BoB: BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data

Haoyu Song[1], Yan Wang, Kaiyan Zhang[1], Wei-Nan Zhang[1]*, Ting Liu[1]
[1]Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, Heilongjiang, China
{hysong,kyzhang,wnzhang,tliu}@ir.hit.edu.cn
yanwang.branden@gmail.com

Abstract

Maintaining consistent personas is essential for dialogue agents. Although tremendous advancements have been brought, the limited-scale of annotated persona-dense data are still barriers towards training robust and consistent persona-based dialogue models. In this work, we show how the challenges can be addressed by disentangling persona-based dialogue generation into two sub-tasks with a novel BERT-over-BERT (BoB) model. Specifically, the model consists of a BERT-based encoder and two BERT-based decoders, where one decoder is for response generation, and another is for consistency understanding. In particular, to learn the ability of consistency understanding from large-scale non-dialogue inference data, we train the second decoder in an unlikelihood manner. Under different limited data settings, both automatic and human evaluations demonstrate that the proposed model outperforms strong baselines in response quality and persona consistency.

1 Introduction

Various approaches have been explored to introduce explicit personas in dialogue models (Qian et al., 2018; Song et al., 2019; Zheng et al., 2020; Liu et al., 2020). The PERSONA can be defined as a composite of elements of identity, such as profiles and background personal facts. In persona-based dialogues, the generated responses are conditioned not only on the dialogue context but also on some predefined personas, so the presenting personality could be more consistent.

Existing persona-based dialogue models heavily utilize a set of persona-related dialogue data (Wolf et al., 2019; Golovanov et al., 2019), such as the PersonaChat (Zhang et al., 2018). This kind of crowd-sourced dataset covers rich persona features,

---

*Wei-Nan Zhang is the corresponding author.

Persona:	I've a son who is in junior high school
Query:	You have any children?
GPT-2:	**No kids**. I work at home depot so I'm busy.

Figure 1: A 12-layer GPT2 finetuned on PersonaChat dataset still generates an inconsistent response.

namely "persona-dense". Nevertheless, the scale of such crowd-sourced datasets is limited by the expensive costs: two annotators are asked to act the part of a given provided persona and chat naturally to get to know each other during the conversation. On the other hand, conversations in daily life are not always persona-related. According to Twitter content analysis, less than 10% messages on Twitter reveal personal anecdote or activities at home or work and even less for personally identifiable information (Naaman et al., 2010; Humphreys et al., 2014). As a result, the large-scale data collected from social media would only contain a limited amount of persona-related dialogues, which is "persona-sparse". The limited-scale of crowd-sourced data and the persona-sparsity in large-scale data present one common challenge: a model trained on limited personalized data cannot sufficiently understand persona consistency. As shown in Figure 1, a 12-layer GPT2 (Radford et al., 2019) finetuned on the PersonaChat dataset still shows a lack of consistency.

After rethinking the essence of persona-based dialogue generation, we can find that it requires the dialogue agent to own the capabilities to 1) understand the persona-response consistency and 2) generate a persona-related response given the dialogue context. Obviously, an ideal dataset that satisfies both features are difficult to annotate. However, once we disentangle persona-based dialogue generation into two sub-tasks: consistency understanding and dialogue generation, it is easy to find

abundant data resources for them. For consistency understanding, we may leverage large-scale non-dialogue inference data, such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) as the training data. As for dialogue generation, we already have various large-scale persona-sparse datasets.

Inspired by the aforementioned motivation, in this work, we explore to learn a consistent persona-based dialogue model from limited personalized dialogues, with the assistance of large-scale non-dialogue inference data. Specifically, the proposed model consists of an encoder $\mathbb{E}$, an auto-regressive decoder $\mathbb{D}_1$ for response generation, and a bidirectional decoder $\mathbb{D}_2$ for consistency understanding. Given personas P and dialogue query Q, the $\mathbb{E}$ and $\mathbb{D}_1$ jointly work in an encoder-decoder manner to capture a typical query to response mapping $F_G(S|Q, P)$, and generate a coarse response representation R_1. Then R_1 and personas P are fed into the bidirectional decoder $\mathbb{D}_2$ to map R_1 to final response representations R_2: $F_U(R_2|S, P)$. Since the consistency understanding part $F_U(R|S, P)$ is independent of the dialogue query Q, it can be learned on non-dialogue inference datasets. Here an unlikelihood training objective (Welleck et al., 2019a) is applied to make contradicted cases in the inference data less likely so that $\mathbb{D}_2$ could acquire the ability of consistency understanding.

We initialize all modules from BERT (Devlin et al., 2019) and name the proposed model BERT-over-BERT (BoB). To verify the effectiveness of our model, we experiment on two limited data scenarios: 1) a persona-dense scenario (Zhang et al., 2018) with low-resource settings (Zhao et al., 2019), and 2) a persona-sparse scenario (Zheng et al., 2019). Both automatic and human evaluations indicate that our model generalizes well under different settings and outperforms strong baselines on most metrics, especially on persona consistency.

Contributions in this work are three-fold:

- We disentangled the task of persona-based dialogue generation into two sub-tasks: consistency understanding and dialogue generation.

- A BERT-based generative framework, BoB, was proposed for training persona-based dialogue models from limited data.

- An unlikelihood training method with non-dialogue inference data was introduced to enhance persona consistency understanding.

2 Related Work

Persona-based Dialogues Recent studies on persona-based dialogue generation focus on a data-driven manner. They learn persona-related features directly from personalized dialogue datasets, either with implicit persona embeddings (Li et al., 2016b) or with explicit profiles (Qian et al., 2018) and personal facts (Mazaré et al., 2018). Following this research line, more sophisticated neural models are emerging, such as modeling mutual-persona (Liu et al., 2020) and multi-stage persona-based dialogue generation (Song et al., 2020a).

Meanwhile, various pre-training methods have also been applied in this field. Wolf et al. (2019) and Golovanov et al. (2019) show that fine-tuning pre-trained GPT on the persona-dense dataset can improve the quality of generated responses. Zheng et al. (2020) propose an attention-routing mechanism in a GPT-based model to control the flow of persona information. Lin et al. (2020) explore how to leverage BERT model for dialogue generation. Different large-scale pretrained chatbots (Roller et al., 2020; Madotto et al., 2020) also show their effectiveness on persona-based dialogues.

Disentangled Representation The concept of "disentangling" can be defined as transformations that only change some properties of the underlying model while leaving all other properties invariant (Higgins et al., 2018). The variational autoencoder (Kingma and Welling, 2013) could be regarded as a disentangled representation learning framework, and various methods are built within it (Kim and Mnih, 2018; Locatello et al., 2019).

Unlikelihood Training Likelihood tries to maximize the probability of target sequence, while unlikelihood corrects known biases by minimizing the probability of negative candidates (Welleck et al., 2019a). Closely related to our work, Li et al. (2020) first explored unlikelihood training in addressing dialogue logical contradictions. They get contradicted dialogues from PersonaChat according to DNLI (Welleck et al., 2019b), a PersonaChat-oriented dialogue inference dataset. Then unlikelihood training is applied to reduce the probability of contradicted responses. Different from Li et al. (2020), with carefully designed decoders, our model could learn from large-scale non-dialogue inference datasets, making it generalizable to different scenarios, such as persona-dense and persona-sparse datasets, as will be seen in our experiments.
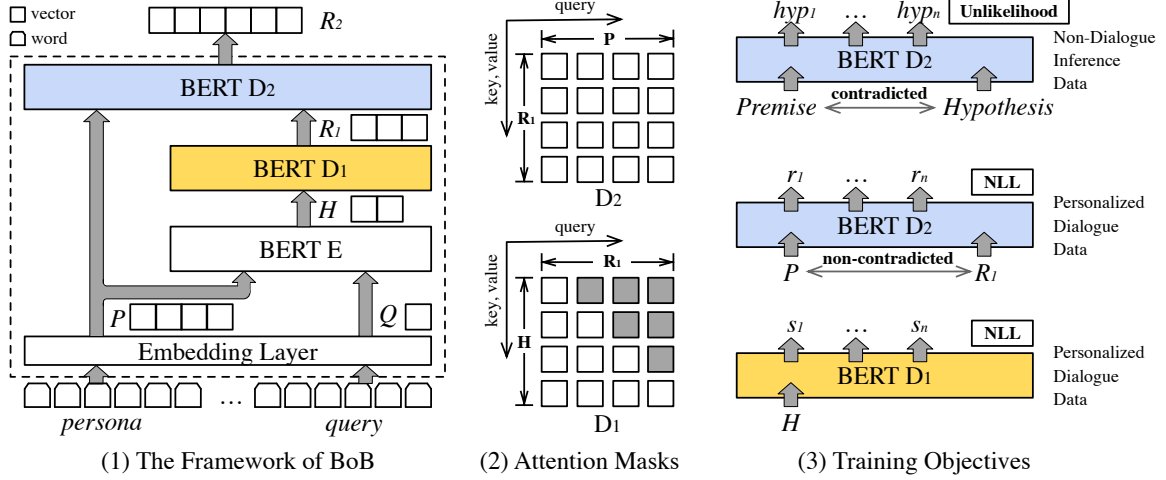
168

Figure 2: (1) The framework of the proposed BoB model, including an encoder (BERT $\mathbb{E}$), a response generation decoder (BERT $\mathbb{D}_1$), and a consistency understanding decoder (BERT $\mathbb{D}_2$). The italics denote the inputs and outputs of each submodule. (2) Transformer attention masks for generation ($\mathbb{D}_1$) and understanding ($\mathbb{D}_2$), and dark square means no attention. (3) Training objectives and the utilized data. NLL denotes negative log-likelihood.

3 Model

3.1 Overview

In this work, our goal is to learn a persona-based dialogue model from limited personalized data. To address the challenges of consistency understanding brought by limited data, we leverage large-scale non-dialogue inference data in our model.

Formally, let $\mathcal{Q} = q_1, q_2, ..., q_n$ denote the dialogue query, $\mathcal{R} = r_1, r_2, ..., r_m$ denote the target response, and $\mathcal{P}$ denote the personas. In addition, let $\mathcal{N}$ denote the non-dialogue inference data, which consists of premise, hypothesis, and their label. The premise and hypothesis are both natural sentences. Note that in the following sections, we use fonts to distinguish between sentences ($\mathcal{P}$, $\mathcal{Q}$, $\mathcal{R}$) and their vector representations (P, Q, R_1, R_2).

The task of the proposed model $\mathbb{M}$ is to generate a persona consistent response $\hat{\mathcal{R}} = \hat{r}_1, \hat{r}_2, ..., \hat{r}_m$, based on both persona $\mathcal{P}$ and query $\mathcal{Q}$, i.e., $\hat{\mathcal{R}} = \mathbb{M}(\mathcal{Q}, \mathcal{P})$. As shown in Figure 2, the proposed model $\mathbb{M}$ consists of three BERT-based submodules: an encoder $\mathbb{E}$, a response decoder $\mathbb{D}_1$, and a consistency understanding decoder $\mathbb{D}_2$. More concretely, $\mathbb{E}$ encodes the embeddings of persona and query, i.e., P and Q, into hidden states H. $\mathbb{D}_1$ performs cross-attention on H in a typical encoder-decoder manner, and generate a coarse representation R_1. $\mathbb{D}_2$ learns consistency understanding from non-dialogue inference data $\mathcal{N}$ and further converts P and R_1 into final representations R_2. At last, a consistent response $\hat{\mathcal{R}}$ could be generated from R_2.

3.2 Disentangling

For response generation, a typical persona-based dialogue model needs persona $\mathcal{P}$ and dialogue query $\mathcal{Q}$ to generate a response. For consistency understanding, a model needs persona $\mathcal{P}$, response $\mathcal{R}$, and the consistency labels between $\mathcal{P}$ and $\mathcal{R}$. However, if we entangle generation and understanding, it is not easy to obtain sufficient annotated data that satisfy the format of $\{\mathcal{P}, \mathcal{Q}, \mathcal{R}, \text{Label}\}$.

Instead, in our model, we design the decoder $\mathbb{D}_2$ to disentangle generation and understanding, where $\mathbb{D}_2$ maps R_1, rather than Q, to R_2. The key to "disentangling" is we can get R_1 without the participation of Q, as R_1 is the representation of $\mathcal{R}$. As a result, the mapping from R_1 to R_2 could be independent of Q. In this way, it becomes possible to 1) learn persona-based dialogue generation from $\{\mathcal{P}, \mathcal{Q}, \mathcal{R}\}$, i.e., the personalized data, and 2) learn consistency understanding from $\{\mathcal{P}, \mathcal{R}, \text{Label}\}$. Moreover, considering the limited amount of such annotated data, we could approximate $\{\mathcal{P}, \mathcal{R}, \text{Label}\}$ by the abundant non-dialogue inference data $\mathcal{N}=\{\text{Premise}, \text{Hypothesis}, \text{Label}\}$, where $\mathcal{P}$ and $\mathcal{R}$ corresponds to the Premise and Hypothesis.

Given data $\mathcal{P}$ and $\mathcal{R}$, suppose $\mathbb{D}_2$ understands persona consistency, it should maximize the likelihood of generating $\mathcal{R}$ if $\mathcal{R}$ is not contradicted to $\mathcal{P}$. Otherwise, it should minimize the likelihood of generating $\mathcal{R}$. Motivated by this observation, we choose to apply unlikelihood training on $\mathbb{D}_2$ to make it understand consistency. The detailed training objectives will be provided in Sec 3.4.

3.3 BERT-over-BERT

3.3.1 Encoder

The encoder $\mathbb{E}$ works like a standard BERT model, which bidirectionally encodes the input embeddings to a sequence of hidden vectors, from which the downstream tasks will be performed on.

In our model, the input consists of persona $\mathcal{P}$ and dialogue query $\mathcal{Q}$. For persona, whether $\mathcal{P}$ is personal facts (e.g., "I have two dogs") or profiles (e.g., "location: Seattle"), we could always convert it into a sequence of words. A special token is placed between persona sequence and dialogue query, and the input is formated as:

$$input = p_1^{(0)}, p_2^{(0)}, ..., p_{u_t}^{(t)}, [s], q_1, q_2, ..., q_n \quad (1)$$

Then the embedding layer will convert $input$ into representations. Following usual practice, the input representations are the sum of the corresponding token, type, and position embeddings, where the type embedding is 0 and 1 for persona and query, respectively. $\mathcal{P}$ and $\mathcal{Q}$ can also get their independent representations. The resulted representations are P and Q, which could be jointly denoted as emb = $e_1^p, e_2^p, ..., e_l^q$, where l is the maximum length of the $input$.

Once we get the input representations, encoder $\mathbb{E}$ will perform multi-head attetnion (Vaswani et al., 2017) on the emb to transform the embeddings into a sequence of hidden vectors H. The multi-head attetnion could be denoted as MultiHead(query, key, value), where scaled dot-product attention is performed on query, key, and value. There are N identical layers in $\mathbb{E}$, for each layer:

$$h^{i+1} = \text{FNN}(\text{MultiHead}(h^i, h^i, h^i)), \quad (2)$$

where h^0 = emb, and FNN is a fully connected feed-forward network containing two linear transformations with a ReLU activation in between. h^N is the final output of encoder $\mathbb{E}$, i.e., H.

3.3.2 Response Generation Decoder

The response generation decoder $\mathbb{D}_1$ is initialized from BERT to inherit its robust language model but works in an auto-regressive decoder manner. First, a cross-attention is inserted between $\mathbb{E}$ and $\mathbb{D}_1$ to pass the context information. Second, a left-to-right mask is applied to $\mathbb{D}_1$ to preserve the auto-regressive generation property.

As the cross-attention does not exist in the BERT model, it is randomly initialized and updated during training. In the cross-attention, the query comes from the previous layer of $\mathbb{D}_1$, and the key and value come from H:

$$r_1^{i+1} = \text{FNN}(\text{MultiHead}(r_1^i, H, H)). \quad (3)$$

This attention is similar to the typical encoder-decoder attention mechanism in sequence to sequence models (Bahdanau et al., 2015), which attends to all positions in the context representations H according to the variations of r_1. In training, r_1^0 is initialized from the embeddings of the target response. At each generation step, future tokens in the target response should not be considered. Therefore, as shown in Figure 2, a left-to-right mask is applied to $\mathbb{D}_1$ to ensure that the predictions can only depend on the known outputs.

$\mathbb{D}_1$ also has N identical layers. And the output of the last layer r_1^N, i.e., R_1, is further fed to $\mathbb{D}_2$.

3.3.3 Consistency Understanding Decoder

Like $\mathbb{E}$ and $\mathbb{D}_1$, the consistency understanding decoder $\mathbb{D}_2$ is also initialized from BERT, from where $\mathbb{D}_2$ initializes a good semantic representation for understanding tasks.

In each layer of $\mathbb{D}_2$, the multi-head attention is performed twice:

$$p^{i+1} = \text{FNN}(\text{MultiHead}(r_2^i, P, P)), \quad (4)$$
$$r_2^{i+1} = \text{FNN}(\text{MultiHead}(p^{i+1}, R_1, R_1)). \quad (5)$$

The resulted r_2^{i+1} in each layer thus fuses information from both P and R_1. The output of the last layer of $\mathbb{D}_2$ is the final representations R_2. With an output layer, e.g. linear layers, upon the R_2, we can get the generated response $\hat{\mathcal{R}}$.

3.4 Training Objectives

We employ negative log-likelihood (NLL) loss and unlikelihood loss for dialogue generation and consistency understanding. A brief illustration is shown in the last column of Figure 2 and detailed descriptions will be provided in this section.

Response Generation In our model, the widely adopted negative log-likelihood loss is applied in the training. For $\mathbb{E}$ and $\mathbb{D}_1$, they read the persona $\mathcal{P}$ and dialogue query $\mathcal{Q}$ to predict the target response $\mathcal{R}$, which yields the raw representations R_1:

$$\begin{aligned}
\mathcal{L}_{NLL}^{\mathbb{D}_1} &= -log(p_\theta(\mathcal{R}|\mathcal{P}, \mathcal{Q})) \\
&= -\sum_{i=1}^{|\mathcal{R}|} log(p_\theta(r_i|\mathcal{P}, \mathcal{Q}, \mathcal{R}_{<i})).
\end{aligned} \quad (6)$$

The generation part in $\mathbb{D}_2$ is also trained by NLL. $\mathbb{D}_2$ reads persona embeddings P and raw representations R_1 to predict the target response $\mathcal{R}$:

$$\begin{aligned}
\mathcal{L}_{NLL}^{\mathbb{D}_2} &= -log(p_\gamma(\mathcal{R}|P, R_1)) \\
&= -\sum_{i=1}^{|\mathcal{R}|} log(p_\gamma(r_i|P, R_1, \mathcal{R}_{<i})).
\end{aligned} \quad (7)$$

Unlikelihood Training Given large-scale non-dialogue inference dataset, we collect positive data $\mathcal{D}^+$ from the entailed category and collect negative data $\mathcal{D}^-$ from the contradicted category:

$$\mathcal{D}^+ = \{(\bar{\mathcal{P}}^{(i)}, \bar{\mathcal{R}}^{(i)+})\}, \quad D^- = \{(\bar{\mathcal{P}}^{(j)}, \bar{\mathcal{R}}^{(j)-})\}, \quad (8)$$

where $\bar{\mathcal{P}}$ and $\bar{\mathcal{R}}$ are premise and hypothesis from the non-dialogue inference data, and their representations in our model are denoted as $\bar{P}$ and $\bar{R}$. For data from $\mathcal{D}^+$, we still apply the NLL loss:

$$\mathcal{L}_{UL}^{\mathbb{D}_2^+} = -\sum_{i=1}^{|\bar{\mathcal{R}}|} log(p_\gamma(\bar{r}_i|\bar{P}, \bar{R}, \bar{\mathcal{R}}_{<i})), \quad (9)$$

For data from $\mathcal{D}^-$, we apply the unlikelihood objective to minimize the likelihood of contradictions:

$$\mathcal{L}_{UL}^{\mathbb{D}_2^-} = -\sum_{i=1}^{|\bar{\mathcal{R}}|} log(1 - p_\gamma(\bar{r}_i|\bar{P}, \bar{R}, \bar{\mathcal{R}}_{<i})), \quad (10)$$

which penalizes every token in the contradicted target. Therefore, the loss $\mathcal{L}_{UL}^{\mathbb{D}_2^-}$ makes generating contradicted responses less likely.

Training Procedure The training steps can be summarized as follows:

1) Response Generation. Given $\mathcal{P}$, $\mathcal{Q}$, and $\mathcal{R}$ from personalized dialogue data, we calculate the response generation loss $\mathcal{L}_1 = \mathcal{L}_{NLL}^{\mathbb{D}_1} + \alpha \mathcal{L}_{NLL}^{\mathbb{D}_2}$;

2) Consistency Understanding. Given $\mathcal{D}^+$ and $\mathcal{D}^-$ from non-dialogue inference data, we calculate the unlikelihood loss $\mathcal{L}_2 = \beta \mathcal{L}_{UL}^{\mathbb{D}_2^+} + (1-\beta)\mathcal{L}_{UL}^{\mathbb{D}_2^-}$;

3) Optimization. Sum up $\mathcal{L}_1$ and $\mathcal{L}_2$. Update parameters with back-propagation.

We initialize our model from the publicly available BERT base model, with 12 layers and hidden size 768. We employ an Adam optimizer with a learning rate of varying from 5e-6 to 5e-5. Empirically, we set α to 5e-3 and β to 0.1. The training of the proposed model was done on an Nvidia Telsa V100 32G GPU. Other details please refer to the released projects.

4 Experiments

4.1 Datasets

To evaluate the performance of the proposed model, we carried out persona-based dialogue generation experiments in a persona-dense scenario and a persona-sparse scenario with two publicly available datasets:

- **PersonaChat** (Zhang et al., 2018) is a crowd-sourced dataset covering rich persona features. The dialogues in this dataset are grounded on specific personal facts. Here we use the ConvAI2 PersonaChat (Dinan et al., 2019), so the results are comparable to existing methods.

- **PersonalDialog** (Zheng et al., 2019) is a large-scale persona-sparse dataset, which is collected from Chinese social media Weibo. This dataset provides persona profiles and dialogues, but the majority of the dialogues are not persona-related. Two testsets are provided: a random testset, which is identically distributed as the training data, and a biased testset, which is manually selected to cover persona-related features.

We summarize the key statistics of two personalized dialogue datasets in Tabel 1.

As aforementioned, we leverage non-dialogue inference data to address the consistency understanding issue brought by limited personalized data. Here we use the non-dialogue inference dataset MNLI (Williams et al., 2018) and its Chinese version CMNLI (Xu et al., 2020) as our auxiliary data. Moreover, to better compare models' performance on persona consistency, we leverage two dialogue inference datasets, DNLI (Welleck et al., 2019b) and KvPI (Song et al., 2020b), for evaluations. The statistics[1] of these inference datasets are summarized in Table 2.

4.2 Compared Methods

The following models, including both non-pretrained and pretrained ones, have been compared in the experiments.

Baselines. Vanilla **Transformer** (Vaswani et al., 2017) is employed as baselines for the experiments on both PersonaChat and PersonalDialog. Personas are concatenated to the dialogue queries.

---

[1]Note that for the DNLI, we only count the tuples that can be restored as {persona, query, response} in our experiments.

Dataset	# Train	# Valid	# Test
PersonaChat	121,880	9,558	7,801
PeronalDialog	5,014,349	423,817	10,000 / 521

Table 1: Statistics of persona-based dialogue datasets.

Dataset	# Entailed	# Neutral	# Contra.
MNLI	130,615	130,590	130,590
CMNLI	130,612	130,555	130,616
DNLI	15,495	20,927	16,488
KvPI	33,114	54,426	31,000

Table 2: Statistics of different inference datasets.

Non-Pretrained Models. Meta-learning has recently been explored in addressing the limited personalized data issue. **CMAML** (Song et al., 2020c) is a meta-learning based method that learns from few shot personas by customizing the model structures. Besides the meta-learning methods, **GDR** (Song et al., 2020a) introduces inference ability on the PersonaChat with a generate-refine framework. However, the two models are elaborately designed for the persona-dense dataset and not appliable for the persona-sparse scenario. Thus we only employ them for experiments on PersonaChat.

Pre-training Models. In the ConvAI2 challenge (Dinan et al., 2019), which utilizes PersonaChat as the competition dataset, **LIC** (Golovanov et al., 2019) is the best performing model. Thus we compare this model in the experiments on both PersonaChat and PersonalDialog. **AttentionRouting** (Zheng et al., 2020) is a pre-training method specially designed for the persona-sparse dataset, and it is also the latest model on PersonalDialog. We also finetune a **GPT2** (Radford et al., 2019) for a thorough comparison on PersonaChat.

4.3 Evaluation Metrics

We focus on two main aspects of the persona-based dialogues: **response quality** and **persona consistency**. To compare different models, we employ both automatic metrics and human evaluations.

Automatic Metrics For dialogue quality, we employ perplexity (**PPL.**) and distinct 1/2 (**Dist.1/2**) following common practice (Zhang et al., 2018; Zheng et al., 2020). Lower perplexity means better language modeling. Distinct 1/2 (Li et al., 2016a) are the ratio of distinct uni-grams / bi-grams, and higher distinct means better reponse diversity.

For persona consistency, we employ two metrics. The first is Consistency Score (**C.Score**) (Madotto et al., 2019), which leverages a referee model to predict consistency and can be defined as:

$$\text{NLI}(r, p_i) = \begin{cases} -1, & \text{if } r \text{ contradicts } p_i, \\ 0, & \text{if } r \text{ is irrelevant to } p_i, \\ 1, & \text{if } r \text{ entails } p_i. \end{cases}$$

$$\text{C.Score}(r) = \sum_{i=1}^{t} \text{NLI}(r, p_i).$$

$$(11)$$

Here the NLI is a pre-trained RoBERTa model (Liu et al., 2019) finetuned with the dialogue inference datasets, i.e., DNLI and KvPI, as described in Table 2. The RoBERT model achieves testset accuracy of 89.3% and 88.9% on DNLI and KvPI, which is aligned to the reported 88.20% (Welleck et al., 2019b) and 88.0% (Song et al., 2020b).

The second metric is Delta Perplexity (**ΔP**), which evaluates consistency from model's internal distributions. Li et al. (2020) first calculates the perplexity of entailed (**p.Ent**) and contradicted (**p.Ctd**) dialogues in the inference dataset. A dialogue model with good understanding ability should assign lower perplexity to the entailed dialogues while higher perplexity to the contradictions. From this intuition, the ΔP can be defined as:

$$\Delta\text{P} = \text{PPL}(\text{Contradicted}) - \text{PPL}(\text{Entailed}),$$

$$(12)$$

where a larger ΔP means the model has a better ability to distinguish entailment from contradiction. In our experiments, we get entailed and contradicted {persona, query, response} tuples from the dialogue inference datasets DNLI and KvPI.

Human Evaluations We recruit two teams (one for English and another for Chinese), each consists of five professional annotators, from a third-party company. These annotators are proficient in language tasks but know nothing about the models. We sample 100 {persona, query, response} tuples for each model's evaluation under every setting.

Human annotators are asked to evaluate dialogue quality from three conventional criteria: fluency (**Flue.**), informativeness (**Info.**), and relevance (**Relv.**). Each criterion is rated on a five-scale, where 1, 3, and 5 indicate unacceptable, moderate, and perfect performance, respectively. The annotators are also instructed to label the consistency (**Per.C.**) between persona and response, where 1 means persona-related and consistent, 0 means irrelevant, and -1 means contradicted.

	PPL	Dist.1	Dist.2	D.AVG	p.Ent	p.Ctd	ΔP	C.Score	Flue.	Info.	Relv.	Per.C.
Transformer	28.8	3.14	17.80	10.47	31.5	35.5	4.0	1.20	3.05	2.57	2.72	0.05
CMAML	36.7	1.00	2.10	1.55	32.3	37.5	5.2	6.96	3.36	2.40	3.09	0.24
GDR	16.7	3.76	23.10	13.43	19.7	32.3	12.6	7.89	3.38	2.74	3.13	0.21
LIC	17.3	6.29	28.99	17.64	13.7	20.4	6.7	14.12	3.70	3.53	3.47	0.39
GPT2	14.4	7.29	28.12	17.71	12.0	20.2	8.2	15.88	3.79	3.22	3.79	0.47
BoB (Ours)	**7.8**	**8.40**	**36.08**	**22.24**	**7.3**	**83.4**	**76.1**	**17.18**	**4.12**	**4.03**	**4.09**	**0.60**

Table 3: Automatic and human evaluation results on the full PersonaChat dataset. The best results are in bold.

	PPL	Dist.1	Dist.2	D.AVG	p.Ent	p.Ctd	ΔP	C.Score	Flue.	Info.	Relv.	Per.C.
Baselines' Best	14.4	7.29	28.99	17.71	12.0	37.5	12.6	15.88	3.79	3.53	3.79	0.47
Ours 1/8 Data	11.6†	7.49†	27.10	17.30	11.3†	83.6†	72.3†	15.87	4.17†	3.48	4.12†	0.62†
Ours 1/4 Data	9.7	7.97	30.20†	19.09†	11.8	85.8	74.0	16.04†	4.19	3.47	4.17	0.60
Ours 1/2 Data	8.9	8.13	33.08	20.61	8.1	81.9	73.8	16.36	4.03	3.70†	3.94	0.61

Table 4: Automatic and human evaluation results of the low resource settings on the PersonaChat dataset. The †
means the minimum amount of data our model needed to outperform baselines' best results.

4.4 Persona-Dense Results

Full PersonaChat We first report the full PersonaChat experimental results in Table 3. Our method achieves better performance consistently across all automatic and human evaluation metrics, which shows the effectiveness of our model.

Among all the metrics, our model obtains significant improvements on PPL and ΔP. The lowest testset PPL means our model has learned a good language model fitting this dataset. Moreover, the highest ΔP shows that our model could more effectively distinguish entailment from contradiction than other baselines, which indicates our model has a better understanding of persona consistency.

Less Personalized Data Now that our model achieves better performance with a large margin on the full PersonaChat dataset, we want to test our model by simulating a low-resource scenario (Zhao et al., 2019), where we gradually reduce the number of examples by halving the training set. We report the low-resource settings' results in Table 4.

As we can see, our model can outperform most of the baselines' best results even by using only 1/8 of the training data. The performance gains largely benefit from the powerful language model of the backbone BERT model. Furthermore, due to the disentangling of generation and understanding, our model presents a stable performance on ΔP regardless of the size of the training set. This is in line with our expectations because the proposed model learns consistency understanding from the

non-dialogue inference data rather than the persona-dense dialogue data. We observe that the method also improves fluency and informativeness. It is mainly due to the introduction of the non-dialogue inference data in the training procedure, which potentially enriches the dialogue language model.

4.5 Validations on Persona-Sparse

We further validate our model on a persona-sparse scenario. To have a more intuitive understanding of "sparsity", we recruit the same annotation team to annotate whether the dataset response is persona-related in the sampled random and biased test data. Results show that only 1% responses are persona-related in the random test data and 28% in the biased test data. We calculate the Fleiss' Kappa among the five annotators and obtain a kappa of 0.774, which means *substantial agreement* (Landis and Koch, 1977). We report the evaluation results on both random and biased testsets in Table 5.

On the random test set, experimental results demonstrate that our model has some advantages over other methods, but no method can consistently outperform the others. One possible reason is that the task has degenerated into the ordinary dialogue generation in the random test set, so our model's advantages can not be effectively leveraged. In contrast, on the biased test set, our model achieves the best performance on most metrics. The good performance on the metrics C.Score and Per.C. indicates that our model can be effectively trained from a dataset with limited personalized dialogues.

	Random Testset						Biased Testset						KvPI
	PPL	C.Score	Flue.	Info.	Relv.	Per.C.	PPL	C.Score	Flue.	Info.	Relv.	Per.C.	ΔP
Trans	43.7	0.95	3.26	2.38	2.72	0.00	83.2	1.04	3.54	2.58	2.84	0.03	3.28
LIC	47.8	**4.08**	3.68	2.66	2.92	**0.02**	43.3	8.25	3.72	3.01	3.04	0.08	2.86
AR	34.2	-2.14	3.71	2.58	**3.02**	-0.03	38.7	11.72	3.78	3.11	3.10	0.13	3.08
Ours	**18.5**	2.10	**3.75**	**2.69**	2.98	0.01	**19.5**	**12.76**	**3.84**	3.13	**3.17**	**0.15**	**85.40**
w/o UL	19.3	-3.13	3.73	2.57	2.93	-0.06	20.1	10.53	3.79	2.92	3.10	0.09	4.10
E+D1	31.7	0.15	3.74	2.68	2.96	-0.01	38.0	9.75	3.74	**3.15**	3.06	0.08	2.80
E	35.5	1.64	3.67	2.57	2.96	0.01	41.1	7.41	3.72	3.05	3.04	0.04	4.60

Table 5: Automatic and human evaluation results on the random testset and biased testset of PersonalDialog, along with the ablation results. Trans denotes Transformer, and AR denotes AttentionRouting. Best results in bold.

	PPL	ΔP	Flue.	Info.	Relv.	Per.C.
Ours	7.8	76.1	4.12	4.03	4.09	0.60
w/o UL	8.1	7.8	3.81	3.50	3.80	0.48
E+D$_1$	23.6	4.9	3.65	3.18	3.60	0.45
E	25.7	7.1	3.69	3.28	3.60	0.42

Table 6: Ablation results of automatic metrics and human evaluations with full PersonaChat dataset.

Persona Query	I've a son who is in junior high school
	You have any children?
GPT2 Ours	No kids. I work at home depot so I'm busy.
	Yes, I have a son in the 8th grade.

Table 7: A generated example from our model.

4.6 Analysis and Ablation Study

In addition to the good performance of the BoB model, we are also curious about **Q1**: what is the key to the BoB model's understanding ability? **Q2**: can the pre-trained models understand persona consistency just through finetuning on the personalized dialogues? And **Q3**: does the extremely low PPL come from the initialization of the BERT model or the architecture of the proposed BoB model?

To better answer the above questions, we ablate the BoB model in the following three ways: 1) **w/o UL**. It removes the unlikelihood objective. 2) **E+D$_1$**. It removes the unlikelihood objective and the second decoder $\mathbb{D}_2$. 3) **E**. It removes the unlikelihood objective and both decoders and thus degenerates into a vanilla BERT model. We report the ablation results on PersonalDialog in Table 5 and full PersonaChat in Table 6. From these results:

Answer to Q1: The key to our model's understanding is the unlikelihood training. In training, our model assigns large perplexity to the contradictions. In generating, the non-contradicted responses are more likely to be generated as they are with much smaller losses. Table 7 shows an example. And as presented in the results, after removing the unlikelihood objective, all ablated models suffer from significant performance degradations in consistency-related metrics, such as Per.C. and ΔP.

Answer to Q2: Pretrained models barely understand consistency from personalized dialogues. According to the poor performances on ΔP, the three BERT-based ablated models can hardly distinguish contradiction from entailment. Although their Per.C. metric still looks good, it may come from just mimicking and copying words rather than understanding. A similar phenomenon also occurs to the pre-trained GPT2, as shown in Table 3. It is also this phenomenon that motivates us to introduce the unlikelihood training into the BoB model.

Answer to Q3: $\mathbb{D}_2$ in the BoB architecture contributes most to the PPL. As shown in both datasets' ablation results, the PPL decreases the most after removing $\mathbb{D}_2$. We can also see an apparent gap between the models with $\mathbb{D}_2$ and the vanilla BERT on PPL. Nevertheless, the BERT model still offers a good initialization for the BoB model to achieve the best performance on different metrics.

4.7 Reproducibility

The implementation for the BoB model is released at https://github.com/songhaoyu/BoB.

5 Conclusions

In this work, we propose a novel BERT-based dialogue model to learn from limited personalized data by disentangling response generation and consistency understanding. Unlikelihood training with non-dialogue inference data is introduced to en-

hance the model's understanding ability. Experiments on two publicly available datasets demonstrate that our model can be trained with limited personalized dialogue data while still obtain significant improvements over strong methods.

Acknowledgments

Ethical Statement

Persona-based dialogue research intends to address the persona inconsistency issue in open-domain dialogue to facilitate human-computer interactions. Giving dialogue system a specific persona is a mainstream to alleviate the inconsistency issue of dialogues under the current stage. The purpose is to endow the dialogue system with self logical consistency rather than imitate specific human beings. Simultaneously, in this work, the data resources we use are all from published works and do not involve privacy issues related to data collection. We also confirm that this work neither automatically infers or attributes identity characteristics to the participants nor categorizes them in the training datasets.

References

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

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058, Florence, Italy. Association for Computational Linguistics.

Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo J. Rezende, and Alexander Lerchner. 2018. Towards a definition of disentangled representations. *CoRR*, abs/1812.02230.

Lee Humphreys, Phillipa Gill, and Balachander Krishnamurthy. 2014. Twitter: a content analysis of personal information. *Information, Communication & Society*, 17(7):843–857.

Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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

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

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4715–4728, Online. Association for Computational Linguistics.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020. Xpersona: Evaluating multilingual personalized chatbot. *arXiv preprint arXiv:2003.07568*.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. 2019. Disentangling factors of variations using few labels. In *International Conference on Learning Representations*.

Andrea Madotto, Zhaojiang Lin, Yejin Bang, and Pascale Fung. 2020. The adapter-bot: All-in-one controllable conversational model. *arXiv preprint arXiv:2008.12579*.

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*, pages 5454–5459, 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*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is it really about me? message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4279–4285. International Joint Conferences on Artificial Intelligence Organization.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu. 2020a. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831, Online. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Zhengyu Zhao, Ting Liu, and Xiaojiang Liu. 2020b. Profile consistency identification for open-domain dialogue agents. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6651–6662, Online. Association for Computational Linguistics.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5190–5196. International Joint Conferences on Artificial Intelligence Organization.

Yiping Song, Zequn Liu, Wei Bi, Rui Yan, and Ming Zhang. 2020c. Learning to customize model structures for few-shot dialogue generation tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5832–5841, Online. Association for Computational Linguistics.

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 *NIPS*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019a. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019b. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2019. Low-resource knowledge-grounded dialogue generation. In *International Conference on Learning Representations*.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *ArXiv*, abs/1901.09672.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9693–9700.