Dialogue Response Generation Using Personal Facts and Personality Traits

Weiwen Su[1] Naoki Yoshinaga[2] Yuma Tsuta[1] Masashi Toyoda[2]
[1]The University of Tokyo [2]Institute of Industrial Science, The University of Tokyo
{su-w,ynaga,tsuta,toyoda}@tkl.iis.u-tokyo.ac.jp

Abstract

Persona-based chatbots assuming a specific persona for chatbots can generate consistent responses given the persona. Existing persona-based dialogue datasets such as PersonaChat and Multi-Session Chat (MSC), however, contain mainly personal facts (*e.g.*, "I like cats.") but lack personality traits ("I am extraverted."). We thus automatically annotate the MSC dataset with personality traits to train persona-based chatbots using personal facts and personality traits. Experimental results on the personality-augmented MSC datasets confirmed that our chatbot improves personality consistency scores, when using a personality-aware reranking.

1 Introduction

Open-domain dialogue systems such as Siri and ChatGPT have become more common in our daily lives. As a daily conversation partner, we expect chatbots to converse like humans with a consistent persona. However, since the conversation data used to train chatbots usually compile conversations from various persons, the resulting chatbots are likely to generate inconsistent responses.

To address those inconsistent responses by data-driven chatbots, researchers consider the identity of speakers to generate responses [1, 2, 3, 4, 5, 6, 7, 8]. To model and control the chatbot persona explicitly, Zhang et al. [3] built PersonaChat, which provides speaker profiles as text descriptions (*e.g.*, "I have a dog."), and most of the following studies on persona-based chatbots utilized this dataset or its extension, Multi-Session Chat (MSC) [9]. Although the existing datasets for persona-based chatbots contain various profiles to describe a speaker persona, those profiles are mainly personal facts such as personal tastes, relatives, social status, and experiences, and barely include personality traits such as agreeableness and extraversion. Meanwhile, Saha et al. [10] predict Big-Five personality traits for speakers in several dialogue datasets to control the style of generated responses. However, their datasets do not contain personal facts.

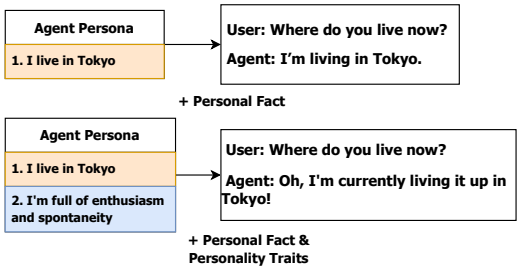

Figure 1 An example of a generation with or without personality traits.

In this study, aiming to investigate the impact of personality traits in persona-based chatbots, we automatically annotate personality traits to existing persona-based datasets, MSC, using Big-Five personality predictor trained on Pandora dataset [11]; we then train and evaluate a persona-based chatbot using the profiles on personal facts and the estimated personality traits of the speaker. An issue here is how to represent predicted personality traits (category with intensity). We adopt the same short text descriptions as the original profiles on personal facts to maintain interpretability and flexibility. To enhance personality consistency, we incorporate a response reranking model [7] to compute the consistency between the personality profiles and the generated utterance to choose the response with the highest personality consistency.

We use the personality-augmented MSC dataset to evaluate the impact of using both profiles on the original personal facts and predicted personality traits in a persona-based chatbot. The automatic and human evaluation confirmed the effectiveness of personality-based profiles, when we use the proposed reranking model.

2 Related Work

Persona-based Response Generation The problem of inconsistent responses of data-driven chatbots was first pointed out by Li et al. [1]. To address this problem, they trained the model with user embeddings from the speakers' dialogue histories to generate more consistent responses. Meanwhile, Zhang et al. [3] proposed PersonaChat, the most commonly used dataset for persona-based chatbots; it compiles conversations between a pair of speakers that role-play given persona, a series of text descriptions (profiles). To address the scarcity of persona in PersonaChat, Majumder et al. [12] expand the original profiles using commonsense knowledge. In the context of long-term conversation, the persona may change over time. Xu et al. [9] thus extends PersonaChat with future conversation sessions, referred to the resulting dataset as MSC. However, the profiles in the PersonaChat and MSC contain mainly personal facts such as personal tastes, relatives, and social status, but lack personality traits.

Personality Controlled Dialogue Generation In dialogue modeling, personality traits such as Big-Five personality are considered to affect speaking styles. In the early stage, Mairesse and Walker [13] leverage a statistical generation model, focusing on extraversion personality. Recently, Wang et al. [14] proposed a seq2seq model for Big-Five personality-conditioned response generation. Saha et al. [10] leverage Big-Five personality and discourse intents as stylistic control codes to generate stylistic responses. Although personal facts and personality traits are two important factors that characterize a speaker, no study attempts to model both persona information into account in response generation, due to the absence of datasets.[1)]

We thus add personality traits as text descriptions to the MSC dataset to train and evaluate persona-based chatbots using both personal facts and personality traits.

3 Approach

In this section, we describe our method to generate a dialogue response based on not only personal facts but also personality traits in chat conversation. The task is defined as, given the dialogue context C, and agent personality traits $P_{\text{personality}}$ and personal facts P_{fact}, the model optimizes the generation of response R to maximize $P(R|C, P_{\text{personality}}, P_{\text{fact}})$ according to the dataset.

1) Very recently, RealPersonaChat [15] datasets have been constructed, including massive Japanese conversations with both personal facts and personality traits provided by the speakers. However, the datasets have not been released at the moment.

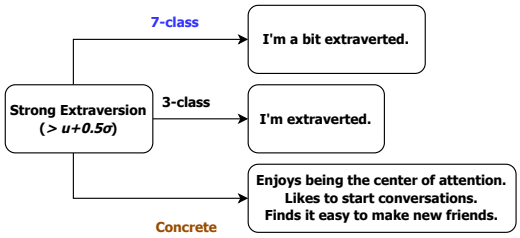

Figure 2 A schema of the three personality verbalization methods. A case of extraversion intensity beyond $u + 0.5\sigma$ is shown.

3.1 Annotating Personality Traits

To augment the existing persona-based dialogue dataset with personality traits, we train a personality detector using Pandora [11] dataset, which consists of Big-Five personality traits and their intensities for 1500+ Reddit users. We develop a RoBERTa [16]-based regression model to predict the target user's intensities of all personality traits at once from the target user's utterances. Following previous work [10], this detector predicts a target user's personality from accumulated target user's utterances, rather than aggregating the predictions from each utterance. Therefore, this model processes several utterances at once like batch processing.

In this study, we explore an effective way to verbalize the detected intensity of personality traits. Specifically, based on the distribution of the personality traits, three verbalization approaches to create personality profiles as shown in Figure 2 are explored in our work:

3-class verbalization Split each Big-Five personality trait into three classes (positive, neutral, and negative) with 0.5 standard deviation σ from the mean u of this dataset. Then verbalize the personality traits in the same format as the personal fact profiles in the dataset using the adjectives of each Big-Five personality and its opposite expression (*e.g.* "Extraverted" and "Introverted"). We do not add personality profiles for the neutral class.

7-class verbalization Rather than the three classes, using 0.5, 1, and 2 σ from u as the threshold, we further divide each personality trait into seven classes by adding adverbs ("a bit", "quite" and non-adverb) of the degree to create more refined personality profiles.

Specifically, three adverbs with original and opposite adjectives result in six classes in addition to the neutral class.

Concrete verbalization Instead of the original and opposite adjectives of Big-Five personality, we take more concrete descriptions of the people owning such personality traits from psychology websites[2] as the personality profiles. Specifically, we sampled the concrete descriptions according to the predicted intensity (33%, 66%, and 100% descriptions among all the descriptions used and combined).

3.2 Response generation based on personal facts and personality traits

Though the personal facts and the personality traits are both in the format of text descriptions, the elements they influence in the conversation are different. Rather than the basic facts and demographic features, the personality traits influence mainly the action pattern, speaking style, and more complicated aspects. Thus, we first concatenate individual descriptions in the personal facts and personality traits respectively, then concatenate the two sequences of descriptions with a special token. We feed the whole descriptions concatenated before the context to the decoder.

Personality-aware Reranking Naively incorporating personality traits may not facilitate the model to fully utilize the given personality traits. Thus, inspired by [7] using reranking to improve consistency between personal facts and generated responses leveraging a model trained on DNLI dataset [17], we propose to reuse the augmented training data of MSC dataset to train a personality consistency prediction model for response reranking. In this task, the response reranking model using RoBERTa [16] regression model calculates the consistency between personality traits and the generated responses. For the training data, we make positive sample pairs of personality traits with utterances in the original dialogue and negative sample pairs of personality profiles with utterances from other dialogues, and use the cosine similarity between the Big-5 personality intensity of those personality traits and intensity of personality traits of the utterance. We use the triplet of the first part of personality traits, the second part of utterances, and the cosine similarity as training data. In the inference stage, given personality profiles and one candidate utterance, the model could predict the consistency score between them.

4 Experiments

In this section, we train chatbots on the augmented MSC dataset to confirm the effectiveness of using both personality traits and personal facts with our reranking method.

4.1 Settings

Datasets We use the MSC dataset [9], augmented by predicted personality traits for evaluation. Since individual dialogue sessions, a series of consecutive utterances, in the MSC datasets can be generated by different pairs of speakers even if they maintain the same personal facts, we ignore multi-session settings in the dataset and handle individual sessions as independent dialogues to guarantee the consistency of speakers' personalities.

Models We adopt DialoGPT[3] [18] as a backbone of persona-based chatbots, which finetunes a GPT-2 [19] on Reddit comment chain data. The models to be compared in the experiment are as follows:

Baseline We finetuned the DialoGPT model on the original MSC datasets with and without the personal facts as baselines. We hereafter referred to them as **Baseline** and **+person. facts**, respectively.

Proposed We fine-tuned three DialoGPT models on the personality-augmented MSC dataset by combining three personality verbalization methods (3-class, 7-class, and concrete verbalization). We referred them to as **3-class**, **7-class**, **concrete** respectively.

We performed the reranking of 5 response candidates for the three proposed models.

Metrics We use perplexity (ppl.) and BLEU-1/2 [20] as basic metrics. We also evaluate the Distinct-1/2 (DIST-1/2) [21] to show whether the generated responses exhibit a certain degree of diversity.

In addition to these generic metrics, following the previous studies, we evaluate the personal fact consistency by consistency score (C. score) [22] which is a textual entailment score computed using a RoBERTa [16] model trained on the DNLI dataset [17]. As for personality consistency, inspired by the utterance-level Pearson correlation used by [10], we compute the Pearson correlation of dialogue-level personality detected from gold and gener-

2) https://www.verywellmind.com/the-big-five-personality-dimensions-2795422

3) https://huggingface.co/microsoft/DialoGPT-small

Table 1 Automatic results. BLEU-1/2 and Distinct 1/2 are scaled by multiplying 100. Pearson correlation is with a p-value < 0.05.

Setting	ppl.	BLEU-1/2	DIST-1/2	C. score	Pers. Corr.
Baseline	18.81	10.86/1.99	**2.87**/28.70	0.332	0.386
+person. facts	18.52	10.86/2.06	2.83/28.73	0.443	0.390
Proposed (Baseline +person. facts and personality traits)					
3-class	18.51	**11.39/2.17**	2.83/28.72	**0.450**	0.646
7-class	**18.46**	11.11/2.04	2.82/**28.75**	0.429	**0.664**
concrete	18.56	11.14/2.11	2.73/28.34	0.443	0.638

Table 2 Human evaluation for response quality.

Setting	Fluency	Coherence	Infomative.	Consist.
+person. facts	3.85	2.77	2.65	0.17
Proposed (7-class)	3.88	2.69	2.57	0.18
Human	4.88	4.78	4.53	0.66

ated responses as personality correlation (Pers. Corr.)

In human evaluation, as for response quality, we choose Fluency, Coherence, and Informativeness following the setting of [7], and ask human subjects to annotate 5-point Likert scales for the three dimensions, where 1 point means bad quality, 3 points mean moderate and 5 points mean a perfect performance. As for personal fact consistency, a point within -1,0,1 is assigned to each response, which means contradicted, neutral, or related to given personal facts.

4.2 Main Results

Table 1 list the automatic evaluation results of the generated responses. We can observe that the personality traits did not change perplexity, Distinct-1/2, and personal fact consistency (C. score) (**Proposed** vs. **+person. facts**). As for BLEU scores, we can see a consistent increase of BLEU-1 with personality traits. As for Personality Correlation, we could observe that the proposed models gain great increase and the 7-class verbalization of personality traits achieves the best personality consistency. In short, the proposed models improve the consistency with assigned personality profiles while maintaining other performance with **+person. facts**.

Table 2 shows the human evaluation for response quality. Considering the automatic results, we chose the best-performing models with the 7-class verbalization of personality traits for evaluation. The results show that the 7-class setting slightly outperformed **+person. facts** in terms of Fluency and Consistency, and the model trained with only personal facts (**+person. facts**) achieved better results in Coherence and Informativeness. The use of personality traits does not show significantly negative influence on the response quality.

Table 3 Ablation studies. BLEU-1/2 and DIST-1/2 are scaled by multiplying 100. Pearson correlation is with a p-value < 0.05.

Setting	ppl.	BLEU-1/2	DIST-1/2	C. score	Pers. Corr.
Proposed w/o Personality-aware Reranking					
3-class	18.51	11.02/1.99	2.80/28.52	0.446	0.400
7-class	18.46	10.88/1.98	2.86/28.74	0.426	0.387
concrete	18.56	11.01/1.98	2.75/28.11	0.426	0.395

4.3 Ablation Studies

To examine the influence of the reranking of response candidates on personality consistency, we also conducted an ablation study as shown in Table 3. The results of the ablation test without reranking (w/o Personality-aware Reranking) allow us to observe the huge impact of response-candidate reranking on personality consistency. Compared with previous results of Personal Facts (in Table 1), the model trained with personality traits could not improve the Personality Correlation. This is probably because the model may have a strong emphasis on personal facts due to the process of creating the MSC dataset, and naively incorporating personality traits did not contribute to the personality consistency. At the moment, personality-aware reranking is vital to improve personality consistency.

5 Conclusions

In this study, we explore the use of personality traits in addition to personal facts in persona-based chat response generation. We augment the existing MSC dataset [9] with personality traits using a personality detector trained on the Pandora dataset. To fully leverage the predicted personality traits, we explore personality verbalization and propose a personality-aware reranking method to pick response candidates with better personality consistency. Experimental results on the personality-augmented MSC dataset show an improvement in personality consistency.

As for future work, we plan to design a more effective model structure to fully utilize personal facts and personality traits at the same time.

Acknowledgement

This work was partially supported by the special fund of Institute of Industrial Science, The University of Tokyo, by JSPS KAKENHI Grant Number JP21H03494, JP21H03445, and by JST, CREST Grant Number JPMJCR19A, Japan.

References

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

[2] Shoetsu Sato, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. Modeling situations in neural chat bots. In Allyson Ettinger, Spandana Gella, Matthieu Labeau, Cecilia Ovesdotter Alm, Marine Carpuat, and Mark Dredze, editors, **Proceedings of ACL 2017, Student Research Workshop**, pp. 120–127, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[3] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 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)**, pp. 2204–2213. Association for Computational Linguistics, July 2018.

[4] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 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**, pp. 4279–4285. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[5] Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In **Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '21, p. 555–564. Association for Computing Machinery, 2021.

[6] Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 167–177. Association for Computational Linguistics, August 2021.

[7] Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. SimOAP: Improve coherence and consistency in persona-based dialogue generation via over-sampling and post-evaluation. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 9945–9959. Association for Computational Linguistics, July 2023.

[8] Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. You impress me: Dialogue generation via mutual persona perception. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1417–1427. Association for Computational Linguistics, July 2020.

[9] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5180–5197. Association for Computational Linguistics, May 2022.

[10] Sougata Saha, Souvik Das, and Rohini Srihari. Stylistic response generation by controlling personality traits and intent. In **Proceedings of the 4th Workshop on NLP for Conversational AI**, pp. 197–211. Association for Computational Linguistics, May 2022.

[11] Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. PANDORA talks: Personality and demographics on Reddit. In **Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media**, pp. 138–152. Association for Computational Linguistics, June 2021.

[12] Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 9194–9206. Association for Computational Linguistics, November 2020.

[13] François Mairesse and Marilyn Walker. PERSONAGE: Personality generation for dialogue. In **Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics**, pp. 496–503. Association for Computational Linguistics, June 2007.

[14] Wanqi Wu and Tetsuya Sakai. Response generation based on the big five personality traits. 2020.

[15] Ao Guo Shota Mochizuki Tatsuya Kawahara Ryuichiro Higashinaka Sanae Yamashita, Koji Inoue. Realpersonachat: A realistic persona chat corpus with interlocutors' own personalities. In **Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation**, 2023.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **ArXiv**, Vol. abs/1907.11692, , 2019.

[17] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3731–3741. Association for Computational Linguistics, July 2019.

[18] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In **ACL, system demonstration**, 2020.

[19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318. Association for Computational Linguistics, July 2002.

[21] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 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**, pp. 110–119. Association for Computational Linguistics, June 2016.

[22] Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. Personalizing dialogue agents via meta-learning. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5454–5459. Association for Computational Linguistics, July 2019.

A Appendix

A.1 Human Evaluation Point Assignment Instruction

We show the concrete point assignment instruction of human evaluation about response quality (Fluency, Coherence, Informativeness) as follows:

Flue.
- 1 point: Extremely difficult to understand, with frequent language issues.
- 2 points: Communication is often unclear, with noticeable language challenges.
- 3 points: Communication is generally clear, with occasional disruptions in fluency.
- 4 points: Communication is very clear, with minimal interruptions or language issues.
- 5 points: Extremely fluent communication, very natural and easy to understand.

Coh.
- 1 point: The text lacks any coherence, making it extremely difficult to follow.
- 2 points: Coherence is often lacking, and the text lacks logical flow.
- 3 points: Text is generally coherent, but there are some logical breaks or disruptions.
- 4 points: The text is very coherent, with strong logical connections between sentences and ideas.
- 5 points: Text is extremely coherent, with seamless logical flow and strong connections.

Info.
- 1 point: Provides very limited information, offering little to no assistance.
- 2 points: Provides limited information, with some helpful aspects.
- 3 points: Provides sufficient information, but may lack depth or detailed explanations.
- 4 points: Provides ample information, with moderate depth, and is helpful to the user.
- 5 points: Provides highly detailed and comprehensive information, greatly assisting the user.