Response Generation based on the Big Five Personality Traits

Wanqi WU† and Tetsuya SAKAI†

† Department of Computer Science and Engineering, Waseda University
3–4–1 Okubo, Shinjuku-ku, Tokyo 169–8555, Japan
E-mail: †wwwangi@moegi.waseda.jp, ††tetsuyasakai@acm.org

Abstract Personality is one's intrinsic property that is hard to change and closely tied with one's natural language expression, so it is a potential key to make a dialogue system consistent. While there exist many studies on generating responses conditioned on certain profiles, little work has been done on incorporating real personalities in dialogue systems. In this study, we firstly constructed a personality identifier by using the myPersonality dataset for detecting trait values from utterances. We then used the constructed identifier to label all the conversations in the Cornell Movie-dialogs Corpus. Finally, we used a GRU-based seq2seq model with attention mechanism to generate responses conditioned on a certain personality by including personality information in the decoder part. The personality model used in our experiment is called the Big Five, which contains five personality dimensions named neuroticism, openness, extraversion, agreeableness and conscientiousness. Our experiment shows that when adjusting the trait values along extraversion and neuroticism dimensions, the generated responses can reflect the prescribed personality slightly while for the other three dimensions there is no improvement observed.

Key words chatbot, dialogue system, personalization, natural language generation

1 Introduction

With technological advancement, an increasing number of industries adopt chatbots aiming at automating business processes or providing users with emotional consolation. However, given the fact that most of the existing chatbots tend to give formulaic and general responses without special user targeting, it is still a great challenge to build a personalized dialogue system that can adapt specific personality according to its users.

As one's intrinsic property, personality is hard to change and closely tied with one's natural language expression, which makes it a potential key to form a consistent dialogue system. A lot of previous studies simplified the issue to generate responses based on users' profiles that contain entities like age and location [1] [2] [3]. However, little work has been done on incorporating real personalities in dialogue systems. So far, there have been many personality models proposed in the field of psychology and one of the most widely used model is called the Big Five model [4]. As shown in Table 1, it contains five personality dimensions named neuroticism (NEU), openness (OPN), extraversion (EXT), agreeableness (AGR) and conscientiousness (CON). Several prior studies have found close correlations between these five traits and linguistic behavior via lexical and syntax analysis [5] [6].

In this study, we mainly investigate how to generate responses conditioned on a certain personality based on the Big Five model. Firstly, we constructed a personality trait value identifier by using a sample of the myPersonality dataset for detecting trait values from utterances. The constructed identifier is then used to label all the conversations in the Cornell Movie-dialogs Corpus. Finally, we constructed a GRU-based seq2seq model with attention mechanism to generate personalized responses by including personality information in the decoder part. Our experiment shows that when adjusting the trait values along extraversion and neuroticism dimensions, the generated responses can reflect the prescribed personality slightly while for the other three dimensions there is no improvement observed.

Table 1 Big Five Personality Model

Trait	Meaning
Extraversion	Sociableness, Energy, Talkativeness, Ability to be articulate, Friendliness, Social confidence
Openness	Imagination, Insightfulness, Varied interests, Creativity, Curiosity
Agreeableness	Altruism, Trust, Modesty, Patience, Consideration
Neuroticism	Pessimism, Moodiness, Jealousy, Anxiety, Instability
Conscientiousness	Thoroughness, Self-discipline, Reliability, Perseverance, Planning

2 Related Work

The current mainstream researches on personalized dialogue system focus on generating responses based on certain profiles. Zheng et al. [2] define personality to be key-value pairs where the keys are limited to gender, age and location. They embed, capture and address these explicit information within seq2seq framework by using a trait fusion module. Moreover, Zhang et al. [7] define personality by using a set of descriptive sentences. Not only basic information like gender and age are revealed in natural language expression, other richer information like interests and occupations are also included. They introduce a generative seq2seq model which encodes profile entries as individual memory representations in a memory network. These representations are added into decoder part to generate the next word in the sequence. However, profiles are not real personality. When they tend to capture abstract personality by using a set of limited concrete terms, some crucial implicit clues may be lost.

To better capture real personality rather than profile-based user information, some researchers has made pilot efforts in this direction. Early work about a famous natural language generation (NLG) system called Personage introduce the first systematic framework that integrate real personality by explicitly defining 40 linguistic features as generation parameters based on psychological knowledge [8]. However, it is designed only for generating utterances that vary along the extraversion dimension and it may not be practically feasible to do such great amount of feature engineering work for each dimension. Another work of Li et al. [9] crafts a seq2seq model that trains embedding for each individual. The hidden units of decoder side are obtained by integrating the speaker embedding so that the information of speaker is encoded and injected into the hidden layer[4]. It is an effective way to implicitly capture speaker-specific information, however, their model can only generate responses for the speakers that are involved in the training data and it requires sufficient dialogue data from each speaker to ensure the reliability of the model which is unrealistic in real usage scenarios.

To address the existing problems stated above, we train a seq2seq model that does not need complicated feature engineering work but can learn high-level features automatically through training process based on the Big Five model. Each user is represented by a five-dimensional personality vector so that the model is not limited to imitating talking patterns of users who are in the training dataset but can be controlled by adjusting the values of the personality vector elements.

3 Proposed Method

3 1 Personality Trait Value Identifier

To the best of our knowledge, currently there is no publicly available dialogue corpus that has Big Five personality annotations at the utterance level. Therefore, our first goal is to train a personality trait value identifier to automatically estimate Big Five trait values from utterances and use that annotated dialogue corpus to train the personality-based response generation model.

3 1. 1 Dataset

We use a small sample of the myPersonality dataset to build the identifier(注1). It was collected from a Facebook App that allowed its users to participate in psychological research by filling in a personality questionnaire and it is the only available dataset that contains both user's written sentences and the corresponding five traits' values.

For pre-processing, we converted all the letters to lowercase, tokenized the sentences, removed the punctuations, numbers, stop-words and posts that have word lengths smaller than 3 and maintain only the words that have a vector representation in word2Vec [10]. The basic statistics of myPersonality sample dataset are shown in Table 2 and Table 3. Moreover, by observing the trait value distributions for five personality dimensions illustrated in Figue 1, we found the data is unbalanced, so we tried to apply SMOTE [11] over-sampling algorithm to deal with this problem.

Table 2 Statistics of the MyPersonality Sample Dataset in terms of Users, Posts and Words

	count
Total Users	156
Total posts	9,916
Total posts after pre-processing	8,540
Total words	140,664
Total words after pre-processing	71,681
Unique words	16,575
Unique words after pre-processing	12,219

Table 3 Statistics of the MyPersonality Sample Dataset in terms of Post Lengths

	Lowest	Average	Highest
Word per post after pre-processing	1	8	45

(注1) : https://github.com/Myoungs/myPersonality-dataset (The authors stopped maintaining the original full dataset since 2018)

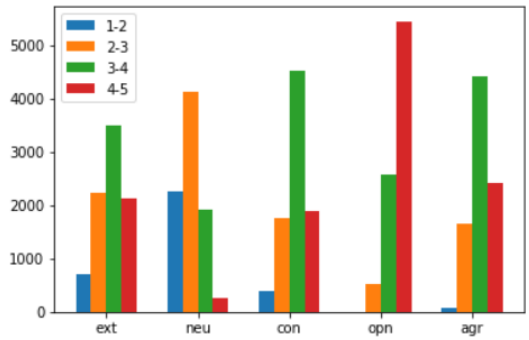

Figure 1 Trait Value Distributions of Users in the MyPersonality Sample Dataset For All Five Personality Dimensions

3.1.2 Construction of Identifier

We treat the construction of identifier as a regression problem and construct five models for five dimensions respectively. In this study, we applied and modified four methods that are proposed in previous studies and compared their performances [12] [13] [14].

Method 1 We average the word embedding representations into a single vector and feed it to a Gaussian Processes model (GP) for training and testing.

Method 2 We represent one sentence with concatenation of maximum, minimum and average values of word embedding vectors and use SVM for prediction. The resultant vector dimensionality for each sentence is 900 (using word2Vec) and grid search is applied to find out the proper hyper-parameters for the model.

Method 3 We aggregate word vectors into sentence vectors and make predictions by using a Convolution Neural Network. More specifically, we extract the n-gram features by applying convolutional filters for $n = 1, 2, 3$ and then concatenate these three vectors to obtain the sentence representation for prediction.

Method 4 We further feed the feature vector obtained from the last hidden layer of CNN to SVM to get the estimated trait values.

3.2 Personality-based Response Generation Model

3.2.1 Dataset

The Cornell Movie-dialogs Corpus(注2) is used to train the chatbot. It contains 220,579 conversational exchanges between 10,292 pairs of movie characters extracted from raw movie scripts. Each utterance in the corpus is annotated with five personality trait values by using the constructed identifier.

3.2.2 Construction of Model

As shown in Figure 2, the model is built within seq2seq model with attention mechanism. The encoder is a 2-layer bidirectional GRU and the decoder is a 2-layer unidirectional

(注2) : https://www.kaggle.com/rajathmc/cornell-moviedialog-corpus

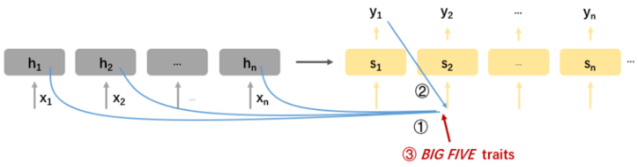

Figure 2 Structure of the Personality-based Response Generation Model

GRU with greedy search. The 5-dimension personality vector is linearly expanded to a 256-dimension vector which is then concatenated with the context vector and the internal hidden state vector to generate the next word in the sequence.

4 Experiments

4.1 Experiment Setup

4.1.1 Trait Value Identifier

For Gaussian Processes, the kernel function is set to be the combination of DotProduct and WhiteKernel. The data is split for training and testing using a 10 Fold Cross-Validation. For SVM, the kernels for all the five dimensions are set to be radial basis function (rbf). Values for the two hyper-parameters, C and *gamma*, that decide the performance of an SVM model are listed in Table 4. C is the penalty parameter that controls the cost of misclassification, while *gamma* adjusts the curvature of the decision boundary.

Table 4 Hyper-parameters of the SVM Used in our Experiment

Hyper-parameter	EXT	NEU	CON	OPN	AGR
C	10	10	10	1	10
Gamma	1	10	1	1	1

For CNN, we apply Relu as the activation function and set the drop ratio as 0.25 at every layer of the neural network.

4.1.2 Response Generation Model

The seq2seq model is constructed using Pytorch and the hyper-parameters are listed in the Table 5.

Table 5 Hyper-parameters Used In the Experiment

Hidden Size	256	Learning Rate	0.0001
Dropout	0.1	Decoder Learning Ratio	5.0
Batch Size	64	Teaching Forcing Ratio	1.0

4.2 Results and Discussion

4.2.1 Trait Value Identifier

Table 6 shows the mean-square error (MSE) of four methods for constructing the trait value identifier, where MSE is defined as the average squared difference between the estimated and actual values. Table 7 shows the statistically significant pairs of four methods obtained by tukey HSD test. In this experiment, the significance level is set to be

0.05. Overall, for EXT, NEU, AGR and OPN dimensions, CNN+SVM method performs the worst among the four and the other three methods have comparable results with subtle differences. For CON dimension, the results show that CNN performs statistically significantly better than SVM, but we cannot distinguish which method among the four performs the best. For current stage, since GP has similar results with SVM but requires a relatively longer time for training and the mse values of CNN change significantly every time we retrain the model due to the high dependency on the choice of initial weight values, we finally choose SVM to construct the five identifiers. Moreover, we apply SMOTE to balance the training data but it results in no difference.

Table 6 MSE of Four Methods For Trait Value Identifier

Traits	GP	SVM	CNN	CNN+SVM
EXT	0.72	0.72	**0.69**	0.82
NEU	**0.58**	**0.58**	**0.58**	0.76
CON	0.53	0.56	**0.52**	0.56
OPN	0.37	**0.36**	0.37	0.41
AGR	0.46	0.45	**0.44**	0.51

Table 7 Statistically Significant Pairs of Four Methods Obtained By Tukey HSD Test

Traits	Model Pair	Diff	P-value
EXT	CNNSVM-GP	0.101	0.00010
	CNNSVM-SVM	0.106	0.00003
	CNNSVM-CNN	0.131	0
NEU	CNNSVM-GP	0.186	0
	CNNSVM-SVM	0.180	0
	CNNSVM-CNN	0.185	0
AGR	CNNSVM-GP	0.053	0.00043
	CNNSVM-SVM	0.061	0.00003
	CNNSVM-CNN	0.066	0
OPN	CNNSVM-GP	0.042	0.00003
	CNNSVM-SVM	0.047	0
	CNNSVM-CNN	0.039	0.00015
CON	SVM-CNN	0.040	0.040

Table 8 shows several example results of the constructed trait value identifiers. Given a sentence, identifiers will estimate the values for all 5 personality dimensions of that sentence ranging from 1 to 5.

For inference, the EXT values are estimated to be 3.77 and 3.30, which are relatively high, for the first sentence "let us go for a picnic!" and the third sentence "I like this place! You know, it is full of flowers!", respectively. It is consistent with public conventional cognition that an extroverted person is more talkative, use more interjections that show passion and is willing to make invitations. Moreover, for the last three sentences, the NEU values are estimated to be high (i.e. 3.90). It also meets the common sense that an emotional person will be more likely to say negative words like "boo", "disappointed", "nightmares" and "hurts".

Table 8 Example Results of The Trait Value Identifier

Sentence	EXT	OPN	CON	AGR	NEU
Let us go for a picnic!	3.77	4.34	4.29	3.55	2.60
Never mind.	2.66	4.33	3.49	3.76	2.60
I like this place! You know, it is full of flowers!	3.30	4.65	3.88	3.95	2.60
10 hour workday? Boo.	2.55	4.15	2.90	3.70	3.90
is so disappointed...why do I always have to be right about these things?	2.09	4.35	3.97	3.65	3.90
I got exactly what I asked for and it hurts. I forgot that nightmares are dreams; too.	2.10	4.17	3.90	3.65	3.90

In summary, the constructed identifiers can estimate reasonable values for EXT and NEU dimensions to some extent while those two values have the tendency to be complementary. However, they tend to approach certain values (e.g. 2.6 or 3.9 for NEU dimension). The same problem exists for the other three dimensions that OPN is always above 4.0, while AGR values fall in the range from 3.5 to 4.0. This problem may be caused by the small size and the imbalance of the training dataset. Since the dataset contains the personality test results of only 156 people, it may not be representative enough to cover all the values ranging from 1 to 5 in each dimension. Another problem is that the meaning of values for OPN, CON and AGR dimensions are not explainable by directly looking at the estimated values and the sentences which requires further investigation.

4 2. 2 Response Generation Model

By using the Cornell-Movie Corpus labeled by the identifiers, we train a chatbot and set the prescribed personality by adjusting the trait values along five personality dimensions. In the experiment, we mainly focus on the performance of the chatbot on two dimensions: extraversion and neuroticism where the identifiers work well. When adjusting the trait values, we fix OPN, CON and AGR values to be 4,3,3, which are the majority values of each dimension to ensure low influence of values of those dimensions on the results.

Tables 9 and 10 show some example responses generated by our chatbot. When we set the trait values to be [2,4,3,3,4] where corresponding dimensions are EXT, OPN, CON, AGR and NEU, the chatbot reveals a negative attitude and is quite emotional that most generated responses contain the word "not" since the prescribed NEU value is high. When the trait value is set to be [4,4,3,3,2] with higher EXT value, the chatbot tends to generate positive responses with more interjections that show passion like "Oh my God" and "Yes!".

However, we notice that responses generated by the standard seq2seq model is more fluent and meaningful than responses generated by the model in which personality is integrated. This may be caused by the fact that when personality vector influences the generation of sentences, the model has a narrower selection of words at each generation step.

Table 9 Example Responses For Trait Value [2,4,3,3,4]

User	**Bot (without personality)**	**Bot (with personality)**
I'm a bit sad.	You are a good man.	You're not gonna make it.
Happy new year!	Yeah.	I'm not asking.
Who do you like the best?	The girl.	I'm not.
Let's talk about your hobby.	What?	I'm not here.
I like to eat fish.	Really?	I don't know.
My boyfriend left me···	I'm sorry.	I'm not asking you.

Table 10 Example Responses For Trait Value [4,4,3,3,2]

User	**Bot (without personality)**	**Bot (with personality)**
I'm a bit sad.	You are a good man.	I'm going to meet you.
Happy new year!	Yeah.	I know.
Who do you like the best?	The girl.	You know what I mean.
Let's talk about your hobby.	What?	Oh my god.
I like to eat fish.	Really?	Yes!
My boyfriend left me···	I'm sorry.	I'm not.

To better compare the performances of the standard seq2seq model and the personality-based model, we conduct a human evaluation. We randomly select 20 utterances from the test dataset for EXT and NEU dimensions respectively and use these utterances to generate two responses by two models. Six judges (five Chinese CS undergraduate and one Chinese CS master level student from Waseda University) are asked to select one response from the two randomly presented responses they think is more related to the target personality trait.

Table 11 Proportions of successful discrimination for six judges

Trait	**1**	**2**	**3**	**4**	**5**	**6**	**Average**
EXT	0.650	0.650	0.600	0.450	0.650	0.500	0.583
NEU	0.450	0.750	0.550	0.550	0.600	0.500	0.567

The proportions of successful discrimination for six judges are listed in table 11, where the average proportion is obtained by averaging six values for the six judges. As a result, the judges can distinguish 58.3% and 56.7% responses that are generated by the personality-based model for EXT and NEU dimensions respectively on average.

The results suggest that constructing a personality-based response generation model by using neural networks is achievable but far from being satisfactory. Extraversion and neuroticism dimensions are better incorporated in the model on the ground that their training data are more normally distributed and the usage of words are more distinctive for two extremes. Future work should be done to find new approaches so that other three dimensions can be incorporated into the model and clearly shown in the generated responses.

5 Conclusion

In this work, we construct a personality identifier for detecting trait values from utterances and by using it we label all the conversations in the Cornell Movie-dialogs Corpus which is finally utilized to train a personality-based response generation model. Our model can generate personalized responses based on the prescribed Big Five personality trait values of two dimensions: extraversion and neuroticism. For the other three dimensions, no clear improvement is observed due to the small size and imbalance of the myPersonality sample dataset used for constructing the trait value identifiers. For future work, we want to build up a new dataset with a wider coverage that contains users' written sentences and their corresponding Big Five trait values at the utterance level in order to train an identifier with higher accuracy so as to lay a solid foundation for the later construction of the personality-based response generation model. Also, we want to try other corpora that is more related to everyday lives since the conversations in movies are sometimes dramatic and unrealistic. Moreover, we would like to simplify and reconstruct the Big Five personality model by extracting useful definitions that have closer correlations with natural language expression to improve the model performance because some dimensions like conscientiousness may not be explicitly shown in one's daily conversations.

Acknowledgement

The first author would like to express my thank to Mr. Zhaohao Zeng of the Sakai Laboratory for all his help and suggestions on wording and phrasing of this paper.

References

[1] Q. Qian, M. L. Huang, et al., "Assigning personality/profile to a chatting machine for coherent conversation generation," Proceedings of the 27th International Joint Conference on

Artificial Intelligence (IJCAI'18), pp. 4279–4285, 2018.

[2] Y. H. Zheng, G. Y. Chen, et al., "Personalized Dialogue Generation with Diversified Traits," arXiv:1901.09672v1, 2019.

[3] J. Li, M. Galley, et al., "A Persona-Based Neural Conversation Model," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Vol. 1, pp. 994–1003, 2016.

[4] J. M. Digman, "Personality structure: Emergence of the five-factor model," Annual review of psychology, Vol. 41, No. 1, pp. 417–440, 1990.

[5] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker, "Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life," Journal of Personality and Social Psychology, Vol. 90, No. 5, pp. 862–877, 2006.

[6] J. W. Pennebaker and L. A. King, " Linguistic styles: Language use as an individual difference," Journal of Personality and Social Psychology, Vol. 77, No. 6, pp. 1296–1312, 1999.

[7] S. Z. Zhang, E. Dinan, et al., " Personalizing Dialogue Agents: I have a dog, do you have pets too?," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 2204–2213, 2018.

[8] F. Mairesse and M. Walker, " PERSONAGE: Personality Generation for Dialogue," Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp 496–503, 2007.

[9] J. W. Li, M. Galley, et al., " A Persona-Based Neural Conversation Model," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 994–1003, 2016.

[10] T. Mikolov, I. Sutskever, et al., "Distributed Representations of Words and Phrases and their Compositionality," NIPS, 2013.

[11] N. V. Chawla, K. W. Bowyer, et al., "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, Vol. 16, No. 1, pp. 321–357, 2002.

[12] P. H. Arnoux, A. B. Xu, et al., "25 Tweets to Know You: A New Model to Predict Personality with Social Media," ICWSM, 2017.

[13] N. Majumder, S. Poria, et al., "Deep Learning-Based Document Modeling for Personality Detection from Text," IEEE Intelligent Systems, Vol. 32, No. 2, pp. 74–79, 2017.

[14] G. Carducci, G. Rizzo, et al., "TwitPersonality: Computing Personality Traits from Tweets Using Word Embeddings and Supervised Learning," Information, Vol. 9, No. 5, pp. 127–147, 2018.