In *AAAI*, Vol. 34. 8878–8885.

[31] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI* (San Francisco, California, USA). 4444–4451.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[33] R. Vedantam, C. Zitnick, and D. Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*. IEEE Computer Society, Los Alamitos, CA, USA, 4566–4575.

[34] Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *arXiv e-prints* (2015), arXiv–1506.

[35] Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 2019. Emotion-Aware Chat Machine: Automatic Emotional Response Generation for Human-like Emotional Interaction. In *CIKM*. Association for Computing Machinery, New York, NY, USA, 1401–1410.

[36] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue Natural Language Inference. In *ACL*. Association for Computational Linguistics, Florence, Italy, 3731–3741.

[37] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL*. Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122.

[38] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv:1901.08149* (2019).

[39] Minghong Xu, Piji Li, Haoran Yang, Pengjie Ren, Zhaochun Ren, Zhumin Chen, and Jun Ma. 2020. A Neural Topical Expansion Framework for Unstructured Persona-Oriented Dialogue Generation. (2020), 2244–2251.

[40] Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2017. Incorporating loose-structured knowledge into conversation modeling via recall-gate LSTM. In *IJCNN*. 3506–3513.

[41] Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tur. 2019. DeepCopy: Grounded Response Generation with Hierarchical Pointer Networks. In *SIGDIAL*. Association for Computational Linguistics, Stockholm, Sweden, 122–132.

[42] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *ACL*. Association for Computational Linguistics, Melbourne, Australia, 2204–2213.

[43] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *ACL*. Association for Computational Linguistics, Online, 270–278.

[44] Sen Zhao, Wei Wei, Zou Ding, and Xian-Ling Mao. 2022. Multi-view Intent Disentangle Graph Networks for Bundle Recommendation. In *AAAI*. 1–7.

[45] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *ACL*. Association for Computational Linguistics, Vancouver, Canada, 654–664.

[46] Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv:1901.09672* (2019).

[47] Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *AAAI*, Vol. 34. 9693–9700.

[48] Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible end-to-end dialogue system for knowledge grounded conversation. *arXiv:1709.04264* (2017).

[49] Ding Zou, Wei Wei, Xian-Ling Mao, Ziyang Wang, Minghui Qiu, Feida Zhu, and Xin Cao. 2022. Multi-level Cross-view Contrastive Learning for Knowledge-aware Recommender System. In *SIGIR*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 1358–1368.