Dual Task Framework for Improving Persona-Grounded Dialogue Dataset

Minju Kim[*1], **Beong-woo Kwak**[*1], **Youngwook Kim**[1], **Hong-in Lee**[1]
Seung-won Hwang[2] **and Jinyoung Yeo**[†1]

[1]Yonsei University
[2]Seoul National University

Abstract

This paper introduces a simple yet effective data-centric approach for the task of improving persona-conditioned dialogue agents. Prior model-centric approaches unquestioningly depend on the raw crowdsourced benchmark datasets such as Persona-Chat. In contrast, we aim to fix annotation artifacts in benchmarking, which is orthogonally applicable to any dialogue model. Specifically, we augment relevant personas to improve dialogue dataset/agent, by leveraging the primal-dual structure of the two tasks, predicting dialogue responses and personas based on each other. Experiments on Persona-Chat show that our approach outperforms pre-trained LMs by an 11.7 point gain in terms of accuracy.

Introduction

In personalized dialogue agents, *persona grounding* has been a long-standing goal to improve both human-likeness and dialogue consistency. For example, when given a profile description such as "**`I am a doctor.`**" and "**`I don't eat meat`**", dialogue agents aim to plausibly respond to dialogue context while endowing the machine with the persona, *e.g.*, "*I am now working at hospital*" and "*I went vegan.*" respectively. To learn and evaluate such grounding, recent research (Kim, Kim, and Kim 2020; Li et al. 2020; Song et al. 2020; Zhang et al. 2019) has proposed to encode personas based on pre-trained language models (PLMs) (Devlin et al. 2019; Liu et al. 2019; Radford et al. 2019), and many persona-conditioned dialogue benchmarks have been released such as Persona-Chat (Zhang et al. 2018), where crowdworkers role-play following the given description of personas to populate dialogues.

In spite of such recent significant progress, we argue that there is much room for improving persona-grounded dialogue agents in the data-centric view. Specifically, as crowdworkers often miss out on stating detailed experience and knowledge, it is reported that several linguistic biases exist in the persona-grounded dialogue datasets, for example, crowdworkers tend to reuse some terms of persona descriptions to script dialogue utterances, *e.g.*, "*I am a doctor.*", in which words trivially overlap to a persona sentence, or merely script utterances as a paraphrase, *e.g.*, "*I work as a doctor.*" or "*I don't eat any meat.*".

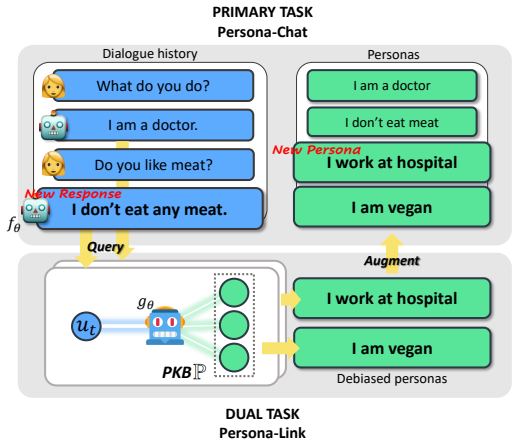


Figure 1: Illustration of the primal-dual task pipeline

Such annotation artifacts can be a bottleneck for learning to ground more engaging utterances and adversarially learning not to make contradictory utterances to personas. In practice, dialogue agents hardly respond in an engaging manner, *e.g.*, "*I am now working at hospital.*", loosening the linguistic ties from the given persona (Ghazarian et al. 2021). Also, on a challenging question such as "*Are you vegan?*" and "*Do you like steak?*", dialogue agents may violate consistency (Welleck et al. 2018), *e.g.*, responding "*No*" and "*Yes*" respectively. A straightforward way to mitigate these phenomena is to manually fix the dataset with additional annotation costs (Bowman and Dahl 2021), by asking another crowdworkers to rewrite/augment the dialogue utterances, but it is too expensive on a large scale.

As motivated above, in this work, we propose to automatically improve the persona-conditioned dialogue dataset to learn better persona-grounded dialogue agents. As illustrated in Figure 1, our key idea is to infer and add new personas from a dialogue, *e.g.*, "*I am a doctor.*" $\mapsto$ **`I work at hospital`** or "*I don't eat any meat.*" $\mapsto$ **`I am vegan`**. For that, we present a novel dual task framework, where the primal task Persona-Chat is learning persona-conditioned dialogue models (*i.e.*, predicting utterances based on persona),

[*]The authors contribute equally to this paper.
[†]Corresponding author (Email: jinyeo@yonsei.ac.kr)

whose personas are augmented by the dual yet secondary task, namely Persona-Link (*i.e.*, predicting personas for utterances reversely). Our framework first learns the Persona-Link models, which use dialogue contexts in Persona-Chat as a query to augment its relevant personas, then learns the Persona-Chat models with the augmented personas for better persona grounding.

Straightforwardly, reversing the Persona-Chat dataset can provide the utterance-to-persona alignments as linking supervisions to train the Persona-Link models. However, using naive closed alignments from individual dialogue episodes may inherit the linguistic bias from Persona-Chat, which disqualifies the ultimate role of debiasing Persona-Chat. To tackle this challenge, we first adapt ideas from semantic matching (Wu et al. 2020) to our dual task, which enables augmenting open alignments out of individual dialogue episodes, *i.e.*, less linguistically-biased alignments from the whole Persona-Chat corpus. Then, furthermore, we also leverage commonsense expansion (Majumder et al. 2020), which enables augmenting more open alignments out of the Persona-Chat corpus itself, in a systematic way.

Our main contributions are summarized as follows: (1) We propose an iterative framework of primal-dual tasks to debias the Persona-Chat dataset/model without any human effort. (2) We automate the learning of Persona-Link models from the Persona-Chat dataset, as a desired form of debiasing Persona-Chat. (3) Our extensive experiments validate that, along with linked personas, the response accuracy significantly increases by 11.7% point on Persona-Chat compared to that of using the raw dataset.

Primal-Dual Task Framework

Primal Task: Persona-Chat

The goal of the Persona-Chat task (Zhang et al. 2018) is to personalize dialogue agents by grounding their utterances to the given personas. The original dataset involves dialogues between pairs of speakers: each speaker is given a hypothetical profile, which is a few persona sentences that describe a character they will imitate, and is instructed to get to know the other. Formally, a profile is defined as a set of persona sentences $\mathcal{P} = \{p_1, ..., p_m\}$ and a dialogue is defined as a set of multi-turn utterances $\mathcal{U} = \{u_1, ..., u_n\}$, and a Persona-Chat dataset $\mathcal{D}_{\text{CHAT}} = \{(\mathcal{P}_i, \mathcal{U}_i)\}_{i=1}^{N}$ where N is the number of dialogues. Based on $\mathcal{D}_{\text{CHAT}}$, when given persona sentences $\mathcal{P}$, dialogue history $\mathcal{U}$, and response space $\mathbb{U}$, the primal task aims at finding a model $f : (\mathcal{P}, \mathcal{U}) \mapsto \mathbb{U}$.

$$f(\mathcal{P}, \mathcal{U}; \theta_{\text{CHAT}}) \triangleq \underset{u \in \mathbb{U}}{\text{argmax}}\ P(u|\mathcal{P}, \mathcal{U}; \theta_{\text{CHAT}}) \tag{1}$$

where θ_{CHAT} is the parameters to be learned for model f_θ. In our setting, we adopt the response selection task (Humeau et al. 2020). As response space $\mathbb{U}$, the model has to pick the correct response from a set of 20 choices, where the remaining 19 were randomly chosen utterances from the evaluation set. Note that in a final system, however, one would retrieve from the entire training set of over 100k utterances, but this is avoided for speed reasons in common evaluation setups (Wolf et al. 2019; Humeau et al. 2020).

Dual Task: Persona-Link

Following the primal-dual structure, the Persona-Link task can be defined as: given an arbitrary dialogue utterance u, the output of a linking model g is its referent persona description in persona space $\mathbb{P}$. That is, the dual task aims at finding a linking model $g : u \mapsto \mathbb{P}$.

$$g(u; \theta_{\text{LINK}}) \triangleq \underset{p \in \mathbb{P}}{\text{argmax}}\ P(p|u; \theta_{\text{LINK}}) \tag{2}$$

where θ_{LINK} is the parameters to be learned for model g_θ. We formalize the dual task as a variant of entity linking system, where an utterance is linked to an entry in the PKB $\mathbb{P}$ of arbitrary size, being populated from the primal dataset $\mathcal{D}_{\text{CHAT}}$, *i.e.*, $\mathbb{P} = \bigcup_{i=1}^{N} \mathcal{P}_i$. To learn model g_θ, we adopt the same neural architecture of the primal task (*i.e.*, Bi-encoder) (Humeau et al. 2020) as a base model with the cross-entropy loss.[1]

Algorithm 1: Primal-Dual Task Framework

Input: Original data $\mathcal{D}_{\text{CHAT}}$
Output: Debiased data $\tilde{\mathcal{D}}_{\text{CHAT}}$, Debiased model $\tilde{\theta}_{\text{CHAT}}$
$\mathcal{D}_{\text{LINK}}, \tilde{\mathcal{D}}_{\text{LINK}} \leftarrow \text{ReverseDataset}(\mathcal{D}_{\text{CHAT}})$ } Persona-Link
$\theta_{\text{LINK}} \leftarrow \text{Train}(g_\theta, \mathcal{D}_{\text{LINK}})$ } Persona-Link
$\tilde{\theta}_{\text{LINK}} \leftarrow \text{Train}(g_\theta, \tilde{\mathcal{D}}_{\text{LINK}}, \theta_{\text{LINK}})$ } Persona-Link
$\tilde{\mathcal{D}}_{\text{CHAT}} \leftarrow \text{AugmentPersona}(\mathcal{D}_{\text{CHAT}}, \tilde{\theta}_{\text{LINK}})$
$\tilde{\theta}_{\text{CHAT}} \leftarrow \text{Train}(f_\theta, \tilde{\mathcal{D}}_{\text{CHAT}})$ } Debiased Persona-Chat
return $\tilde{\mathcal{D}}_{\text{CHAT}}, \tilde{\theta}_{\text{CHAT}}$

Primal-Dual Task Pipeline

We introduce a new framework that exploits the duality (Xia et al. 2017) of Persona-Chat and Persona-Link tasks where the input (utterance) and output (persona) of the dual task are roughly the inverse of its primal task.

In the training phase of the primal task, our ultimate goal is to transform the original dialogue dataset $\mathcal{D}_{\text{CHAT}}$ into the debiased dataset $\tilde{\mathcal{D}}_{\text{CHAT}}$ by augmenting new personas, then train the debiased dialogue model f_θ from $\tilde{\mathcal{D}}_{\text{CHAT}}$, which reduces the model dependence on linguistic bias between personas and utterances. Algorithm 1 describes its overall procedure. In the proposed framework, the linking model g_θ should be learned in advance to augment plausible personas per a given utterance. However, the key challenge is collecting the supervisory information as the training data of g_θ (denoted as $\mathcal{D}_{\text{LINK}}$ or $\tilde{\mathcal{D}}_{\text{LINK}}$), which is capable of injecting the debiasing ability to g_θ. As illustrated in Figure 2, we mainly discuss this in the subsequent section.

Once an arbitrary linking model g_θ is obtained as the desired debiasing function, we leverage g_θ also in the inference phase of the primal task, not only in training time. Specifically, as illustrated in Figure 1, a new persona is interactively augmented per new response to ground the dialogue agents to debiased persona-response alignments. This procedure is repeated until the dialogue ends.

[1]For more flexible application, inspired by (Wu et al. 2020), we relax the collective linking in Eq. (2) into finer-level linking computing the utterance-to-persona score, *i.e.*, $p(p_i|u_i; \theta_{\text{LINK}})$.

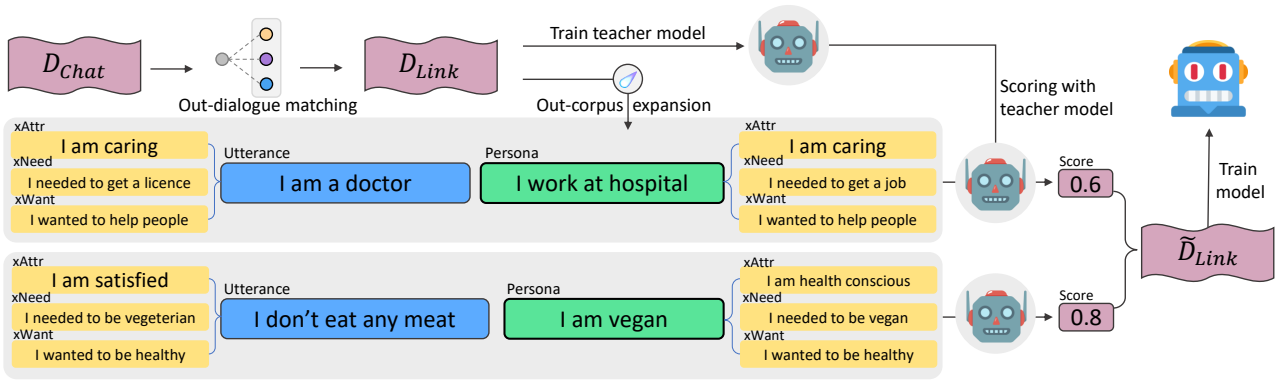


Figure 2: Overall procedure of learning linking models at semantic and commonsense level for Persona-Link

Reversing Persona-Chat to Persona-Link

A straightforward way to learn the dual task, Persona-Link, is to reverse input and output sides of the primal task each other (Wang et al. 2021; Su, Chuang, and Chen 2020; Zhu, Cao, and Yu 2020). However, in contrast to prior work, such a naive approach disqualifies the ultimate role of debiasing Persona-Chat since the Persona-Link data and model can inherit even the linguistic biases presented in the Persona-Chat dataset. Thus, the Persona-Link data should involve the desired asymmetric characteristics with Persona-Chat. We now present the three phases for learning Persona-Link models.

Phase 1: Out-dialogue Semantic Matching

Basically, as supervisory data for Persona-Link, we can collect the plausible alignments of persona and utterances in individual dialogue episodes of the Persona-Chat dataset, by using their semantic relationship. Specifically, we leverage natural language inference (NLI), a task of determining whether a hypothesis (*e.g.*, persona) can be inferred from the given premise (*e.g.*, utterance). The hypothesis sentence is classified into three categories: `Entailment` (true), `Contradiction` (false), and `Neutral` (undetermined). We adopt a RoBERTa model (Liu et al. 2019) trained on MNLI (Wang et al. 2019). By NLI, the minimal sampling to populate $\mathcal{D}_{\text{LINK}}$ is to consider the *in-dialogue* matching of persona-utterance pair (u, p) as candidates, where u and p are derived from the same dialogue episode. However, as mentioned earlier, such (u, p) pairs in individual dialogue episodes reportedly suffer from trivial word overlaps.

In contrast, we perform the *out-dialogue* matching of all possible combinations in $\mathcal{D}_{Chat}$, which is stochastically less overlapped at lexical level. That is, such a simple strategy can contribute to mitigating the linguistic bias in the NLI-driven alignments, which we call seed Persona-Link dataset.

Definition 1 *(Seed Persona-Link Dataset) Given* $\mathcal{D}_{\text{CHAT}} = \{(\mathcal{P}_j, \mathcal{U}_j)\}_{j=1}^{N}$*, a seed Persona-Link dataset can be defined as* $\mathcal{D}_{\text{LINK}} = \{(u_i, p_i, y_i)\}_{i=1}^{M}$ *where* u_i *and* p_i *are drawn from* $\mathcal{D}_{\text{CHAT}}$*, i.e.,* $u_i \in \bigcup_{j=1}^{N} \mathcal{U}_j$ *and* $p_i \in \bigcup_{j=1}^{N} \mathcal{P}_j$*, and binary label* y_i *is captured by a NLI classifier. If* (u_i, p_i) *is inferred as* `Entailment`*,* $y_i = 1$*, and* $y_i = 0$ *otherwise.*

Phase 2: Out-corpus Commonsense Expansion

Linguistic bias comes from the limited knowledge and experience of individual crowdworkers, which motivates the out-dialogue matching of personas and annotated utterances. Beyond the Persona-Chat corpus, if someone would have world-level knowledge, she may willingly annotate more engaging, less linguistically-biased utterances for a given persona. For example, utterance "*I don't eat any meat.*" can involve commonsense attributes '*I needed to be vegeterian*' or '*I wanted to be healthy*', which enable drawing a new persona "`I am vegan`" from PKB, which are not captured by semantic matching (*i.e.*, predicted as Neutral).

If a linking model is trained on sentence pairs of only similar semantics, such a potentially relevant mapping cannot be captured for generalization in inference time. Thus, since personas and utterances are instances of world events that often imply real-world consequences or richer information, we propose to exploit such commonsense attributes as "anchors". It may ensure the learned representations are well associated via reasoning effects (Liu et al. 2020; Majumder et al. 2020) beyond the semantically close alignments.

Specifically, we populate an augmented dataset $\tilde{\mathcal{D}}_{Link}$ that expands pairs in $\mathcal{D}_{Link}$ based on commonsense knowledge. For commonsense expansion, we capture implicit attributes of either personas and utterances and annotate them as metadata, using GPT2 based commonsense knowledge generators (Hwang et al. 2021) (Appendix A). The commonsense attributes are surrounded by special tokens and concatenated into a single sequence with persona or utterance.

Definition 2 *(Commonsense Persona-Link Dataset) Given* $\mathcal{D}_{\text{LINK}} = \{(u_i, p_i, y_i)\}_{i=1}^{M}$*, its commonsense-expanded set is defined as* $\tilde{\mathcal{D}}_{\text{LINK}} = \{(\tilde{u}_i, \tilde{p}_i, y_i)\}_{i=1}^{M}$ *where expanded samples* $\tilde{u}_i$ *and* $\tilde{p}_i$ *are obtained by a commonsense expansion function* $\psi : (u_i \text{ or } p_i) \rightarrow (\tilde{u}_i \text{ or } \tilde{p}_i)$*. This function provides tuples that belong to nine relation types spanning over cause-effect interrelations between events:* `xAttr`*,* `xEffect`*,* `xIntent`*,* `xNeed`*,* `xReact`*,* `xWant`*,* `oEffect`*,* `oReact`*,* `oWant`*— where a prefix 'x' indicates an effect or cause on the person and 'o' denotes the same on others. We present more details of the commonsense expansion function in Appendix A.*

Dialogue model	Linking model	R@1/20 ↑	R@5/20 ↑	MRR ↑	Contradict@1 ↓
Bi-encoder	N/A	0.814	0.973	0.882	0.075
Cross-encoder		0.884	0.991	0.930	0.040
Bi-encoder	Manual Paraphrasing	0.758	0.961	0.876	0.088
	Paraphrasing	0.769	0.963	0.851	0.094
	Bi-encoder	0.881	0.987	0.927	0.053
	Cross-encoder	0.881	0.988	0.928	0.043
Bi-encoder	Persona-Link (small PKB)	0.904	0.992	0.943	0.039
	Persona-Link (large PKB)	**0.931**	**0.993**	**0.959**	**0.027**

Table 1: The performance of dialogue models on Persona-Chat testset

Phase 3: Learning with Label Regularization

As many commonsense attributes are commonly shared between utterance and persona, they are likely to be relevant to each other although the NLI-driven label indicates negative. On the other hand, not always such an assumption is true since generated commonsense attributes are often ambiguous or over-claimed. These concerns motivate us to better regularize learning on $\tilde{\mathcal{D}}_{\text{LINK}}$ by well-calibrated soft labels. That is, the expanded set $\tilde{\mathcal{D}}_{\text{LINK}}$ on the denser space may not strictly follow the pre-annotated binary labels from NLI.

To address this, we argue that the linking model trained on $\mathcal{D}_{\text{LINK}}$ can be a good reference point. Specifically, we first train a linking model θ_{LINK} only with the semantic-level Persona-Link dataset $\mathcal{D}_{\text{LINK}}$, which can further perform inference over new data samples $(\tilde{u}, \tilde{p})$ to compute their outputs as new labels $P(\tilde{p}|\tilde{u}; \theta_{\text{LINK}})$. By regularizing by the semantic feature space of θ_{LINK}, the commonsense-expanded dataset can have well-calibrated labels for encoding commonsense attributes. We thus train another linking model $\tilde{\theta}_{\text{LINK}}$ with dual goals of following not only the original hard labels but also new soft labels as:

$$\theta_{\text{LINK}} = \text{argmin}_\theta \sum_{(u,p,y)\in\mathcal{D}_{\text{LINK}}} \mathcal{L}(y, P(p|u;\theta)) \quad (3)$$

$$\tilde{\theta}_{\text{LINK}} = \text{argmin}_\theta \sum_{(\tilde{u},\tilde{p},y)\in\tilde{\mathcal{D}}_{\text{LINK}}} \mathcal{L}(y, P(\tilde{p}|\tilde{u};\theta)) + \lambda \cdot \mathcal{L}(P(\tilde{p}|\tilde{u};\theta_{\text{LINK}}), P(\tilde{p}|\tilde{u};\theta)) \quad (4)$$

where λ is a preference weight with the distillation loss. To compute the linking score $P(p|u;\theta)$ and the cross-entropy loss $\mathcal{L}$, based on the Bi-encoder architecture allowing for fast and real-time inference, we follow the optimization procedure (Logeswaran et al. 2019; Humeau et al. 2020; Jeong et al. 2021) of retrieval-based models (*e.g.*, information retrieval, entity linking, and response selection): The network is trained to maximize the score of the correct persona p with respect to the (randomly sampled) personas of the same batch (*i.e.*, *in-batch negatives*).

Inference Once a linking model $\tilde{\theta}_{\text{LINK}}$ is learned, given a utterance u in Persona-Chat, we follow Eq. (2) to augment a new persona into the original dialogue episode. Considering that the development/test sets of Persona-Chat are unseen, as PKB $\mathbb{P}$, we only use a list of personas involved in the training set of Persona-Chat at least once.

PERSONA-CHAT Evaluation

As bias analysis in benchmark datasets is a non-trivial problem (Bowman and Dahl 2021; Torralba and Efros 2011), we measure the response quality of dialogue models as a proxy of bias mitigation. We measure Recall@k/N and MRR as the model performance to the gold utterance, where $N = 20$ and $k = [1, 5]$. Another metric is Contradict@1, indicating the textual disagreement judged by an NLI model: the proportion of contradictory responses in the top-1 candidates returned by dialogue agents, *i.e.*, consistency error ratio.

Performance with White-box Training

For the experiment on white-box settings (*i.e.*, augmenting personas in both training and inference time), we consider following six models including two Persona-Link variants as the candidates to be analyzed.

1) **Paraphrasing**: As unsupervised data augmentation (Xie et al. 2020), we consider an off-the-shelf paraphrasing system in (Mallinson, Sennrich, and Lapata 2017), where personas were translated into a foreign language and back-translated as paraphrases.

2) **Manual Paraphrasing**: As labor-intensive augmentation, we use manually revised persona sentences additionally presented in Persona-Chat, where extra workers rephrased them to remove trivial word overlap.

3) **Bi-encoder**: As an IR baseline of linking model (Wu et al. 2020) trained on $\mathcal{D}_{\text{LINK}}$, we use two transformers that separately encode utterance and persona, and multi-sentence scoring over the pre-computed cache of personas.

4) **Cross-encoder**: We consider using a single transformer (Humeau et al. 2020) that uses a richer self-attention mechanism, jointly encoding utterance and persona in $\mathcal{D}_{\text{LINK}}$ to obtain the final representation, which is impractical for real-time use.

5) **Persona-Link (small PKB)**: Our proposed linking model with a smaller Persona Knowledge Base (PKB), caching 1K personas which are randomly sampled from the primal dataset $\mathcal{D}_{\text{CHAT}}$.

6) **Persona-Link (large PKB)**: Using the same model, we present the original version of our approach, where all 5K personas from $\mathcal{D}_{\text{CHAT}}$ are plugged into its PKB.

Table 1 shows the overall performance of dialogue models trained with different linking models. Models paired with

Linking model	Dialogue model: Pretrained on Wikipedia				Pretrained on Reddit			
	R@1/20↑	R@5/20 ↑	MRR ↑	Ctrd.@1 ↓	R@1/20↑	R@5/20 ↑	MRR ↑	Ctrd.@1 ↓
No persona (0%)								
N/A	0.591	0.870	0.711	0.174	0.659	0.900	0.766	0.144
Bi-encoder	0.699	0.936	0.802	0.120	0.773	0.960	0.853	0.094
Persona-Link	**0.733**	**0.956**	**0.828**	**0.113**	**0.805**	**0.970**	**0.876**	**0.078**
Incomplete persona (80%)								
N/A	0.768	0.953	0.848	0.094	0.825	0.973	0.889	0.073
Bi-encoder	0.768	0.966	0.850	0.094	0.845	0.984	0.905	0.064
Persona-Link	**0.785**	**0.972**	**0.864**	**0.093**	**0.858**	**0.985**	**0.913**	**0.058**
Original persona (100%)								
N/A	0.814	0.973	0.882	0.075	0.868	0.986	0.919	0.054

Table 2: The response quality of Persona-Chat models, observing debiased personas added at test time. In advance, in-dialogue persona sentences were randomly removed, and link models augmented in-context personas based on dialogue histories.

Persona-Link variants significantly outperform other baselines for all measures. This suggests that training dialogue model with well augmented personas helps the model to be more persona-grounded and consistent which is a primary goal in persona-based dialogue literature. Furthermore, as the performance gain increases with the size of PKB used for Persona-Link, we believe that populating bigger PKB improves the performance. Another observation is that Persona-Link variants outperform both manual and automatic paraphrasing approaches on every measure. Even though paraphrasing techniques might contribute to richer representation of personas, we argue that the paraphrasing approaches are limited to mitigating the trivial word overlap while utterances can be expressed in much flexible way in persona-based dialogue. While Bi- and Cross-encoder linking models result in better performance compare to the paraphrasing-based linking models due to out-dialogue matching which aims to mitigate linguistic bias, Persona-Link variants outperform in both measures. Especially, we observe all Persona-Link variants outperform Cross-encoder which compromises high computational cost for performance. We further analyze how Persona-Link achieves better performance in Persona-Link Evaluation section.

Performance with Black-box Training

To demonstrate that the augmented persona is not overclaimed by out-dialogue matching or out-corpus expansion, we compare the quality of augmented persona using dual task models. For that, we simulate a setting where none or few persona information is provided and new personas are augmented only in inference time by using off-the-shelf dialogue agents (*i.e.*, learned without augmenting personas in training time). Inspired by the evaluation method of measuring the relevance of generated response in (Su et al. 2020), we use the pre-trained dialogue agents as our diagnosis model assuming that the response quality of the model reflects the suitability of augmented persona in given dialogue contexts. That is, the suitability of the augmented persona to the context would be low if the augmented personas, which replace partial or entire persona of dataset, cause the failure of pre-trained dialogue agents. For the diagnosis agents, we adopt two dialogue models of Bi-encoder; each pretrained on Wikipedia (BERT-like; Devlin et al. 2019) or Reddit (an adapted setup for dialogue; Mazaré et al. 2018), and both fine-tuned on $\mathcal{D}_{\text{CHAT}}$. Note that, during fine-tuning, the dialogue agents only observe the original personas. As dual task models, we adopt two linking models; Bi-encoder trained on $\mathcal{D}_{\text{LINK}}$ and our Persona-Link trained on $\tilde{\mathcal{D}}_{\text{LINK}}$.

Specifically, the original persona sentences were removed at a given percentage; **No persona** eradicates the whole persona information and **Incomplete persona** discards 20%. In test time, linking models generate personas based on the utterance histories and merge them into the sample for testing dialogue models[2]. With persona augmentation, Table 2 shows the overall performance of dialogue models. In each setting, we report the base performance with information limit, where no linking model is available, *i.e.*, N/A. In the last row, we report the upper bound performance, where dialogue models access full information of **Original persona**.

First, compared to N/A, we observe linking models significantly improve (+14.6% at most) response quality by backtracking personas from utterances for the augmentation. Dialogue models seem to benefit from augmented personas, even though the models haven't seen them in training process, suggesting that personas provided by Persona-Link successfully keep contextual relevance to the given dialogues. Second, Persona-Link models outperform Bi-encoder models in all settings. As the only difference between the two is commonsense expansion with regularized training, our approach seems to enrich dual task with common knowledge beyond the alignments in Persona-Chat. Thirdly, the performance gain is the greatest in the persona-free setup, which is common in real-world scenarios. This suggests that our persona augmentation may boost dialogue agents in bootstrapping user profiles on-the-fly. As observing only debiased personas improves a trained agent, we conclude that our approach efficiently counteracts linguistic bias in Persona-Chat.

Now, we examine how persona augmentation calibrates

[2]For fair comparison, all utterances on the development set remained unseen by linking models. Also, for faster computation, all utterances in dialogue were concurrently processed in batch.

Agent's personas:
My mom is a secretary.
I am a bodybuilder.
I have one brother.
(-) I like to eat a lot of meat.
(+) I am a meat eater.
(+) My favorite meal is steak.
Dialogue history:
User: Hi there, how are you tonight?
Agent: Great. I just finished a huge steak. How are you?
User: I am good, drinking some scotch.
Response without augmented personas:
Agent: Cool! What are your hobbies?
Response with augmented personas:
Agent: I am major meat eater to build muscles.

Table 3: The model calibration in response retrieval task based on online persona augmentation. After removing persona (-), a linking model augmented personas (+), which calibrates the prediction of dialogue agent, with any parameter updates.

model prediction in response retrieval, by showing an example in Table 3. After removing a persona in advance, a linking model augments its own personas into context based on dialogue history. Observing additional information relevant to utterances (*i.e.*meal) and linguistically debiased (*i.e.*steak), the model prediction was calibrated online to retrieve the gold response, which is more consistent and engaging. Note that we use the same dialogue model to retrieve both responses. (See Appendix C for details).

PERSONA-LINK Evaluation

In this section, we conduct additional experiments to evaluate the proposed process for the dual task.

Overall Linking Performance

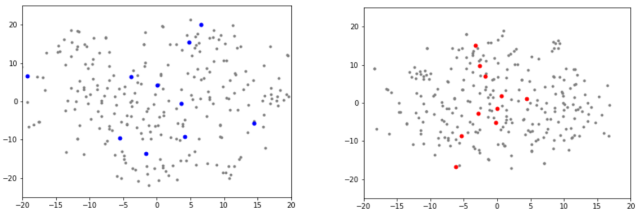
To study the model accuracy to the dual task, we perform comparative experiments of Persona-Link. To find ground truth pairs for Persona-Link, we manually collect 300 utterance-persona pairs from the test data in Persona-Chat, including 230 unique ground truth personas with seven annotators, which achieve a very good agreement (Cohen's kappa = 0.86). The number of utterances per persona is 1.16. As in Persona-Chat evaluation, we measure Recall@k and MRR as the model performance to the gold persona.

We consider five IR baselines trained on $\mathcal{D}_{\text{LINK}}$: **Cosine Similarity**, **BM25** (Robertson et al. 1995), **K-NRM** (Xiong et al. 2017) **Conv-KNRM** (Dai et al. 2018), and **Bi-encoder** (Humeau et al. 2020). Also, to investigate the impact of commonsense expansion and label regularization, we consider two variants of our approach using alternative expansion methods: **EDA** (Wei and Zou 2019), and **Paraphrasing** (Mallinson, Sennrich, and Lapata 2017). More details for these baselines are presented in Appendix B.

Table 4 shows the overall linking performance. First, we observe that among IR baselines trained on $\mathcal{D}_{\text{LINK}}$, Bi-encoder achieves the best performance, which validates our choice of base model for the dual task. Second, all models trained on $\mathcal{D}_{\text{LINK}}$ show poorer results than Ours and its variants using expansion methods with regularized training. This means that leveraging expanded dataset and calibrating labels into soft labels is more appropriate than naive IR models. Finally, Ours outperforms its variants (and other linking models as well) in all metrics. We note that while utterance or persona remains semantically similar by EDA or paraphrasing, Ours expands some commonsense attributes so that potentially relevant mapping can be captured. This supports the efficacy of exploiting commonsense as anchors, which may associates learned representations beyond the semantically close alignments. In Figure 3, we visualize the utterance embeddings without (Figure 3a) and with (Figure 3b) commonsense expansion. We observe that the utterance embeddings relevant to a persona gets closer with commonsense expansion, which can be a possible explanation for the performance boost of Ours exploiting commonsense over its variants.

Linking Model	R@1	R@10	MRR
Cosine Similarity	0.108	0.349	0.190
BM25	0.491	0.792	0.595
K-NRM	0.294	0.591	0.384
CONV-KNRM	0.345	0.628	0.439
Bi-encoder	0.583	0.871	0.684
Ours (EDA)	0.599	0.885	0.709
Ours (Para)	0.610	0.885	0.715
Ours	**0.669**	**0.922**	**0.759**

Table 4: Linking performance on Persona-Link



(a) w/o commonsense expansion

(b) w/ commonsense expansion

Figure 3: t-SNE visualization on the development set. The colored points indicate the utterance embeddings relevant to persona "I have a marketing job".

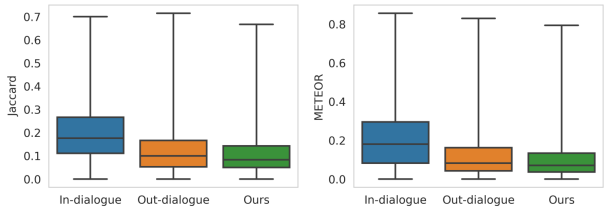


Figure 4: Similarity with respect to linking method.

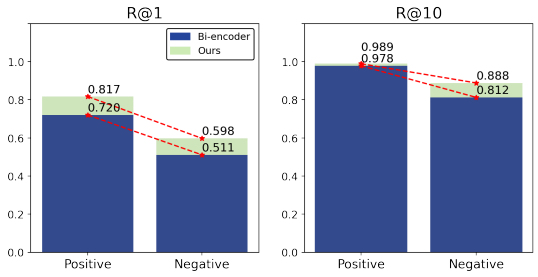


Figure 5: Performance with respect to PI result.

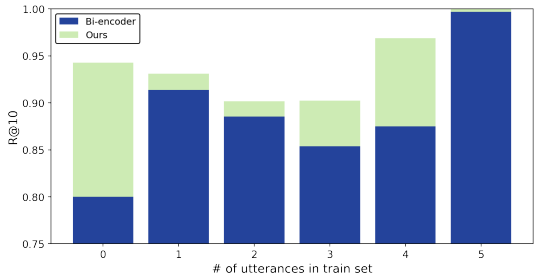


Figure 6: Performance with respect to data size.

Analysis on Linguistic Bias

In addition to the linking performance, we further analyze the linked utterance-to-persona alignments by linking models to see their linguistic bias. We plot the linguistic similarity between utterance-to-persona alignments in Figure 4. Following (Park et al. 2019), we calculate Jaccard similarity and METEOR, widely known metric for lexical similarity and semantic similarity, respectively. Jaccard similarity measures word co-occurrence and METEOR measures exact, stem, synonym, and paraphrase matches. For both metrics, an alignment with lower similarity can be seen as the one with less linguistic bias. Two models drop the original similarity from 0.2 to 0.1 on average, indicating the efficacy of out-dialogue semantic matching. Ours shows a slight decrease compared to out-dialogue model which confirms that Ours successfully leverages the commonsense expansion to align utterances to debiased persona.

We compare the performance of θ_{LINK} and $\tilde{\theta}_{\text{LINK}}$ in semantically different pair sets to investigate whether commonsense expansion helps the model link to debiased persona. Here, we use a BERT model fine-tuned with a popular paraphrase identification (PI) dataset called Quora Question Pairs to split our utterance-persona pairs into semantically equivalent and non-equivalent sets. With this model, only 35.5% of utterance-persona pairs are inferred as positive (*i.e.*, paraphrase). For analysis, in Figure 5, we split the overall utterance-persona pairs into positive and negative sets by the PI model, and ablate the performance of linking model with or without using commonsense knowledge and label regularization. As a result, while $\tilde{\theta}_{\text{LINK}}$ shows better performance than θ_{LINK} in all sets, $\tilde{\theta}_{\text{LINK}}$ achieves much better performance in negative set of R@10. This demonstrates the advantage of using commonsense knowledge, which gives the model the ability to link semantically different pairs that leads $\tilde{\theta}_{\text{LINK}}$ to link utterance to debiased persona.

Next, we evaluate how commonsense expansion affects when only a limited number of examples are provided in train time. Figure 6 shows the linking performance grouped by frequency of utterance-persona pairs in the training data in terms of R@10. Here, the performance means how well the model links utterances to their gold persona when the utterance-persona pair of the gold persona is not provided or few in train time. We observe that $\tilde{\theta}_{\text{LINK}}$ has more balanced performance for any persona groups, which is the effect of using commonsense knowledge, which can work as pivots when the utterance and persona has similar commonsense in common. Especially, $\tilde{\theta}_{\text{LINK}}$ shows the significant improvement for unseen personas (w/o any aligned training data) compared to θ_{LINK}, which shows that $\tilde{\theta}_{\text{LINK}}$ may be applicable to zero-shot linking. Since $\tilde{\theta}_{\text{LINK}}$ shows great performance to unseen utterance-persona pairs, we suggest to use our linking model with a large set of personas.

Related Work

Building personalized dialogue agents has been a popular task recently. Zhang et al. (2018) introduced the Persona-Chat dataset, a crowd-sourced conversation dataset with persona information, to improve model engagingness and consistency. Based on such data, recent works focus on improving persona-grounded dialogue agent's performance (Kim, Kim, and Kim 2020; Li et al. 2020; Song et al. 2020; Huang, Zhu, and Gao 2020; Zhang et al. 2019; Luo et al. 2019). Yavuz et al. (2019) designed the DeepCopy model, which leverages copy mechanism to incorporate persona texts. Song et al. (2019) integrated persona texts into the Per-CVAE model for generating diverse responses. However, prior work reports several linguistic bias which hamper learning to ground more engaging and consistent utterances.

Few recent works focused on augmenting grounding with commonsense knowledge with successful applications in open-domain dialogue agent (Ghazvininejad et al. 2018; Moon et al. 2019). Ghazvininejad et al. (2018) generalizes the Seq2Seq approach to neural conversation models by combining conversational and KG data. Moon et al. (2019) propose DialKG walker model that learns the transitions of dialogue contexts as structured traversals over KG. In this work, we extend this effort into improving persona-grounded dialogue dataset in data-centric view via augmenting relevant personas.

Conclusion

In this work, we propose a primal-dual task framework to debias the Persona-Chat dataset and its dialogue model without any human effort. By exploiting external LMs and knowledge distillation in a systematic way of overcoming the linguistic bias in Persona-Chat, our linking model was effective in improving 11.7% in dialogue accuracy from BERT Bi-encoder model using the original Persona-Chat dataset. We hope future research to leverage the graph structure of PKB with commonsense attributes for better persona linking.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020-11-0863). Jinyoung Yeo is a corresponding author.

References

Bowman, S.; and Dahl, G. 2021. What Will it Take to Fix Benchmarking in Natural Language Understanding? In *Proceedings of ACL*.

Dai, Z.; Xiong, C.; Callan, J.; and Liu, Z. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of WSDM*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Ghazarian, S.; Liu, Z.; Chakrabarty, T.; Ma, X.; Galstyan, A.; and Peng, N. 2021. DiSCoL: Toward Engaging Dialogue Systems through Conversational Line Guided Response Generation. *arXiv preprint arXiv:2102.02191*.

Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; Yih, W.-t.; and Galley, M. 2018. A knowledge-grounded neural conversation model. In *Proceedings of AAAI*.

Huang, M.; Zhu, X.; and Gao, J. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems*.

Humeau, S.; Shuster, K.; Lachaux, M.-A.; and Weston, J. 2020. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *Proceedings of ICLR*.

Hwang, J. D.; Bhagavatula, C.; Le Bras, R.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In *Proceedings of AAAI*.

Jeong, M.; Choi, S.; Yeo, J.; and Hwang, S.-w. 2021. Label and Context Augmentation for Response Selection at DSTC8. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Kim, H.; Kim, B.; and Kim, G. 2020. Will I Sound like Me? Improving Persona Consistency in Dialogues through Pragmatic Self-Consciousness. In *Proceedings of EMNLP*.

Li, A. W.; Jiang, V.; Feng, S. Y.; Sprague, J.; Zhou, W.; and Hoey, J. 2020. ALOHA: Artificial Learning of Human Attributes for Dialogue Agents. In *Proceedings of AAAI*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*.

Liu, Y.; Yang, T.; You, Z.; Fan, W.; and Yu, P. S. 2020. Commonsense Evidence Generation and Injection in Reading Comprehension. In *Proceedings of SIGDIAL*.

Logeswaran, L.; Chang, M.-W.; Lee, K.; Toutanova, K.; Devlin, J.; and Lee, H. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of ACL*.

Luo, L.; Huang, W.; Zeng, Q.; Nie, Z.; and Sun, X. 2019. Learning personalized end-to-end goal-oriented dialog. In *Proceedings of AAAI*.

Majumder, B. P.; Jhamtani, H.; Berg-Kirkpatrick, T.; and McAuley, J. 2020. Like hiking? You probably enjoy nature: Persona-grounded Dialog with Commonsense Expansions. In *Proceedings of EMNLP*.

Mallinson, J.; Sennrich, R.; and Lapata, M. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of EACL*.

Mazaré, P.-E.; Humeau, S.; Raison, M.; and Bordes, A. 2018. Training millions of personalized dialogue agents. In *Proceedings of EMNLP*.

Moon, S.; Shah, P.; Kumar, A.; and Subba, R. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of ACL*.

Park, S.; Hwang, S.-w.; Chen, F.; Choo, J.; Ha, J.-W.; Kim, S.; and Yim, J. 2019. Paraphrase diversification using counterfactual debiasing. In *Proceedings of AAAI*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Robertson, S. E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M. M.; Gatford, M.; et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp*.

Song, H.; Zhang, W.-N.; Cui, Y.; Wang, D.; and Liu, T. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of IJCAI*.

Song, H.; Zhang, W.-N.; Hu, J.; and Liu, T. 2020. Generating persona consistent dialogues by exploiting natural language inference. In *Proceedings of AAAI*.

Su, H.; Shen, X.; Zhao, S.; Xiao, Z.; Hu, P.; Zhong, R.; Niu, C.; and Zhou, J. 2020. Diversifying Dialogue Generation with Non-Conversational Text. In *Proceedings of ACL*.

Su, S.-Y.; Chuang, Y.-S.; and Chen, Y.-N. 2020. Dual Inference for Improving Language Understanding and Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 4930–4936.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR 2011*, 1521–1528. IEEE.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of ICLR*.

Wang, J.; Chen, K.; Shou, L.; Wu, S.; and Chen, G. 2021. Effective Slot Filling via Weakly-Supervised Dual-Model Learning. In *Proceedings of AAAI*.

Wei, J.; and Zou, K. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of EMNLP-IJCNLP*.

Welleck, S.; Weston, J.; Szlam, A.; and Cho, K. 2018. Dialogue natural language inference. In *Proceedings of ACL*.

Wolf, T.; Sanh, V.; Chaumond, J.; and Delangue, C. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Wu, L.; Petroni, F.; Josifoski, M.; Riedel, S.; and Zettlemoyer, L. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of EMNLP*.

Xia, Y.; Qin, T.; Chen, W.; Bian, J.; Yu, N.; and Liu, T.-Y. 2017. Dual supervised learning. In *International Conference on Machine Learning*, 3789–3798. PMLR.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised data augmentation for consistency training. *Proceedings of NeurIPS (workshop)*.

Xiong, C.; Dai, Z.; Callan, J.; Liu, Z.; and Power, R. 2017. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. In *Proceedings of SIGIR*.

Yavuz, S.; Rastogi, A.; Chao, G.-L.; and Hakkani-Tur, D. 2019. DeepCopy: Grounded Response Generation with Hierarchical Pointer Networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 122–132.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of ACL*.

Zhang, Y.; Gao, X.; Lee, S.; Brockett, C.; Galley, M.; Gao, J.; and Dolan, B. 2019. Consistent dialogue generation with self-supervised feature learning. *arXiv preprint*.

Zhu, S.; Cao, R.; and Yu, K. 2020. Dual learning for semi-supervised natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 1936–1947.