You Impress Me: Dialogue Generation via Mutual Persona Perception

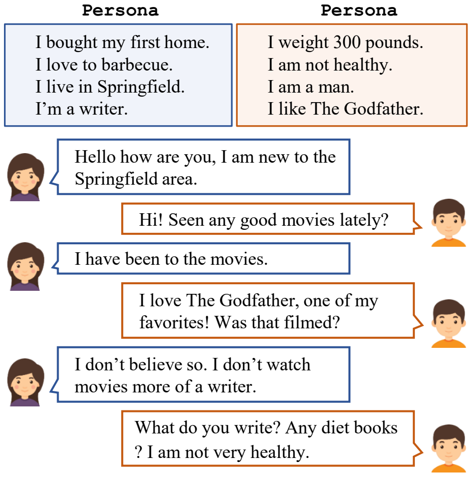
Qian Liu[†*], Yihong Chen[◊*], Bei Chen[§], Jian-Guang Lou[§],
Zixuan Chen[♠*], Bin Zhou[†], Dongmei Zhang[§]
[†]School of Computer Science and Engineering, Beihang University, China
[◊]UCL Centre for Artificial Intelligence, University College London, United Kindom
[♠]School of Computer Science, Fudan University, China
[§]Microsoft Research, Beijing, China
[†]{qian.liu, zhoubin}@buaa.edu.cn; [§]{beichen, jlou, dongmeiz}@microsoft.com;
[◊]yihong.chen@cs.ucl.ac.uk; [♠]remch183@outlook.com

Abstract

Despite the continuing efforts to improve the engagingness and consistency of chit-chat dialogue systems, the majority of current work simply focus on mimicking human-like responses, leaving understudied the aspects of modeling understanding between interlocutors. The research in cognitive science, instead, suggests that understanding is an essential signal for a high-quality chit-chat conversation. Motivated by this, we propose $\mathcal{P}^2$ Bot, a transmitter-receiver based framework with the aim of explicitly modeling understanding. Specifically, $\mathcal{P}^2$ Bot incorporates mutual persona perception to enhance the quality of personalized dialogue generation. Experiments on a large public dataset, Persona-Chat, demonstrate the effectiveness of our approach, with a considerable boost over the state-of-the-art baselines across both automatic metrics and human evaluations.

1 Introduction

Thanks to the advance in neural models and the accessibility of massive datasets, open-domain dialogue (i.e. chit-chat) systems have made great progress towards mimicking human-like responses. Nevertheless, there still exist some serious challenges in building personalized chatbots that can deliver engaging conversations and gain user trust (Song et al., 2019). For example, current chit-chat systems tend to generate uninformative responses (Li et al., 2016b). Moreover, they are usually lack of coherent personality traits due to the fact that training dialogues actually come from a diverse set of speakers (Zhang et al., 2018b).



I bought my first home.
I love to barbecue.
I live in Springfield.
I'm a writer.

I weight 300 pounds.
I am not healthy.
I am a man.
I like The Godfather.

Hello how are you, I am new to the Springfield area.

Hi! Seen any good movies lately?

I have been to the movies.

I love The Godfather, one of my favorites! Was that filmed?

I don't believe so. I don't watch movies more of a writer.

What do you write? Any diet books ? I am not very healthy.

Figure 1: A clippled dialogue from Persona-Chat.

Several attempts have been made to alleviate the above issues. Methods like special reward shaping to reduce generic responses (Li et al., 2016b) and representing the speakers with latent variables (Li et al., 2016a) were introduced to improve the engagingness of chit-chat systems. A more straightforward approach, which equips chit-chat systems with predefined personas, was proposed accompanied by a novel dataset, Persona-Chat (Zhang et al., 2018b). Figure 1 shows a clipped dialogue from Persona-Chat. Two interlocutors meet for the first time and are having a conversation in order to get to know each other. What makes Persona-Chat unique is that personas of both interlocutors are explicitly described using several profile sentences, facilitating the training of chatbots with configurable and persistent personalities.

Persona-Chat has fueled a growing interest in developing methods for personalized dialogue

[*] Work done during an internship at Microsoft Research.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1417–1427
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

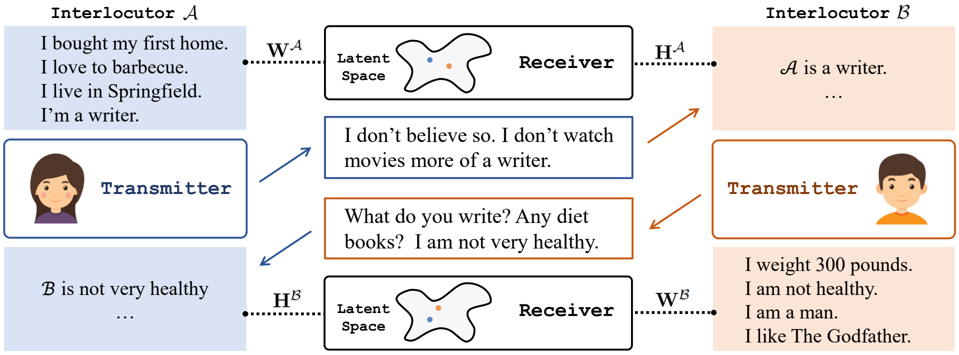

Interlocutor $\mathcal{A}$
I bought my first home.
I love to barbecue.
I live in Springfield.
I'm a writer.
$\mathbf{W}^{\mathcal{A}}$
Latent Space
Receiver
$\mathbf{H}^{\mathcal{A}}$
Interlocutor $\mathcal{B}$
$\mathcal{A}$ is a writer.
…
Transmitter
I don't believe so. I don't watch movies more of a writer.
What do you write? Any diet books? I am not very healthy.
Transmitter
$\mathcal{B}$ is not very healthy
…
$\mathbf{H}^{\mathcal{B}}$
Latent Space
Receiver
$\mathbf{W}^{\mathcal{B}}$
I weight 300 pounds.
I am not healthy.
I am a man.
I like The Godfather.

Figure 2: The overview of $\mathcal{P}^2$ Bot (see text).

generation. Mazaré et al. (2018) incorporated additional data from Reddit to train the model. Wolf et al. (2019b) fine-tuned pretrained language model (Radford et al., 2018) to improve the dialogue generation. Although both works demonstrate promising results, they focus more on mimicking the style of human-like responses, leaving understudied the aspects of explicitly modeling understanding between interlocutors. Our work, instead, takes the perspective of understanding modeling.

According to the research in cognitive science, effective communication creates similar activation maps in the brains of both interlocutors (Hasson et al., 2012), suggesting that understanding between interlocutors is an essential signal for a high-quality chit-chat conversation. For instance, in the conversation shown in Figure 1, the two interlocutors foster understanding either by raising persona-related topics, *"Seen any good movies lately?"*, or by revealing their own personas through answering questions, *"I don't watch movies more of a writer."*. The efforts to build understanding keep the conversation flowing.

Taking into account the above, we propose Persona Perception Bot ($\mathcal{P}^2$ Bot), explicitly modeling the understanding between interlocutors with a transmitter-receiver framework. Distinguished from traditional methods, $\mathcal{P}^2$ Bot highlights a novel concept, **mutual persona perception**, which is better suited to describe the information exchange process that empowers the interlocutors to get to know each other. In order to train $\mathcal{P}^2$ Bot for personalized dialogue generation, we employ supervised training and self-play fine-tuning piloted by reward signals characterizing mutual persona perception. Experiments on the Persona-Chat dataset demonstrate the superiority of our approach over the baselines in both automatic metrics and human evaluations[1].

2 Methodology Overview

The central idea of $\mathcal{P}^2$ Bot is to explicitly model understanding between interlocutors and enhance dialogue generation via mutual persona perception. It comprises two components, **Transmitter** and **Receiver**, respectively responsible for dialogue generation and mutual persona perception. Figure 2 gives an overview of $\mathcal{P}^2$ Bot: interlocutor $\mathcal{A}$ has a persona $\mathbf{w}^{\mathcal{A}}$, described with L profile sentences $\{w_1^{\mathcal{A}}, \cdots, w_L^{\mathcal{A}}\}$. When she first meets the other interlocutor $\mathcal{B}$, they are going to know each other through a N-turn dialogue $(x_1^{\mathcal{A}}, x_1^{\mathcal{B}}, \cdots, x_N^{\mathcal{A}}, x_N^{\mathcal{B}})$, where $x_n^{\mathcal{A}}$ denotes the utterance that $\mathcal{A}$ says in n-th turn and N denotes the number of total turns. Given the entire dialogue history up to n-th turn $\mathbf{h}_n^{\mathcal{A}} = (x_1^{\mathcal{A}}, \cdots, x_{n-1}^{\mathcal{B}})$, **Transmitter** generates $x_n^{\mathcal{A}}$ according to the distribution $p(x_n^{\mathcal{A}} \,|\, \mathbf{w}^{\mathcal{A}}, \mathbf{h}_n^{\mathcal{A}})$, and transmits it to $\mathcal{B}$. The same process applies to $\mathcal{B}$, keeping the conversation flowing.

As the conversation goes on, impressions are gradually built via utterances. For example, when $\mathcal{A}$ says *"I don't watch movies more of a writer."*, the impression that *"$\mathcal{A}$ is a writer."* is left on $\mathcal{B}$'s mind. As mentioned above, a successful conversation helps interlocutors know each other, which means $\mathcal{B}$'s impression of $\mathcal{A}$ should correspond to $\mathcal{A}$'s persona and vice versa. **Receiver** aims to measure the proximity between the built impressions and the actual personas. Specifically, as demonstrated by the dashed black lines in Figure 2, Receiver first projects impressions and personas into a latent space, and then measures the relevance between them based on the *impression encoding* (e.g. $\mathbf{H}^{\mathcal{A}}$, $\mathcal{B}$'s impression on $\mathcal{A}$, projected from $\mathcal{A}$'s

[1] Our code is available at `https://github.com/SivilTaram/Persona-Dialogue-Generation`

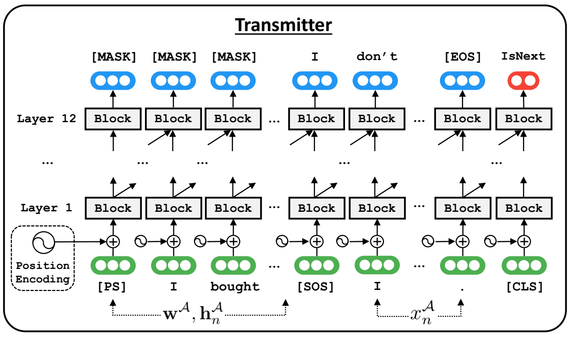


Figure 3: The overall architecture of Transmitter. "Block" is short for "Transformer Block". Arrows ↗ bridge the current block to subsequent blocks of its following layer. Position encoding is to incorporate position information into block by assigning an embedding for each absolute position in the sequence. Here we omit the architecture inside the block, and refer the readers to Vaswani et al. (2017) for more details. [MASK] tokens are ignored in the training objective.

utterances $\mathbf{x}^{\mathcal{A}}$), and *persona encoding* (e.g. $\mathbf{W}^{\mathcal{A}}$, projected from $\mathcal{A}$'s persona $\mathbf{w}^{\mathcal{A}}$)[2]. The relevance scores serve as mutual persona perception rewards, and are further incorporated into the training of Transmitter. Details of the two components are presented in Section 3 and 4.

3 Transmitter

Following previous work (Li et al., 2016b; Zhang et al., 2018b), we treat dialogue generation as a sequence generation problem. Concretely, we employ the pretraining transformer language model introduced in Radford et al. (2018) (i.e. GPT) to initialize Transmitter. The entire training procedure consists of two steps: (1) **Supervised Dialogue Generation**. We optimize Transmitter via maximum likelihood estimation (MLE) on the supervised dialogue generation task. (2) **Self-play Model Fine-tuning**. We simulate dialogues between two randomly paired interlocutors, encouraging Transmitter to learn a policy that maximizes reward signals via reinforcement learning (RL) (Sutton et al., 1999). The design of the reward function considers both language modeling and our proposed mutual persona perception.

3.1 Supervised Dialogue Generation

As illustrated in Figure 3, Transmitter follows the overall architecture of 12 stacked transformer layers to encode context and generate response. Here, the context contains the persona $\mathbf{w}^{\mathcal{A}}$, the dialogue history $\mathbf{h}_n^{\mathcal{A}}$, and several special tokens (e.g. [PS] which indicates the start of persona). Given a training instance $(\mathbf{w}^{\mathcal{A}}, \mathbf{h}_n^{\mathcal{A}}, x_n^{\mathcal{A}})$, the training objective of MLE is to maximize the conditional log-likelihood as:

$$\mathcal{L}_{\text{mle}} = \sum_t \log p_\theta(x_{n,t}^{\mathcal{A}} \,|\, \mathbf{w}^{\mathcal{A}}, \mathbf{h}_n^{\mathcal{A}}, x_{n,<t}^{\mathcal{A}}), \quad (1)$$

where θ is the parameter of Transmitter. $x_{n,t}^{\mathcal{A}}$ means the t-th token in $x_n^{\mathcal{A}}$, and $x_{n,<t}^{\mathcal{A}}$ indicates the token sequence before t-th token. Equation 1, hereafter simplified as $\log p_\theta(x_n^{\mathcal{A}} \,|\, \mathbf{w}^{\mathcal{A}}, \mathbf{h}_n^{\mathcal{A}})$, applies to both $\mathcal{A}$ and $\mathcal{B}$, and we mention $\mathcal{A}$ for the sake of brevity (the same as below).

During inference, beam search is applied to store top-ranked response candidates $\{\hat{x}_n^{\mathcal{A}}\}$, and Transmitter subsequently chooses as prediction the one that maximizes the length-normalized score:

$$x_n^{\mathcal{A}*} = \arg\max_{\hat{x}_n^{\mathcal{A}}} \frac{\log p_\theta(\hat{x}_n^{\mathcal{A}} \,|\, \mathbf{w}^{\mathcal{A}}, \mathbf{h}_n^{\mathcal{A}})}{|\hat{x}_n^{\mathcal{A}}|}. \quad (2)$$

Besides the sequence generation task, inspired by Wolf et al. (2019b), we set up an auxiliary task, **Next Utterance Prediction**. Apart from training Transmitter to generate responses, we also train it to discriminate whether the response is the next utterance of the given context. Concretely, we append a special token [CLS] to the tail of the generated tokens. A classifier is built on top of the token's hidden state in the last transformer layer, as indicated by the red rounded rectangle in Figure 3. In training, for each response, we randomly sample a distractor and train the classifier to give a higher score on the response than the distractor. In inference, the classifier is used to rank response candidates together with Equation 2. Denoting as $y_n = 1$ the signal indicating the generated response $\hat{x}_n^{\mathcal{A}}$ is predicted as the next utterance, Equation 2 is extended as:

$$\begin{aligned} x_n^{\mathcal{A}*} = \arg\max_{\hat{x}_n^{\mathcal{A}}} \Bigg(& \alpha \cdot \frac{\log p_\theta(\hat{x}_n^{\mathcal{A}} \,|\, \mathbf{w}^{\mathcal{A}}, \mathbf{h}_n^{\mathcal{A}})}{|\hat{x}_n^{\mathcal{A}}|} \\ & + (1-\alpha) \cdot \log p_\theta(y_n = 1 | \mathbf{w}^{\mathcal{A}}, \mathbf{h}_n^{\mathcal{A}}, \hat{x}_n^{\mathcal{A}}) \Bigg), \end{aligned} \quad (3)$$

where α is a hyper-parameter.

3.2 Self-play Model Fine-tuning

Although supervised dialogue generation alone can be used to mimic human-like responses, it does not inherently target at understanding. Therefore, we

[2] We take $\mathcal{A}$ as an example, and all are similar to $\mathcal{B}$.

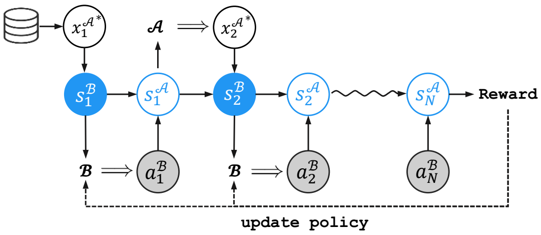


Figure 4: The illustration of the self-play procedure. Arrows ⇒ represent the process of dialogue generation driven by Transmitter. Note that $x_1^{\mathcal{A}^*}$ is directly taken from the dataset as it is difficult to generate high-quality utterances without any dialogue history.

further fine-tune Transmitter using reinforcement learning with the goal of maximizing mutual persona perception. Analogous to Lewis et al. (2017), we apply **self-play** to simulate the communication between two Transmitters, both of which have been trained as described in Section 3.1.

Specifically, we have the two Transmitters communicate with each other for several turns. One Transmitter serves as a user with the parameters frozen, while the other is a learnable **agent**. The parameter of the learnable agent, θ, is fine-tuned during the self-play. Without loss of generality, in our experiments, we let interlocutor $\mathcal{A}$, who starts a conversation, be the user, and correspondingly $\mathcal{B}$ be the learnable agent.

Here we introduce some necessary formulations for modeling our problem with reinforcement learning. A **state** contains the persona and the dialogue history. For example, the state for $\mathcal{B}$ at turn n is defined as $s_n^{\mathcal{B}} = \{\mathbf{w}^{\mathcal{B}}, \mathbf{h}_n^{\mathcal{B}}\}$. An **action** $a_n^{\mathcal{B}}$ is the response to be generated. The action space is infinitely large as the response can be arbitrary long. Taking $s_n^{\mathcal{B}}$ as input, the parameter θ defines a **policy** $p_\theta(a_n^{\mathcal{B}}|s_n^{\mathcal{B}})$, through which the learnable agent generates its response.

As illustrated in Figure 4, when it is $\mathcal{B}$'s turn to speak, $\mathcal{B}$ receives $s_n^{\mathcal{B}}$ and picks $a_n^{\mathcal{B}}$ according to the policy p_θ. As for $\mathcal{A}$, it receives $s_n^{\mathcal{A}}$ and generates the response $x_n^{\mathcal{A}^*}$ to simulate a user. $\mathcal{A}$ and $\mathcal{B}$ alternately produce responses till the number of turns exceeds the given limit. Once a complete dialogue is generated, the reward is collected to optimize θ using policy gradient (Sutton et al., 1999). Denoting as $R(a_n^{\mathcal{B}})$ the reward $\mathcal{B}$ gets at turn n (more details are provided later), we can optimize it by maximizing the following objective:

$$\mathcal{L}_{\text{rl}} = \mathbb{E}_{a_n^{\mathcal{B}} \sim p_\theta(a_n^{\mathcal{B}}|s_n^{\mathcal{B}})}[R(a_n^{\mathcal{B}})]. \tag{4}$$

Applying likelihood ratio trick, θ is updated by ascending the following gradient:

$$\nabla_\theta \mathcal{L}_{\text{rl}} = \mathbb{E}_{a_n^{\mathcal{B}} \sim p_\theta(a_n^{\mathcal{B}}|s_n^{\mathcal{B}})} \nabla_\theta \log p_\theta(a_n^{\mathcal{B}}|s_n^{\mathcal{B}}) R(a_n^{\mathcal{B}}). \tag{5}$$

As aforementioned, the space of action $a_n^{\mathcal{B}}$ is infinite. In practice, REINFORCE algorithm (Williams, 1992) is leveraged to approximate Equation 5 by sampling $a_n^{\mathcal{B}}$ from policy $p_\theta(a_n^{\mathcal{B}}|s_n^{\mathcal{B}})$. Furthermore, subtracting a baseline (Weaver and Tao, 2001), here the mean reward of a mini-batch, is applied on $R(a_n^{\mathcal{B}})$ to reduce variance. The agent samples tokens one by one through *multinomial sampling* over the output distribution of $\mathcal{B}$, until the special token `[EOS]` is sampled or exceeding the maximum allowed decoding step (e.g. 32). Compared to beam search sampling, multinomial sampling provides more diversities.

3.3 Reward Shaping (RS)

As described in Section 1, we believe that a high-quality chit-chat conversation should highlight both human language modeling and mutual persona perception. Bearing this in mind, we design three rewards to address language style, discourse coherence and mutual persona perception respectively.

RS.1 Language Style The generated responses should conform to human language styles, which we believe can be evaluated by a pretrained language model (i.e. GPT). After length normalization, the score for $a_n^{\mathcal{B}}$ is given as:

$$R_1(a_n^{\mathcal{B}}) = \frac{1}{|a_n^{\mathcal{B}}|} \sum_t \log p_{\text{lm}}(a_{n,t}^{\mathcal{B}} \,|\, a_{n,<t}^{\mathcal{B}}), \tag{6}$$

where $a_{n,t}^{\mathcal{B}}$ and $a_{n,<t}^{\mathcal{B}}$ have similar denotation as the previously mentioned $x_{n,t}^{\mathcal{A}}$ and $x_{n,<t}^{\mathcal{A}}$.

RS.2 Discourse Coherence The language score is evaluated individually, without considering the discourse coherence. However, a reasonable response should establish links in meaning with context, which is also an important aspect of human-like responses. To take into account the discourse coherence, we employ the well-trained Next Utterance Predictor (mentioned in Section 3.1). The reward is given by the log probability of $a_n^{\mathcal{B}}$ being the next utterance of $s_n^{\mathcal{B}}$:

$$R_2(a_n^{\mathcal{B}}) = \log p_\theta(y_n = 1 \,|\, a_n^{\mathcal{B}}, s_n^{\mathcal{B}}). \tag{7}$$

RS.3 Mutual Persona Perception RS.1 and RS.2 only steer the agent training process towards human-like responding. They do not explicitly encourage understanding between interlocutors. Therefore, we meticulously design the reward to characterize mutual persona perception. Contrast from RS.1 and RS.2, mutual persona perception is a long-term goal throughout the whole dialogue, meaning that the effect of current action might only play out some time later. For instance, receiving *"what are your hobbies?"* from $\mathcal{B}$, it is highly likely that $\mathcal{A}$'s response is relevant to $\mathcal{A}$'s hobbies. This suggests that, not only $\mathcal{A}$'s response but also $\mathcal{B}$'s initial question contributes to mutual persona perception. Denoting as γ the discount factor indicating how far ahead $\mathcal{B}$ looks, the reward of mutual persona perception for $a_n^{\mathcal{B}}$ is defined as:

$$R_3(a_n^{\mathcal{B}}) = r(a_n^{\mathcal{B}}) + \sum_{k=n+1}^{N} \Big(\gamma^{2(k-n)-1} r(x_k^{\mathcal{A}^*}) + \gamma^{2(k-n)} r(a_k^{\mathcal{B}}) \Big), \tag{8}$$

where $r(a_n^{\mathcal{B}})$ is the persona perception score that $\mathcal{B}$ obtains in n-th turn, and $r(x_k^{\mathcal{A}^*})$ is defined likewise. $r(a_n^{\mathcal{B}})$ can be computed using a score function:

$$r(a_n^{\mathcal{B}}) = \text{score}(a_n^{\mathcal{B}}, \mathbf{w}^{\mathcal{B}}). \tag{9}$$

In $\mathcal{P}^2$ Bot, the score function comes from Receiver, which will be elaborated in Section 4. The final reward $R(a_n^{\mathcal{B}})$ for $a_n^{\mathcal{B}}$ is a weighted sum of the rewards listed above:

$$R = \lambda_1 R_1 + \lambda_2 R_2 + \lambda_3 R_3, \tag{10}$$

where λ_1, λ_2 and λ_3 are hyper-parameters.

4 Receiver

Receiver is devised to measure the proximity between the built impressions and the actual personas, implemented by negative sampling. Specifically, in training, we randomly sample a persona distractor $\mathbf{w}^{\mathcal{Z}}$. Receiver is trained to identify the real persona $\mathbf{w}^{\mathcal{A}}$ from $\{\mathbf{w}^{\mathcal{A}}, \mathbf{w}^{\mathcal{Z}}\}$. In inference, for each utterance, Receiver is responsible for providing a reasonable relevance score, to model our proposed mutual persona perception. The score subsequently joins the self-play fine-tuning on Transmitter as part of the rewards, as in Equation 8.

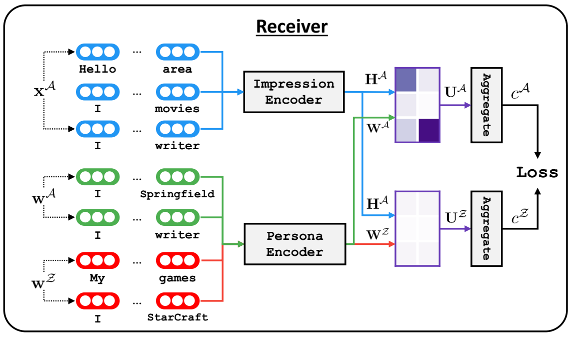

Receiver
$\mathbf{x}^{\mathcal{A}}$
Hello
area
I
movies
I
writer
$\mathbf{w}^{\mathcal{A}}$
I
Springfield
I
writer
$\mathbf{w}^{\mathcal{Z}}$
My
games
I
StarCraft
Impression Encoder
Persona Encoder
$\mathbf{H}^{\mathcal{A}}$
$\mathbf{W}^{\mathcal{A}}$
$\mathbf{H}^{\mathcal{A}}$
$\mathbf{W}^{\mathcal{Z}}$
$\mathbf{U}^{\mathcal{A}}$
$\mathbf{U}^{\mathcal{Z}}$
Aggregate
Aggregate
$c^{\mathcal{A}}$
$c^{\mathcal{Z}}$
Loss

Figure 5: The overall architecture of Receiver (see text).

4.1 Training

As illustrated in Figure 5, Receiver contains two different encoders for impression and persona respectively. Initialized by BERT (Devlin et al., 2019), both encoders provide deep contextualized representations for each token. Then we average all the representations, yielding a fixed d-dimensional vector for one sentence. In this way, feeding $(x_1^{\mathcal{A}}, x_2^{\mathcal{A}}, \cdots, x_N^{\mathcal{A}})$ into the impression encoder consecutively, we obtain the impression encoding $\mathbf{H}^{\mathcal{A}} \in \mathbb{R}^{N \times d}$. The persona encoding $\mathbf{W}^{\Delta} \in \mathbb{R}^{L \times d}$ is produced likewise, where $\Delta \in \{\mathcal{A}, \mathcal{Z}\}$. The relevance score matrix $\mathbf{U}^{\Delta}$ is computed via the scaled dot product (Vaswani et al., 2017):

$$\mathbf{U}^{\Delta} = \frac{\mathbf{H}^{\mathcal{A}} (\mathbf{W}^{\Delta})^{\top}}{\sqrt{d}}, \in \mathbb{R}^{N \times L}. \tag{11}$$

In essence, Receiver is expected to capture fine-grained correlations between the persona and the dialogue. However, we do not have access to the golden fine-grained correlations. The only thing we know is that, compared with $\mathbf{W}^{\mathcal{Z}}$, $\mathbf{H}^{\mathcal{A}}$ is more correlated to $\mathbf{W}^{\mathcal{A}}$. Since the comparison is at a coarse granularity, we gather $\mathbf{U}^{\Delta}$ into the cumulative score c^{Δ} through an aggregate function Agg, as shown in Figure 5. To encourage $c^{\mathcal{A}}$ while at the same time depress $c^{\mathcal{Z}}$, we design a marginal loss $\mathcal{L}_{\text{rec}}$, which makes $c^{\mathcal{A}}$ larger than $c^{\mathcal{Z}}$ by a margin m. Moreover, considering that an utterance generally relates to zero or one profile, L_1 regularization is enforced to make $\mathbf{U}^{\Delta}$ sparse. Combining all of these, the training loss for Receiver is:

$$\mathcal{L}_{\text{rec}} = \max(0, m + c^{\mathcal{Z}} - c^{\mathcal{A}}) + \beta \cdot |\mathbf{U}^{\Delta}|_1, \tag{12}$$

where β is a hyper-parameter for penalty.

As for Agg, one straightforward way is to average over all positions of $\mathbf{U}^{\Delta}$. However, it maximizes every entry in $\mathbf{U}^{\mathcal{A}}$, including all those that

Category	Model	Original			Revised		
		Hits@1(%)↑	ppl↓	F1(%)↑	Hits@1(%)↑	ppl↓	F1(%)↑
Retrieval	KV Profile Memory	54.8	-	14.25	38.1	-	13.65
	Dually Interactive Matching	78.8	-	-	**70.7**	-	-
Generative	Generative Profile Memory	10.2	35.01	16.29	9.9	34.94	15.71
	Language Model	-	50.67	16.30	-	51.61	13.59
	SEQ2SEQ-ATTN	12.5	35.07	16.82	9.8	39.54	15.52
Pretrain Fintune	Lost In Conversation	17.3	-	17.79	16.2	-	16.83
	Transfertransfo	**82.1**	17.51	19.09	-	-	-
	$\mathcal{P}^2$ BOT (Our)	81.9 [0.1]	**15.12** [0.16]	**19.77** [0.08]	68.6 [0.2]	**18.89** [0.11]	**19.08** [0.07]

Table 1: Automatic evaluation results of different methods on the PERSONA-CHAT dataset. The standard deviation $[\sigma]$ (across 5 runs) of $\mathcal{P}^2$ BOT is also reported. All the results were evaluated on the dev set since the test set was not publicly available.

should not be activated (e.g. relevance scores between unrelated profile sentences and utterances), introducing unnecessary noise into the training of Transmitter. To alleviate the problem, we choose to implement Agg as a controllable weighted function, which summarizes $\mathbf{U}^{\Delta}_{n,:}$ as:

$$Agg(\mathbf{U}^{\Delta}_{n,:}) = \frac{\sum_{k=1}^{L} \exp(\mathbf{U}^{\Delta}_{n,k}/\tau) \cdot \mathbf{U}^{\Delta}_{n,k}}{\sum_{k=1}^{L} \exp(\mathbf{U}^{\Delta}_{n,k}/\tau)}, \quad (13)$$

where *temperature* $\tau > 0$ is a tunable parameter (Hinton et al., 2015) controlling the evolution of Agg. In the beginning, Agg behaves close to average pooling. As τ anneals, Agg gradually focuses more on the highest relevance score. In this way, noise reduces as training goes on. Finally, c^{Δ} is given by:

$$c^{\Delta} = \frac{1}{N} \sum_{n=1}^{N} Agg(\mathbf{U}^{\Delta}_{n,:}). \quad (14)$$

4.2 Inference

Given $x^{\mathcal{A}}_n$ and $\mathbf{w}^{\mathcal{A}}$, Receiver employs the following function to obtain $x^{\mathcal{A}}_n$'s persona perception score, further modeling mutual persona perception as in Equation 9:

$$\text{score}(x^{\mathcal{A}}_n, \mathbf{w}^{\mathcal{A}}) = \frac{Agg\big(\mathbf{H}^{\mathcal{A}}_{n,:}(\mathbf{W}^{\mathcal{A}})^{\top}\big)}{\sqrt{d}}, \quad (15)$$

where $\mathbf{H}^{\mathcal{A}}_{n,:}$ and $\mathbf{W}^{\mathcal{A}}$ are the impression encoding and persona encoding for $x^{\mathcal{A}}_n$ and $\mathbf{w}^{\mathcal{A}}$ respectively.

5 Experiment

We conducted experiments on the dataset PERSONA-CHAT, assessing $\mathcal{P}^2$ BOT using both automatic metrics and human evaluations. To verify the effectiveness of our proposed mutual persona perception, we perform a thorough model analysis in Section 5.3. Finally, we probe Receiver's capability on perceiving persona in Section 5.4.

5.1 Implementation Details

PERSONA-CHAT dataset contains 8,939 / 1,000 multi-turn dialogues conditioned on 1,155 / 100 personas for train / dev. Each persona is described with at least 5 profile sentences. To make it more challenging, PERSONA-CHAT also provides *revised* personas by rephrasing, generalizing or specializing the *original* ones. For example, *"I am overweight."* is revised from *"I weight 300 pounds."*.

Our implementation was based on PyTorch (Paszke et al., 2019), ParlAI (Miller et al., 2017), and HuggingFace's transformers library (Wolf et al., 2019a). We used Adam (Kingma and Ba, 2015) optimizer with a learning rate of 6.25e-5 for both Receiver and Transmitter in supervised learning. In the training of Receiver, τ reduced linearly from 10 to 0.5. In the self-play phase of Transmitter, the learning rate was set as 1e-6. The hyperparameters m, α, β, γ, λ_1, λ_2 and λ_3 were set as 0.4, 0.1, 1e-4, 0.5, 0.4, 0.1 and 0.5 respectively. The supervised training of Transmitter lasted for 2 epochs, and the self-play fine-tuning comprised 2000 dialogues, where the number of turns was 3. The beam search size was set as 2.

5.2 Methods Comparison

Our baselines fall into three categories: retrieval-based, generative-based and pretrain-finetune-based models. Among the retrieval-based baselines, *KV Profile Memory* (Zhang et al., 2018b) was the official baseline which employed the memory network along with profile information, and

Model	1 (%)	2 (%)	3 (%)	4 (%)	Avg
Lost In Conversation	26.3	**48.7**	22.0	3.0	2.017
Transfertransfo	**41.7**	25.3	**28.7**	4.3	1.956
$\mathcal{P}^2$ Bot (Our)	18.9	26.3	28.6	**26.2**	**2.621**

Table 2: Human evaluation results.

Dually Interactive Matching Network (Gu et al., 2019) proposed a dual matching architecture to match between the responses and their corresponding contexts. *Language Model*, *Generative Profile Memory* (Zhang et al., 2018b) and Seq2Seq with attention mechanism (Bahdanau et al., 2015) were implemented as generative baselines for dialogue generation. The remaining methods were all pretrain-finetune-based. *Transfertransfo* (Wolf et al., 2019b)[3] achieved the state-of-the-art performance on automatic metrics, while *Lost In Conversation*[4] topped the human evaluations (Dinan et al., 2019). Analogous to our approach, they employed the pretrained language model GPT to initialize their models, and then fine-tuned it on the dataset.

Table 1 shows the experimental results on automatic metrics. Following Zhang et al. (2018b), we reported the official automatic metrics to evaluate the methods: **Hits@1**, **Perplexity (ppl)** and **F1**. Given 20 response candidates, Hits@1 is the probability that the real response ranks the highest according to the model. Perplexity measures the negative log likelihood of the correct sequence output by the model, lower values indicating better performance. F1 is the harmonic mean of word-level precision and recall. As observed, our approach outperforms almost all baselines and achieves new state-of-the-art performance on ppl and F1, with highly competitive performance on Hits@1. In the revised mode, our approach still achieves the best performance, obtaining a relative improvement of 13.4% on F1 against the strongest baseline. It is worth noting that we also tried to employ F1 as the reward, but the result is far from satisfactory.

As mentioned in Dinan et al. (2019), no automatic metric is perfect for evaluating such an open-domain task. Hence, we also performed crowd-sourced **human evaluations** on the state-of-the-art baselines (i.e. Transfertransfo & Lost In Conversation) and our proposed $\mathcal{P}^2$ Bot. Concretely, on the original dev set, we randomly sampled 200 responses generated by these methods and asked each worker to rate them. The rating ranges from 1 to 4. 1 means the response is good only in terms of grammar and sentence structure; 2 means in addition to valid grammar, the response is also coherent with the context; 3 means the coherent response is meanwhile interesting and informative, instead of just a simple response like "Yes"; And 4 means the response is consistent with the persona of the interlocutor, which is of extreme importance for the task of reflecting whether the model can effectively utilize the persona information. As shown in Table 2, the results are consistent with the automatic evaluation results, demonstrating the superiority of $\mathcal{P}^2$ Bot against the baselines. We also conducted Wilcoxon signed-rank tests between our method and the baselines and the results show the improvements are significant with $p < 0.05$.

Variant	Hits@1(%)↑	F1(%)↑	BLEU(%)↑
$\mathcal{P}^2$ Bot-S	68.7	18.14	0.56
- Persona	65.5	17.77 (-2.0%)	0.57 (+ 1.8%)
- Next	17.6	18.11 (-0.1%)	0.55 (- 1.8%)
+ RS.1	68.4	18.32 (+0.9%)	0.60 (+ 7.1%)
↪ + RS.2	68.6	18.41 (+1.5%)	0.61 (+ 8.9%)
↪ + RS.3	68.6	19.08 (+5.2%)	0.75 (+33.9%)

Table 3: Variant analysis results on Persona-Chat revised mode, along with relative improvements (shown inside brackets) compared with $\mathcal{P}^2$ Bot-S. BLEU refers to the cumulative 4-gram BLEU score. "-Persona" means dialogue generation without personas; "-Next" ablates the auxiliary task mentioned in Section 3.1; "+RS.1" means only using Language Style score as the reward in the self-play fine-tuning phase; "↪ +RS.2" means adding Discourse Coherence to the reward on the basis of RS.1; "↪ +RS.3" is equivalent to our proposed $\mathcal{P}^2$ Bot.

5.3 Model Analysis

Variant Analysis We conducted variant analysis on $\mathcal{P}^2$ Bot to investigate the influence of RS.1, RS.2 and RS.3. Another metric **BLEU** (Papineni et al., 2002), which evaluates the quality of response, was introduced to make the analysis more comprehensive. We show the variant analysis results in Table 3, where $\mathcal{P}^2$ Bot-S is the variant of $\mathcal{P}^2$ Bot which is trained only in the supervised setting. As expected, the results on Hits@1 validate the important role of the auxiliary task. Across all the variants, the gains in BLEU and F1 are very small, revealing the difficulty in improving them. Nevertheless, solely by adding RS.3, we obtained a 25% relative improvement on BLEU, indicating the effectiveness of our proposed mutual persona

[3] http://github.com/huggingface/transfer-learning-conv-ai

[4] http://github.com/atselousov/transformer_chatbot

PERSONA	i. I love new kids on the block. ii. I was born in the early 80's. iii. I also like old school hip hop. iv. My favorite toy as a child as my lite brite.	i. I am a blonde girl with really short hair. ii. I love wearing skinny jeans and leggings. iii. I 'm rather skinny as I like to stay in shape. iv. My favorite hobbies are listening to music and playing video games.	i. I've an allergy to water. ii. I'm a famous twitch streamer. iii. I helped design the game StarCraft. iv. I am working on building a robot to be my body double.
CONTEXT	Do you like clowns? They horrify me!	What video game? I relax by knitting. Love it!	Cool. I did hardcore exercise four hours everyday.
* HUMAN	No! My favorite toy as a kid was lite brite because I was so scared!	I was playing Fallout. What do you look like? I' m blonde thin.	That's a lot of exercising. I hate exercising. I just work on my robot most days .
* $\mathcal{P}^2$ BOT	I do not like them at all. I am a little old for them .	I play RPGs. I like to wear skinny jeans and tees .	I do not exercise at all. I'm busy with my robot project .
* TRANSFERTRANSFO	I do not have time for clowns. Do you like to listen to music ?	My fav color is blue, I have a lot of friends in my group.	I work at a computer company. I could make you an Android!
* LOST IN CONVERSATION	I love clowns. they are my favorite.	I love HALO 3, what do you knit?	That sounds like a lot of fun !

Table 4: Sampled responses(*) by Human, $\mathcal{P}^2$ BOT and the state-of-the-art baselines.

Model	Original		Revised	
	Hits@1↑	MRR↑	Hits@1↑	MRR↑
Random	3.1	0.2	3.1	0.2
IR	67.5	20.9	9.7	2.2
Receiver	**93.8**	**37.5**	**78.2**	**16.6**

Table 5: Experimental results on Persona Perception.

perception. Similar conclusions can be drawn from the trend of F1.

Case Study For a more comprehensive comparison, we show in Table 4 some randomly sampled responses of different methods. The results suggest the responses generated by our approach are more human-like. As observed, benefiting from our proposed mutual persona perception, the responses of $\mathcal{P}^2$ BOT are more consistent, engaging and informative. For instance, in the last example in Table 4, the response *"I'm busy with my robot project"* explicates why the speaker does not exercise, meanwhile revealing that he is working on the robot, as depicted in his persona.

Error Analysis Though our approach works well in most cases, we observed that the self-play simulation might fall into repeated cycles after rounds of training, as the challenge mentioned by Li et al. (2016b). Another issue is that the bots sometimes ask redundant questions in our approach, which might be due to inappropriate hyperparameters in reward shaping.

5.4 Persona Perception Probing

Receiver plays an important role in our approach, and we are interested in its capability on perceiving persona. Therefore, we conducted experiments on a synthesized dataset. We constructed the dataset by sampling 31 persona distractors for each dialogue in PERSONA-CHAT. Two widely used ranking metrics were used to evaluate the performance: **Hits@1** and **Mean Reciprocal Rank (MRR)**. Hits@1 is the same metric as the one mentioned in Section 5.2, except that the candidate size is 32. Given a dialogue and the complete set of profile sentences, MRR is the average reciprocal ranks of the dialogue-relevant profile sentences. Two simple baselines Random and IR (Sordoni et al., 2015) were chosen for comparison. Table 5 shows the experimental results of different methods on the synthesized dataset. As observed, our approach achieved excellent results on both original and revised modes. For example, compared with the IR baseline, our approach achieved an absolute improvement of 26.3% on Hits@1 in the original mode. In addition, the surprising results in the revised mode further demonstrate Receiver's capability to perceive rephrased persona.

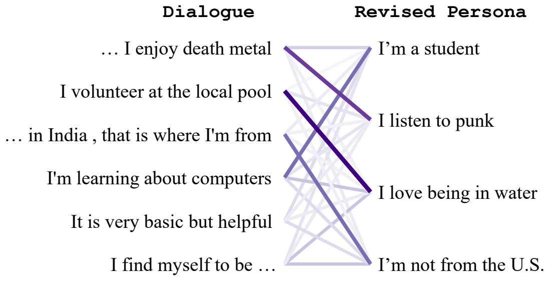


Figure 6: Visualization of the relevance scores between a sampled dialogue and its corresponding revised persona. Deeper color means higher score. We omit some context due to space limitation.

To further understand the trained Receiver, we visualize the relevance scores between a sampled dialogue and its corresponding revised persona in Figure 6. As illustrated, the relevance scores between related profile sentences and dialogue utterances are significantly higher. For example, the utterance *"I volunteer at the local pool"* from the interlocutor implies the profile *"I love being in the water"*, and our Receiver successfully captures the relevance between them.

6 Related Work

Methods to build open-domain dialogue systems generally fall into two major categories: retrieval-based and generative-based. Retrieval-based methods retrieve response candidates and rank them based on the matching scores with the dialogue (Sordoni et al., 2015; Wu et al., 2017; Gu et al., 2019). Generative-based methods typically use SEQ2SEQ model as the backbone (Sutskever et al., 2014; Bahdanau et al., 2015; Serban et al., 2017; Wolf et al., 2019b), where the encoder extracts the information in an utterance and the decoder generates the response. Our work adopts a similar architecture. Besides supervised learning, researchers also explore reinforcement learning based methods. Lewis et al. (2017) applied reinforcement learning for negotiation dialogues and showed it outperforms supervised learning when negotiating with humans. Yang et al. (2018) proposed to generate dialogue responses by dual learning based domain adaptation. Zhang et al. (2018a) built a coherence model to provide the reward signal for penalizing dull responses. Liu et al. (2019) employed reinfrocement learning to learn an intermediate structure span. Our approach differs from this line of work in that we focus on improving personalized dialogues via mutual persona perception, which has not yet been explored before.

More recently, under the topic of dialogue personalizing, Zemlyanskiy and Sha (2018) proposed a post-processing method to re-rank candidates generated by beam search, while Olabiyi et al. (2019) employed adversarial approaches to solve the consistency problem on interlocutors' names. Madotto et al. (2019) applied meta-learning to quickly adapt to new speakers, and Tigunova et al. (2019) extracted user attributes from daily dialogues. Compared with them, our work enhances persona based dialogue generation from a novel perspective.

Furthermore, researchers explored to generate diverse responses conditioned on persona (Song et al., 2019, 2020). Personalization in goal-oriented dialogue systems has also received some attention (Joshi et al., 2017; Luo et al., 2019). The researches focus more on making the goal-oriented bots adjust the response according to different user profiles, while we aim to endow bots with persistent personalities.

7 Conclusion & Future Work

We propose $\mathcal{P}^2$ BOT, a transmitter-receiver framework which explicitly models understanding between interlocutors. Under this framework, mutual persona perception is incorporated as a reward signal to achieve the personalized dialogue generation. Experiments on a large public dataset PERSONA-CHAT demonstrate the effectiveness of our approach. For future work, we would like to extend Receiver to conversational recommender systems. After turns of chatting, the agent should be able to infer the user's persona, based on which personalized contents can be recommended.

Acknowledgments

We thank all the anonymous reviewers for their valuable comments. This work was supported in part by National Natural Science Foundation of China (U1736217 and 61932003), and National Key R&D Program of China (2019YFF0302902).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, May 7-9, 2015, San Diego, CA, USA*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2019*, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2). *CoRR*, abs/1902.00098.

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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, Hong Kong, China. Association for Computational Linguistics.

Uri Hasson, Asif A Ghazanfar, Bruno Galantucci, Simon Garrod, and Christian Keysers. 2012. Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in cognitive sciences*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Chaitanya K. Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *CoRR*, abs/1706.07503.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, May 7-9, 2015, San Diego, CA, USA*.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, Copenhagen, Denmark. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, Berlin, Germany. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, Austin, Texas. Association for Computational Linguistics.

Qian Liu, Bei Chen, Haoyan Liu, Jian-Guang Lou, Lei Fang, Bin Zhou, and Dongmei Zhang. 2019. A split-and-recombine approach for follow-up query analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, Hong Kong, China. Association for Computational Linguistics.

Liangchen Luo, Wenhao Huang, Qi Zeng, Zaiqing Nie, and Xu Sun. 2019. Learning personalized end-to-end goal-oriented dialog. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, The 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, January 27 - February 1, 2019, Honolulu, Hawaii, USA*. AAAI Press.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, Florence, Italy. Association for Computational Linguistics.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, Brussels, Belgium. Association for Computational Linguistics.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2017*, Copenhagen, Denmark. Association for Computational Linguistics.

Oluwatobi Olabiyi, Anish Khazane, Alan Salimov, and Erik T. Mueller. 2019. An adversarial learning framework for a persona-based multi-turn dialogue model. *CoRR*, abs/1905.01992.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL 2002*, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Curran Associates, Inc.

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

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence, AAAI 2019, February 4-9, 2017, San Francisco, California, USA*. AAAI Press.

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020, February 7-12, 2020, New York City, New York, USA*. AAAI Press.

Haoyu Song, Weinan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019, August 10-16, 2019, Macao, China*.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, NIPS 2014, December 8-13 2014, Montreal, Quebec, Canada*.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12: Annual Conference on Neural Information Processing Systems 1999, NIPS 1999, November 29 - December 4, 1999, Denver, Colorado, USA,*. The MIT Press.

Anna Tigunova, Andrew Yates, Paramita Mirza, and Gerhard Weikum. 2019. Listening between the lines: Learning personal attributes from conversations. In *Proceedings of the World Wide Web Conference, WWW 2019, May 13-17, 2019, San Francisco, CA, USA*. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, NIPS 2017, December 4-9, 2017, Long Beach, CA, USA*.

Lex Weaver and Nigel Tao. 2001. The optimal reward baseline for gradient-based reinforcement learning. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, UAI 2001, August 2-5, 2001, University of Washington, Seattle, Washington, USA*. Morgan Kaufmann.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019a. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019b. TransferTransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, Vancouver, Canada. Association for Computational Linguistics.

Min Yang, Wenting Tu, Qiang Qu, Zhou Zhao, Xiaojun Chen, and Jia Zhu. 2018. Personalized response generation by dual-learning based domain adaptation. *Neural Networks*.

Yury Zemlyanskiy and Fei Sha. 2018. Aiming to know you better perhaps makes me a more engaging dialogue partner. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018*, Brussels, Belgium. Association for Computational Linguistics.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018a. Reinforcing coherence for sequence to sequence model in dialogue generation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. 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 2018*, Melbourne, Australia. Association for Computational Linguistics.