DEIM Forum 2021 E33-2

User-oriented Dialogue Response Generation Using Multi-task Learning with Personality Prediction

Weikun MA†, Naoki YOSHINAGA††, and Masashi TOYODA††

† Information Science and Technology, The University of Tokyo, 7–3–1, Hongo, Bunkyo-ku, Tokyo 113–8656, Japan
†† Institue of Industrial Science, The University of Tokyo, 4–6–1 Komaba, Meguro-ku, Tokyo 153–8505, Japan
E-mail: †, ††{maweikun, ynaga, toyoda}@tkl.iis.u-tokyo.ac.jp

Abstract Modeling the personality of users helps chatbots to generate more personalized responses that are more specific and fit to the target users. Since existing approaches require conversation histories of the target users to train a model for generating responses personalized to them, it is difficult to generate responses personalized to users unseen in the training data. In this study, we propose a method that learns to instantly capture users' personalities from their dialogue contexts and generates personalized responses. We first create a profile-augmented dialogue dataset using conversational logs and the associated users' profile descriptions collected from Twitter, and then train our model to respond each utterance using multi-task learning with personality estimation. In inference, unseen users can be successfully processed without their profiles. We conduct both automatic evaluation and human evaluation to show the effectiveness of our approach.

Key words personalization, dialogue system, multi-task learning, twitter

1 Introduction

Open-domain dialogue systems have wide applications including chatbots and virtual personal assistants. Thanks to the spread of microblogs, such as Twitter, and the accessibility to massive dialogue logs, there has been a growing research interest in neural-based end-to-end dialogue systems [1] [2] [3]. In particular, the Sequence-to-sequence (Seq2seq) framework [4] [5], which learns a mapping from the input to the output response, has drawn wide attention.

The content of conversations deeply depends on personalities of the involved interlocutors. Therefore, personalized dialogue response generation has been widely concerned as a research topic. Researchers so far have concentrated on endowing the dialogue agent with a steady personality to improve the consistency. Some methods try to manipulate the input of the decoder to obtain responses personalized to a given personality [6] [7] [8]. To utilize users' personalities, the method that attaches user IDs to dialogue contexts has been proposed [9]. Besides, Liu+ assume that the personality of each user has been known before, and the model can take advantage of these personalities [10]. However, all of these methods require user-specific data in inference to generate responses, and without the data, it becomes difficult to train a good model.

To solve the unseen user problem of the user-oriented dialogue agent to generate personalized responses, we propose to instantly estimate personalities from given utterances to generate personalized responses. We noticed that two human interlocutors are capable of perceiving the implicit personal information from the dialogue flow and of giving appropriate responses accordingly. Hence our method manages to achieve the same capability without any personality data in advance. We collect a vast number of dialogue logs from Twitter, along with the profiles of the associated users who start the conversation, and construct a profile-augmented dialogue dataset (see Sec. 2). As our main contribution, we employ multi-task learning where response generation is the main task and personality estimation is the auxiliary task (Sec. 3). We take the estimated personality as a high-dimensional vector or as a category. Accordingly we have two types of auxiliary tasks: personality induction (regression task) and personality classification, respectively. Because the auxiliary task is only considered in training, our model is capable of handling conversations with unseen users. The normal Seq2seq model is selected as our baseline, as the proposed model has nothing different to a normal Seq2seq model in test stage.

To verify the effectiveness of our methods, we train our model using Twitter conversation logs and evaluate the generated responses by automatic metrics, BLEU and RUBER [11], and human evaluation (Sec. 4). An improvement is found in RUBER score with 0.401 for the better one of two proposed models and 0.397 for the baseline. According to the results of human evaluation, the proposed model also beats the baseline on fluency and user relevance.

Our contributions are as follows: (1) Propose multi-task learning with personality induction method to capture the user's personality at a fine-grained level; (2) Propose multi-task learning with personality classification method to model the user's personality at a coarse level.

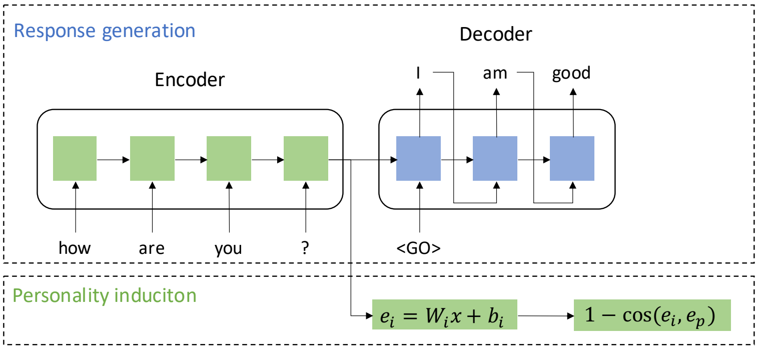

Figure 1: The overall architecture of multi-task learning with personality induction.

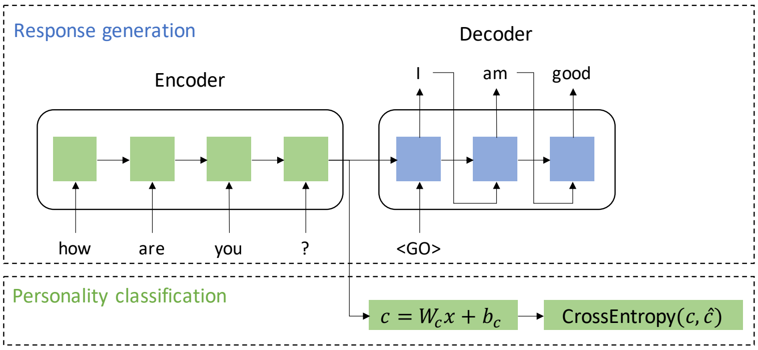

Figure 2: The overall architecture of multi-task learning with personality classification.

2 Profile-augmented Dialogue Dataset

We construct a profile-augmented dialogue dataset based on a large volume of tweets and associated replies collected from Twitter. For convenience, we take a tweet and its reply as a pair of dialogue, and all the dialogues are single-turn. Besides, the profiles of the involved users in the context of each pair of dialogue is also stored along with the dialogue. Hence each item of data in the dataset has three components (see Table 1 for a concrete example):

(i) Profile: a description text written by a user to introduce himself.

(ii) Context: a tweet posted by a user, who is taken as the initiator of a conversation and his profile will be collected at the same time.

(iii) Response: a reply posted by another user, who is taken as the respondent in a one-turn conversation.

The profile-augmented dialogue dataset has a large volume with more than 8.3 million pairs of dialogue, as well as profiles of all the associated users who start the conversation. The detailed statistics of the dataset are shown in Table 2.

Profile	Dad, husband, President, citizen.
Context	Here are some of my favorite songs of the year. As usual, I had some valuable consultation from our family music guru, Sasha, to put this together.
Response	That is not even Wizkid's best song on the album.

Table 1: An example of data format in the large-scale Twitter dataset.

	train	validation	test
no. of dialog pairs	8,327,300	1,000,000	500,000
avg length of utterances	12.3	12.3	12.3
no. of users	354,772	173,119	122,042

Table 2: Statistics of the Large-scale Twitter Dataset after oversampling.

3 Multi-task Learning with Personality Estimation

Our method aims to generate responses personalized to the target user, and thus learn to respond users while guessing their personalities. Different from the existing studies [9] [10], which require the profiles or conversation histories of the target user, our method can handle various unseen users, and generate personalized responses without any data specific to the target user. To realize this requirement, we employ multi-task learning where the response generation is the main task and personality estimation is the auxiliary task. We combine their objective functions and train the two tasks together. The personality estimation task takes as input a dialogue context, and outputs the personality of the user who starts the conversation. We take the output personality 1) as a high-dimensional vector or 2) as a category. Correspondingly there are two auxiliary tasks: personality induction and personality classification. The model only considers the auxiliary task during training stage, which means the input of our model is exactly the same as an ordinary Seq2seq model while conducting the inference. Consequently, our model is capable of coping with the conversation with an unseen user and giving a response personalized to that user. The following parts unfold some details of the main task and the two auxiliary tasks.

3.1 Dialogue Response Generation

As the main task in multi-task learning, dialogue response generation models the utterances and interactive structure of the dialogue, and generating adequate response to maximize the log probability [2] [3]. We denote a pair of dialogue as a context c and a response r. Each r consists of a sequence of N tokens, i.e. $r = \{w_1, w_2, ..., w_N\}$, where w_i is a random variable sampled from the vocabulary V and represents the i-th token. The model parameterized by θ is trained to maximize the probability of the generated response r given a context c. The computation is shown as follows:

$$P_\theta(r) = \prod_{i=1}^{N} P_\theta(w_i | w_1, w_2, ..., w_{i-1}, c)$$

where $P_\theta(r)$ denotes the probability of the generated response.

Seq2seq model [4] [5] is composed of a RNN-encoder and a RNN-decoder. We employ it to solve this problem. We choose Long Short-Term Memory (LSTM) [12] to serve as the encoder and decoder. The LSTM-encoder maps a sequence of context tokens to a vector obtained from the last hidden state of the topmost layer. Then the vector is given to the LSTM-decoder for decoding and generating the response in an auto-regressive way. The auto-regressive decoding means that after each time step, the generated token is fed to the decoder again as the input of the next time step.

3.2 Personality Induction

We formulate the personality induction task as a regression task. We define the output personality as a high-dimensional vector, named personality embedding, which contains the personality of the user starting the conversation. The goal of this task is to make the encoded dialogue context similar to the personality embedding. The model for this task consists of an encoder and a linear transformation layer (Figure 1). The encoder is shared with the dialogue generation task. The loss function is defined as follows.

$$Loss = 1 - cos(e_i, e_p)$$

where $cos(\cdot)$ represents cosine similarity, e_p is the personality embedding, and e_i is the linearly transformed encoded dialogue context. By inducing the encoded context with personality embedding, the representation of context would contain information about personality.

To train the model of this task, we compute personality embeddings by adding up all the embeddings for words in each user profile. Then, we fix these emebddings during the whole training stage.

3.3 Personality Classification

We formulate the personality classification task as a classification task. Given a dialogue context, the goal of this task is to predict the class of the user who initiate the conversation. The model of this part is composed of an encoder and a linear transformation layer (see Figure 2). By minimizing the cross entropy loss of this task, the encoder learns to extract personality features from the dialogue context. Compared to the regression task, this task models the user's personality from a higher level.

In order to obtain the user categories as the target of classification, we cluster profile embeddings computed by SIF algorithm [13] and use the cluster ID of each user as their categories.

4 Experiments

We conducted experiments on our profile-augmented dialogue dataset to verify the effectiveness of our proposed multi-task learning method. As for evaluation, because it is difficult for automatic metrics to assess the relevance between the generated response and the target user, we employ both automatic evaluation and human evaluation.

4.1 Models

The entire model is built with Fairseq framework [14]. The encoder and the decoder are both 2-layer BiLSTM with 512-dimensional hidden states for each layer. The encoder of personality classification is shared with the main task, and the feedforward network has a 512-dimensional input layer, a 512-dimensional hidden layer and a 10-dimensional output layer, for all the users being clustered into 10 classes in preprocessing. We use the pre-trained 300-dimensional English word embeddings trained with fastText [15] to initialize the embedding layer. All the embeddings are set to updatable during training. For the combination of two objective functions, we times the loss of the auxiliary task by a constant number 50 and add them up. This weighted summation helps to bring the magnitude of two objectives into the same level and strengthens the impact of the auxiliary task.

4.2 Baseline

As mentioned above, the main part of our multi-task learning model is a Seq2seq model. If we only consider the inference stage, our proposed model has nothing different to a normal Seq2seq model. From this point of view, we choose a general Seq2seq model [16] as our baseline. The encoder and decoder of the baseline are also 2-layer BiLSTM with 512-dimensional hidden states for each layer. We train it on our large-scale Twitter dataset as same as the proposed model.

4.3 Automatic Evaluation

Although we use BLEU score the evaluate to results, it is not robust to evaluate responses having less word overlap with the ground

Model	cos	acc.	BLEU	RUBER
Seq2seq	-	-	0.12	0.397
+Induction	0.83	-	0.10	0.382
+Classification	-	0.782	**0.13**	**0.401**
Induction	**0.84**	-	-	-
Classification	-	**0.802**	-	-

Table 3: The experiment results of single task models, multi-task models and the baseline evaluated by BLEU and RUBER.

truth. In terms of the empirical observations that the ground truth is not the only answer and the context itself provides useful information to judge a response, Tao+ has proposed the Referenced metric and Unreferenced metric Blended Evaluation Routine (RUBER) [11]. According to its experimental results, RUBER score has a higher correlation with human evaluation. Therefore we decide to use RUBER as our automatic metric. RUBER exploits a BiGRU to compute the unreferenced score based on a dialogue context and the generated response. We trained our own unreferenced scorer as a single-layer BiGRU with 100-dimensional hidden state. For the referenced scorer we use the same fastText word embedding mentioned above.

Additionally, to show the effectiveness of multi-task learning, we compute the similarity of personality induction results and the accuracy of personality classification results. We also conduct ablation test by measuring the results of single-task model and of multi-task model.

The results of automatic evaluation is shown in Table 3. We can see that there is a small drop in both the similarity and the accuracy of the multi-task models against the single-task models. We consider that compared to the main task, the auxiliary task is more sensible to the disturbance introduced by the sharing of parameters. Hence the performance of two auxiliary tasks would decrease by a little. If we look at the BLEU score, we would find that each method have acheived a low value on BLEU, which means the generated responses are quite different from the reference. It makes the RUBER more important. In terms of RUBER scores, we observe that it is more effective to formulate the personality estimation as a classification task. Such a difference might be caused by the difference of complexity between regression and classification tasks.

Proposed	**Fluency**		**User relevance**	
	Win(%)	**Lose(%)**	**Win(%)**	**Lose(%)**
vs. Baseline	**51.8**	48.2	**53.3**	46.7
vs. Human	36.8	**63.2**	24.1	**75.9**

Table 4: Results of human evaluation on fluency and user relevance.

4.4 Human Evaluation

We conduct a human evaluation to compare the multi-task learning with personality classification method and the baseline. We adopt the comparative method for human evaluation. Annotators are asked to choose a winner from two parallel results. The results of ground truth, baseline model, and the proposed model are considered as three parallel evaluation targets. We regard each unique combination containing two of the three targets as a comparison pair. We evaluate the results from two aspects: fluency and user relevance. Fluency measures the grammatical correctness. User relevance measures to what extent the response is related to the personality of the user. Annotators are supposed to know the personality of a user based on his profile, as well as the context of the dialogue. Then they evaluate the results according to their understanding of the user. We invited three annotators to evaluate 300 random samples. One of them is a native English speaker, and another two have good language proficiency in English. We shuffle the order of responses from different models and the ground truth.

Table 4 shows the results of human evaluation. Consistent with the results of automatic metric, our classification model beats the baseline on fluency with a win rate of 51.8%. By employing multi-task learning with personality prediction, we obtained a win rate of 53.3% on user relevance against the baseline. Though the Seq2seq model, which is the baseline here, is the elementary model in the field of open-domain dialogue system, the results have shown that multi-task learning method can actually enhance the power of the Seq2seq model. Therefore, we expect that the same method could work on other advanced models.

5 Conclusion

To solve the unseen user problem of user-oriented dialogue response generation, we propose to use multi-task learning with personality estimation. We formulate personality estimation as two different types of tasks: personality induction (a regression task) or personality classification (a classification task). The auxiliary task shares the encoder parameters with the main task, dialogue response generation. In this way, the power to capture users' personalities learned by the auxiliary task can be shared with the main task. Experiments on our profile-augmented dialogue dataset demonstrates the effectiveness of our approach. For future work, we would like to apply our method to large-scale pre-trained models like DialoGPT.

Acknowledgements

The research was supported by NII CRIS collaborative research program operated by NII CRIS and LINE Corporation.

References

[1] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 583–593,

2011.

[2] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, p. 3776–3783, AAAI Press, 2016.

[3] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Beijing, China), pp. 1577–1586, Association for Computational Linguistics, July 2015.

[4] O. Vinyals and Q. V. Le, "A neural conversational model," in *ICML Deep Learning Workshop*, 2015.

[5] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan, "A neural network approach to context-sensitive generation of conversational responses," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Denver, Colorado), pp. 196–205, Association for Computational Linguistics, May–June 2015.

[6] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. 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)*, (Melbourne, Australia), pp. 2204–2213, Association for Computational Linguistics, July 2018.

[7] P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes, "Training millions of personalized dialogue agents," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (Brussels, Belgium), pp. 2775–2779, Association for Computational Linguistics, Oct.-Nov. 2018.

[8] W. Zeng, A. Abuduweili, L. Li, and P. Yang, "Automatic generation of personalized comment based on user profile," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, (Florence, Italy), pp. 229–235, Association for Computational Linguistics, July 2019.

[9] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Berlin, Germany), pp. 994–1003, Association for Computational Linguistics, Aug. 2016.

[10] Q. Liu, Y. Chen, B. Chen, J.-G. Lou, Z. Chen, B. Zhou, and D. Zhang, "You impress me: Dialogue generation via mutual persona perception," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 1417–1427, Association for Computational Linguistics, July 2020.

[11] C. Tao, L. Mou, D. Zhao, and R. Yan, "Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017.

[14] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, (Minneapolis, Minnesota), pp. 48–53, Association for Computational Linguistics, June 2019.

[15] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[16] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, pp. 3104–3112, 2014.