Neural Response Generation for Customer Service based on Personality Traits

Jonathan Herzig, **Michal Shmueli-Scheuer**, **Tommy Sandbank** and **David Konopnicki**

IBM Research - Haifa
Haifa 31905, Israel
{hjon,shmueli,tommy,davidko}@il.ibm.com

Abstract

We present a neural response generation model that generates responses conditioned on a target personality. The model learns high level features based on the target personality, and uses them to update its hidden state. Our model achieves performance improvements in both perplexity and BLEU scores over a baseline sequence-to-sequence model, and is validated by human judges.

1 Introduction

Automated conversational agents are becoming popular for various tasks, such as personal assistants, shopping assistants, or as customer service agents. Automated agents benefit from adapting their personality according to the task at hand (Reeves and Nass, 1996; Tapus and Mataric, 2008) or to the customer (Herzig et al., 2016). Thus, it is desirable for automated agents to be capable of generating responses that express a target personality.

Personality is defined as a set of traits which represent durable characteristics of a person. Many models of personality exist while the most common one is the *Big Five* model (Digman, 1990) , including: *Openness, Conscientiousness, Extraversion, Agreeableness,* and *Neuroticism*. These traits were correlated with linguistic choices including lexicon and syntax (Mairesse and Walker, 2007).

In this paper we study how to encode personality traits as part of neural response generation for conversational agents. Our approach builds upon a sequence-to-sequence (SEQ2SEQ) architecture (Sutskever et al., 2014) by adding an additional layer that represents the target set of personality traits, and a hidden layer that learns high-level personality based features. The response is then generated conditioned on these features.

Example 1
Customer: *Why isn't your service working???*
Consensus-agent: *We are aware of the issue and are working to fix it.*
Agreeableness-agent: *We're here to help! Are you getting any error messages or codes?*

Example 2
Customer: *You guys messed up my delivery today.*
Consensus-agent: *Please contact us if you don't get it by the end of the week.*
Conscientiousness-agent: *Please email us with your tracking #, details and contact #. We'll check on it.*

Figure 1: Examples of a customer utterance followed by a consensus agent response and a high agreeableness or high conscientiousness response.

Specifically, we focus on conversational agents for customer service; in this context, many studies examined the effect of specific personality traits of human agents on service performance. Results indicate that **conscientiousness** (a person's tendency to act in an organized or thoughtful way) and **agreeableness** (a person's tendency to be compassionate and cooperative toward others) correlate with service quality (Blignaut et al., 2014; Sackett, 2014).

Figure 1 shows examples of customer utterances, followed by two automatically generated responses. The first response (in each example), is generated by a standard SEQ2SEQ response generation system that ignores personality modeling and in effect generates the consensus response of the humans represented in the training data. The second response is generated by our system, and is aimed to generate

Proceedings of The 10th International Natural Language Generation conference, pages 252–256,
Santiago de Compostela, Spain, September 4-7 2017.

data for an agent that expresses a high level of a specific trait. In example 1, the agreeableness-agent is more compassionate (expresses empathy) and is more cooperative (asks questions). In example 2, the conscientiousness-agent is more thoughtful (will "check the issue").

We experimented with a dataset of 87.5K real customer-agent utterance pairs from social media. We find that leveraging personality encoding improves relative performance up to 46% in BLEU score, compared to a baseline SEQ2SEQ model. To our knowledge, this work is the first to train a neural response generation model that encodes target personality traits.

2 Related Work

Generating responses that express a target personality was previously discussed in different settings. Early work on the PERSONAGE system (Mairesse and Walker, 2007; Mairesse and Walker, 2008; Mairesse and Walker, 2010; Mairesse and Walker, 2011) presented a framework projecting different traits throughout the different modules of an NLG system. The authors explicitly defined 40 linguistic features as generation parameters, and then learned how to weigh them to generate a desired set of traits. While we aim at the same objective, our methodology is different and does not require feature engineering. Our approach utilizes a neural network that automatically learns to represent high level personality based features.

Neural response generation models (Vinyals and Le, 2015; Shang et al., 2015) are based on a SEQ2SEQ architecture (Sutskever et al., 2014) and employ an encoder to represent the user utterance and an attention-based decoder that generates the agent response one token at a time. Models that aim to generate a coherent persona also exist. Li et al. (2016) modified a SEQ2SEQ model to encode a persona (the character of an artificial agent). The main difference with our work is that we focus on modeling the expression of specific personality traits and not an abstract character. Moreover, their persona-based model can only generate responses for the agents that appear in the training data, while our model has no such restriction. Finally, Xu et al. (2017) generated responses for customer service requests on social media using standard SEQ2SEQ, while we modify it to generate a target personality.

3 Sequence-to-Sequence Setup

We review the SEQ2SEQ attention based model on which our model is based.

Neural response generation can be viewed as a sequence-to-sequence problem (Sutskever et al., 2014), where a sequence of input language tokens $x = x_1, \ldots, x_m$, describing the user utterance, is mapped to a sequence of output language tokens $y_1, \ldots, y_n$, describing the agent response.

The **encoder** is an LSTM (Hochreiter and Schmidhuber, 1997) unit that converts $x_1, \ldots, x_m$ into a sequence of context sensitive embeddings $b_1, \ldots, b_m$. An attention-based **decoder** (Bahdanau et al., 2015; Luong et al., 2015) generates output tokens one at a time. At each time step j, it generates y_j based on the current hidden state s_j, then updates the hidden state s_{j+1} based on s_j and y_j. Formally, the decoder is defined by the following equations:

$$s_1 = \tanh(W^{(s)} b_m), \tag{1}$$

$$p(y_j = w \mid x, y_{1:j-1}) \propto \exp(U[s_j, c_j]), \tag{2}$$

$$s_{j+1} = LSTM([\phi^{(out)}(y_j), c_j], s_j), \tag{3}$$

where $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, n\}$ and the context vector, c_j, is the result of global attention (see (Luong et al., 2015)). The matrices $W^{(s)}$, $W^{(a)}$, U, and the embedding function $\phi^{(out)}$ are decoder parameters. The entire model is trained end-to-end by maximizing $p(y \mid x) = \prod_{j=1}^{n} p(y_j \mid x, y_{1:j-1})$.

4 Personality Generation Model

The model described in section 3 generates responses with maximum likelihood which reflect the consensus of the agents that appear in the training data. This kind of response does not characterize a specific personality and thus can result in inconsistent or unwanted personality cues. In this section we present our PERSONALITY-BASED model (Figure 2) which generates responses conditioned on a target set of personality traits values which the responses should express. The target set of personality traits is represented as a vector p, where p_i represents the desired value for the i^{th} trait. This value encodes

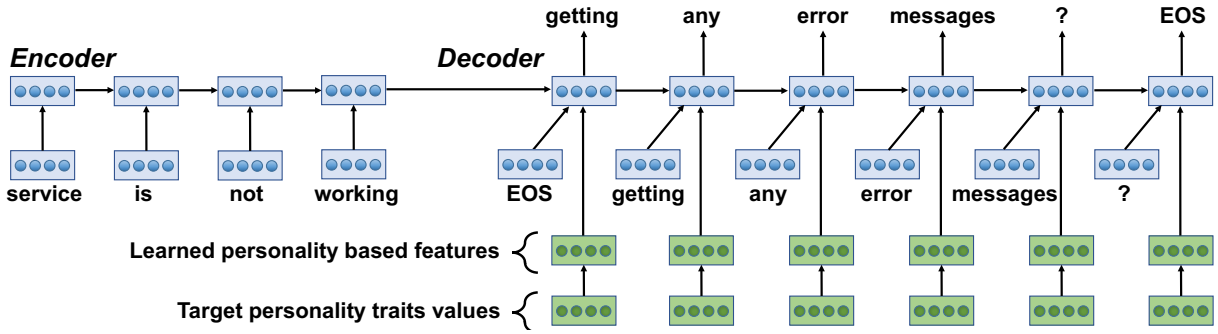

Figure 2: Architecture for the personality based generation model.

how strongly should this trait be expressed in the response. Consequently, the size of p depends on the selected personality model (e.g., five traits for the Big Five model).

As in (Mairesse and Walker, 2011), we argue that personality traits are exhibited as different types of stylistic linguistic variation. Thus, our model's response is conditioned on generation parameters which are based on personality traits. In comparison to (Mairesse and Walker, 2011) where generation parameters were defined manually, we learn these high-level features automatically during training. We introduce a personality based features hidden layer $h_p = \sigma(W^{(p)}p + b)$, where $W^{(p)}$ and b are parameters learned by the model during training. Each personality feature h_i is a weighted sum of the targeted traits values (following a sigmoid activation). Now, at each token generation, the decoder updates the hidden state conditioned on the personality traits features h_p, as well as on the previous hidden state, the output token and the context. Formally, Equation 3 is changed to:

$$s_{j+1} = LSTM([\phi^{(out)}(y_j), c_j, h_p], s_j), \quad (4)$$

Conditioning on h_p captures the relation of text generation to the underlining personality traits.

5 Experiments

Data. Our model is designed to generate text conditioned on a target set of personality traits. Specifically, we verified its performance in a scenario of customer service. For our experiments we utilized the dataset presented in (Xu et al., 2017), which exhibits a large variety of customer service properties. This dataset is a collection of $1M$ conversations over customer service Twitter channels of 62 different brands which cover a large variety of product categories. Several preprocessing steps were performed for our purposes:

We first split the data to pairs consisting of a single customer utterance and its corresponding agent response. We removed pairs containing non-English sentences. We further removed pairs for agents that participated in less than 30 conversation pairs, so we would have sufficient data for each agent to extract their personality traits (see below). This resulted in $87.5K$ conversation pairs in total including 633 different agents (138 ± 160 pairs per agent on average).

Following (Sordoni et al., 2015; Li et al., 2016) we used BLEU (Papineni et al., 2002) for evaluation. Besides BLEU scores, we also report perplexity as an indicator of model capability. For implementation details, refer to Appendix A.

Results. We experimented with two different settings to measure our model's performance.

Warm Start: In the first experiment, data for each agent in the dataset was split between training, validation and test data sets with a fraction of $80\%/10\%/10\%$, respectively. We then extracted the agents' personality traits using an external service (described in Appendix B), from the training data for each agent. These personality traits values are then used during the model training as the values for the personality vector p. In this setting, since all the agents that appear in the test data appear also in the training data, we can also test the performance of (Li et al., 2016), which learns a persona vector for each agent in the training data.

The results in table 1 show that the standard SEQ2SEQ model achieved the lowest performance in terms of both perplexity and BLEU score while the competing models which learn a representation

Model	Perplexity	BLEU
SEQ2SEQ	11.49	6.3%
PERSONA-BASED (Li et al., 2016)	9.25	15.55%
PERSONALITY-BASED	9.62	12.46%

Table 1: Warm start performance.

for the agents achieved higher performance. The PERSONA-BASED model achieved similar perplexity but higher BLEU score than our model. This is reasonable since PERSONA-BASED is not restricted to personality based features. However, this model can not generate content for agents which do not appear in the training data, and thus, it is limited.

Cold Start: In our second experiment, we split the dataset such that 10% of the agents only formed the validation and test sets (half of each agent's examples for each set). Data for the other 90% of the agents formed the training set.

In this setting, data for agents in the test set does not appear in the training set. These agents represent new personality distributions we would like to generate responses for. Note that, we extracted target personality traits for agents in the training set using their training data, or, for agents in the test set, using validation data. In this setting, it is not possible to test the PERSONA-BASED model since no representation is learned during training for agents in the test set. Thus, we only compare our model to the baseline SEQ2SEQ model. Table 2 shows that, in this setting, we get better performance by utilizing personality based representation: our model achieves a relative 6.7% decrease in perplexity, and a 46% relative improvement in BLEU score. Results from both experiments demonstrate that we can better model the linguistic variation in agent responses by conditioning on target personality traits.

Human Evaluation. We conducted a human evaluation of our PERSONALITY-BASED model using a crowd-sourcing service. This evaluation measures whether the responses generated by our model are correlated with the target personality traits. We focused on two personality traits from the *Big Five* model that are important to customer service: *agreeableness* and *conscientiousness* (Blignaut et al., 2014; Sackett, 2014). We extracted 60 customer utterances from the validation set of the **cold start** setting described above. We selected customer utterances that convey a negative sentiment, since responses to this kind of utterances vary much. After sentences were selected, we generated corresponding agent responses in the following way. We generated a high-*trait* target personality distribution (*trait* was either *agreeableness* or *conscientiousness*), where *trait* was set to a value of 0.9, and all other traits to 0.5. Similarly, we created a low-*trait* version where *trait* was set to 0.1. For each trait and customer utterance we generated a response for the high-*trait* and low-*trait* versions.

Model	Perplexity	BLEU
SEQ2SEQ	21.04	3.19%
PERSONALITY-BASED	19.64	4.67%

Table 2: Cold start performance (agents in the test data do not appear in the training data).

Each triplet (a customer utterance followed by high-*trait* and low-*trait* generated responses) was evaluated by five master level judges. To get the judges familiar with personality traits, we first presented clear definitions of the two traits, followed by several examples (from the task's domain), and explanation. Following Li et al. (2016) methodology, the two responses were presented in a random order, and judged on a 5-point zero-sum scale. A score of 2 (-2) was assigned if one response was judged to express the trait more (less) than the other response, and 1 (-1) if one response expressed the trait "somewhat" more (less) than the other. Ties were assigned a score of zero.

The judges rated each pair, and their scores were averaged and mapped into 5 equal-width bins. After discarding ties, we found that the high-*trait* responses generated by our PERSONALITY-BASED model were judged either more expressive or somewhat more expressive than the low-*trait* corresponding responses in 61% of cases. If we ignore the somewhat more expressive judgments, the high-*trait* responses win in 17% of cases.

6 Conclusions and Future Work

We have presented a personality-based response generation model and tested it in customer care tasks, outperforming baseline SEQ2SEQ model. In future work, we would like to generate responses adapted to the personality traits of the customer as well, and to apply our model to other tasks such as education systems.

References

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Linda Blignaut, Leona Ungerer, and Helene Muller. 2014. Personality as predictor of customer service centre agent performance in the banking industry: An exploratory study. *SA Journal of Human Resource Management*, 12(1).

John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, and Anat Rafaeli. 2016. Predicting customer satisfaction in customer support conversations in social media using affective features. In *UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*, pages 115–119.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. A persona-based neural conversation model. In *ACL*.

M. Luong, H. Pham, and C. D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421.

Franois Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue. In *ACL*, pages 496–503.

François Mairesse and Marilyn A Walker. 2008. Trainable generation of big-five personality styles through data-driven parameter estimation. In *ACL*, pages 165–173.

François Mairesse and Marilyn A. Walker. 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Model. User-Adapt. Interact.*, 20(3):227–278.

François Mairesse and Marilyn A. Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Byron Reeves and Clifford Nass. 1996. How people treat computers, television, and new media like real people and places. *CSLI Publications and Cambridge*.

Walmsley P. T. Sackett, P. R. 2014. Which personality attributes are most important in the workplace? *Perspectives on Psychological Science*, 9(5):538–551.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *CoRR*, abs/1506.06714.

I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Adriana Tapus and Maja J Mataric. 2008. Socially assistive robots: The link between personality, empathy, physiological signals, and task performance. In *AAAI Spring Symposium: Emotion, Personality, and Social Behavior*, pages 133–140.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media.

A Implementation Details

We tuned hyper-parameters based on validation set perplexity for both the baseline SEQ2SEQ and our PERSONALITY-BASED models. We used an LSTM with 800 hidden cells, and a personality based layer with 40 hidden cells. We trained the model for 15 epochs with an initial learning rate of 0.1, and halved the learning rate every epoch, starting from epoch 7. After training was finished we picked the best model according to validation set perplexity. We initialized parameters by sampling from the uniform distribution $[-0.1, 0.1]$. The log likelihood of the correct response was maximized using stochastic gradient descent with a batch size set to 64, and gradients were clipped with a threshold of 5. Vocabulary size is limited to $50,000$. Dropout rate is set to 0.2. At test time, we used beam search with beam size 5. All models were implemented in Torch.

B Personality Traits Detection

To extract personality traits for agents in our experiments we utilized the IBM Personality Insights service, which is publicly available. This service infers three models of personality traits, namely, *Big Five*, *Needs* and *Values* from social media text. It extracts percentile scores for 52 traits[1].

[1] www.ibm.com/watson/developercloud/doc/personality-insights/models.html