UvA-DARE (Digital Academic Repository)

Automatic Evaluation of Neural Personality-based Chatbots

Xing, Y.; Fernández, R.

Link to publication

Citation for published version (APA):

Automatic Evaluation of Neural Personality-based Chatbots

Yujie Xing and **Raquel Fernández**
Institute for Logic, Language and Computation
University of Amsterdam
`yujie.xing@outlook.com` `raquel.fernandez@uva.nl`

Abstract

Stylistic variation is critical to render the utterances generated by conversational agents natural and engaging. In this paper, we focus on sequence-to-sequence models for open-domain dialogue response generation and propose a new method to evaluate the extent to which such models are able to generate responses that reflect different personality traits.

1 Introduction

The advent of deep learning methods has led to the development of data-driven conversational agents for informal open-domain dialogue (see Serban et al., 2016a, for a review). These chatbot systems model conversation as a sequence-to-sequence (SEQ2SEQ) problem (Sutskever et al., 2014) and rely on large amounts of unannotated dialogue data for training. We investigate whether such models are able to generate responses that reflect different personality traits. We test two kinds of models: The speaker-based model by Li et al. (2016b), where response generation is conditioned on the individual speaker, and a personality-based model similar to Herzig et al. (2017), where generation is conditioned on a personality type.

Evaluating the output of chatbot systems is remarkably difficult (Liu et al., 2016). To make progress in this direction with regards to personality aspects, we propose a new statistical evaluation method that leverages an existing personality recogniser (Mairesse et al., 2007), thus avoiding the need for specialised corpora or manual annotations. We adopt the Big Five psychological model of personality (Norman, 1963), also called OCEAN for the initials of the five personality traits considered: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroti-

cism. Each of the traits is represented by a scalar value on a scale from 1 to 7.

In the remainder of the paper, we introduce the models we examine and describe our new evaluation method. Our results show that the models are able to generate output that reflects distinct personalities, over a baseline encoding chance personality variation. We conclude with a brief discussion on related work.

2 Dialogue Generation Models

The generation models we make use of are standard SEQ2SEQ models consisting of an encoder LSTM, an attention mechanism, and a decoder LSTM (Sutskever et al., 2014; Bahdanau et al., 2015). The model processes context-response pairs, where the context $X = x_1, x_2, \ldots, x_m$ corresponds to the latest utterance(s) in the dialogue and the response $Y = y_1, y_2, \ldots, y_n$ is the utterance to be generated next. The probability of the response Y given the context X is predicted as:

$$p(Y|X) = \prod_{t=1}^{n} p(y_t|y_1, \ldots, y_{t-1}, X) \quad (1)$$

The attention mechanism by Yao et al. (2015) is used over the hidden states of the encoder LSTM to generate a context vector c_t that determines the relative importance of the words in the context utterance at each decoding step t. Then the probability of each word w_k ($k \in |V|$, where V is the vocabulary) to be the next word at step t is predicted with a softmax function:

$$P_t(w_k) = \frac{exp((W)_k \cdot f(c_t, h_t))}{\sum_{k=1}^{|V|} exp((W)_k \cdot f(c_t, h_t))} \quad (2)$$

where h_t is the hidden state of the decoder LSTM and f is an activation function. The weights of matrix $W \in \mathbb{R}^{|V| \times d}$ are learned during training, with d being the number of hidden cells.

Both the Speaker Model and the Personality Model we describe below include 4-layer LSTMs with 1024 hidden cells per layer.

189

2.1 Speaker Model

Our starting point is the persona-based model by Li et al. (2016b).[1] In this model, each speaker is associated with an embedding $\boldsymbol{v}_s$ learned during training. Whenever a response by speaker s is encountered during training, the corresponding embedding $\boldsymbol{v}_s$ is inserted into the first hidden layer of the decoder LSTM at each time step (i.e., conditioning each word in the utterance). The hidden states h_t of the decoder LSTM is thus calculated as follows (where y_t^* is the embedding of the response word at time t, and g stands for the LSTM cell operations):

$$h_t = g(h_{t-1}, y_t^*, c_{t-1}, \boldsymbol{v}_s) \qquad (3)$$

Li et al. (2016b) evaluated their model regarding individual content (factual) consistency. Our goal is to evaluate whether the model preserves individual stylistic aspects related to personality traits.

2.2 Personality Model

We modify the Speaker Model to allow for the generation of responses reflecting different personality types. To this end, instead of leveraging speaker embeddings, we estimate the OCEAN scores for each speaker and insert a personality embedding $\boldsymbol{v}_o$ into the first layer of the LSTM decoder.[2]

OCEAN scores are 5-dimensional vectors o, where each dimension ranges from 1 to 7. We normalise them to the range $[-1, 1]$ and then embed them with a linear layer: $\boldsymbol{v}_o = W_o \cdot \frac{o-4}{3}$, where $W_o \in \mathbb{R}^{5 \times d}$ is learned during training, thus learning relationships between OCEAN trait values and properties of the utterances. Whenever a response with personality traits o is encountered, we insert $\boldsymbol{v}_o$ into the first hidden layer of the decoder LSTM. Thus, the hidden states h_t are now calculated as:

$$h_t = g(h_{t-1}, y_t^*, c_{t-1}, \boldsymbol{v}_o) \qquad (4)$$

This version of the model is similar to Herzig et al. (2017).[3] The authors focus on the customer service domain and evaluate the model output's style

for only two personality traits with human evaluation. In contrast, we deal with open-domain chat and assess all OCEAN traits globally, using the automatic method we describe Section 4.

3 Experimental Setup

3.1 Dataset

We use transcripts from two American situation comedy television series: *Friends*[4] and *The Big Bang Theory*.[5] We consider only those characters who contribute a minimum of 2000 turns, which results in 13 characters (6 from *Friends* and 7 from *The Big Bang Theory*). We assign a unique speaker id to each character. In addition, we estimate the personality of each character as follows: for each character, we randomly select 50 samples of 500 utterances each, and estimate the OCEAN scores for each sample using the personality recogniser by Mairesse et al. (2007), which exploits linguistic features from 'Linguistic Inquiry and Word Count' (Pennebaker and King, 1999) and the MRC Psycholinguistic database (Coltheart, 1981).[6] We assign to each character the OCEAN score resulting from taking the arithmetic mean of the estimated scores for the corresponding 50 samples.

We consider every two consecutive turns in a scene to be a context-response pair and annotate each response with either the speaker id or the speaker's OCEAN scores. The resulting dataset contains ∼86k context-response pairs, of which around 2000 pairs were randomly selected and reserved for validation.

3.2 Training

Given the relatively small size of the TV-series dataset, following Li et al. (2016b) we use the OpenSubtitles dataset (Tiedemann, 2009) to pretrain the model. OpenSubtitles is a large open-domain repository containing over 50M lines from movie subtitles. Since this data does not include information on which character is the speaker of each line, we simply take each two consecutive lines to be a context-response pair. Due to limitations regarding computational power, we lever-

---

[1]See http://github.com/jiweil/Neural-Dialogue-Generation. We reimplemented the model in PyTorch.

[2]The procedure for assigning OCEAN scores to a given speaker is explained in the next section.

[3]Our personality model is a modified version of our reimplementation of the code by Li et al. (2016b) (see footnote 1). The code by Herzig et al. (2017) is not readily available.

[4]https://sites.google.com/site/friendstvcorpus/

[5]https://bigbangtrans.wordpress.com/

[6]We choose this recogniser because it can estimate numerical scores for each OCEAN trait, instead of binary classifications, and it's open source. For more details, we refer the reader to Mairesse et al. (2007).

age only a subset of the dataset: $\sim$1.8M pairs for training and $\sim$75k pairs for validation.

We train a standard SEQ2SEQ model for 15 iterations on the OpenSubtitles training set, until perplexity becomes stable in the validation set. We then initialise the Speaker and Personality models using the parameters learned with OpenSubtitles and train them on the TV-series training set for 30 more iterations, until the perplexity in the corresponding validation set stabilises.

We use the same settings as Li et al. (2016b) for training: We set the batch size to 128, the learning rate to 1.0 (halved after the 6th iteration), the threshold for clipping gradients to 5, and the dropout rate to 0.2. Vocabulary size is 25,000 and the maximum length of an input sentence is 50. All parameters (including the speaker embeddings in the Speaker Model) are initialised sampling from the uniform distribution on $[-0.1, 0.1]$.

3.3 Testing

For testing, we again leverage OpenSubtitles to extract a large subset of $\sim$2.5M utterances not present in the training or validation sets. Using each of the utterances in this set as context, we let the trained Speaker and Personality models generate responses for each of the 13 characters, employing Stochastic Greedy Sampling (Li et al., 2017). Since general responses are a known problem in neural response generation chatbots (Sordoni et al., 2015; Serban et al., 2016b; Li et al., 2016a; Zhang et al., 2018) and our goal is to focus on personality-related stylistic differences, we remove the most frequent 100 responses common to all characters/personalities. After this cleaning step, we end up with $\sim$700k responses per character/personality. We refer to the clean set of generated responses as the evaluation set.

4 Evaluation Method

We propose a new evaluation method to measure whether persona-based neural dialogue generation models are able to produce responses with distinguishable personality traits for different characters and different personality types.

Using the evaluation set, for each character we randomly select 250 samples of 500 responses and calculate the OCEAN scores for each sample. Recall that the OCEAN scores correspond to 5-dimensional vectors. We label each of these 250 vectors with the corresponding character. This gives us 13 gold classes—one for each character—with 250 datapoints each. We then use a support vector machine classifier[7] to test to what extent the OCEAN scores estimated from the generated responses allow us to recover the gold character classes. We compute results using 5-fold cross-validation (training on 80% of the set and testing on the remaining 20% once for each fold). We report average scores over ten iterations of this procedure (i.e., 5×10).

We consider a baseline obtained by randomising the gold character label in the set of generated responses, which indicates the level of performance we may expect by chance. In addition, we use the procedure described above to discriminate between characters using their original (gold) utterances from the transcripts, rather than model-generated responses. This serves as a sanity check for the personality recogniser used to estimate the OCEAN scores—if the recogniser cannot detect personality differences among the characters in the original transcripts, it is not reasonable to expect that the models will be able to generate responses with different personality styles—and provides an upper bound for the performance we can expect to achieve when evaluating generated responses.

Given that the particular personality recogniser we use (Mairesse et al., 2007) was not optimised for dialogues from TV-series transcripts, as an additional sanity check we compare its performance on the original (gold) utterances with a bag-of-words (BoW) approach. This allows us to test whether the recogniser may only be detecting trivial patterns of word usage.[8] We select the top 200 most frequent words over the original utterances as features, without removing words typically considered stop words such as pronouns or discourse markers, since they may be personality indicators. Then we run the same classification procedure using these BoW representations.

5 Results

In Table 1, we report average F1 score per character (including precision and recall) for the Speaker

Friends	Precision	Recall	F1	*Big Bang Theory*	Precision	Recall	F1
Baseline	0.16 (σ=.01)	0.16 (σ=.01)	0.16	Baseline	0.15 (σ=.01)	0.15 (σ=.02)	0.15
Gold	0.61 (σ=.12)	0.61 (σ=.16)	0.61	Gold	0.69 (σ=.11)	0.69 (σ=.16)	0.69
Speaker	0.32 (σ=.02)	0.32 (σ=.05)	0.32	Speaker	0.46 (σ=.20)	0.47 (σ=.23)	0.47
Personality	0.22 (σ=.04)	0.23 (σ=.09)	0.23	Personality	0.29 (σ=.19)	0.31 (σ=.24)	0.30

Table 1: Average scores for 6 characters in *Friends* (left) and 7 characters in *The Big Bang Theory* (right)

and the Personality models, as well as the baseline and gold data. The results for these four conditions are all statistically significantly different from each other.[9]

5.1 Lower and Upper Bounds

The first thing to note is that the results on the gold transcripts are higher than the baseline, reaching 61% F1 score on *Friends* and 69% on *The Big Bang Theory*. This indicates that the evaluation method is able to distinguish between the different personalities in the data reasonably well. Apparently, *The Big Bang Theory* characters are more distinct from each other than those in *Friends*.

When we use the BoW approach on the gold transcripts instead of the representations by the personality recogniser, we obtain significantly lower results: 23% F1 score on *Friends* and 19% on *The Big Bang Theory*.[10] The personality recogniser thus detects patterns that go beyond what can be captured with BoW representations.

5.2 Speaker and Personality Models

We find that the responses generated by the Speaker model display consistent personality variation above baseline, although a significant level of the personality markers found in the original data seems to be lost (32% vs. 61% and 47% vs. 69%). The results obtained for the Personality model are significantly above baseline as well (23% vs. 16% and 30% vs. 15%). We also see that the personality traits found in the responses generated by the Personality model yield lower distinguishability than those by the Speaker model. This is to be expected, since the Personality model generates responses for a personality type, which

should be more varied (and hence less distinguishable) than those by an individual speaker.

An advantage of the Personality model, however, is that in principle it allows us to generate responses for novel, predefined personalities that have not been seen during training. To test this potential, we create five extreme personality types by setting up the score of one of the OCEAN traits to a high value (6.5) and all remaining four traits to an average value (3.5). We then let the model generate responses to all the utterances in the evaluation set for each of these extreme personalities and evaluate the extent to which the responses differ in style following the same procedure as in the previous experiment. Table 2 shows the results.

	Precision	Recall	F1
Baseline	0.19	0.19	0.19
Average	0.53 (σ=.07)	0.53 (σ=.09)	0.53
Open	0.46	0.46	0.46
Conscientious	0.59	0.62	0.61
Extravert	0.63	0.65	0.64
Agreeable	0.53	0.50	0.51
Neurotic	0.44	0.42	0.43

Table 2: Average scores for personality types with high value for different OCEAN personality traits

We find that the generated responses are distinguishable with 53% average F1 score. This indicates that the model has learned to generalise beyond the training data. Table 3 shows some examples of generated responses.

Joey (*Friends*):	*Oh, of course I love you, baby.*
Raj (*Big Bang*):	*I don't love you.*
Open:	*You are beautiful!*
Agreeable:	*Oh I, I love you too.*

Table 3: Responses to *Do you love me?* by the Personality model for personality types of given characters and extreme types not seen during training

---

[9] Significance is tested with a two-independent-sample t-test on the results of 10 iterations, first using Levene's test to assess the equality of variances and then applying Welch's or Student's t-test accordingly.

[10] We also run this experiment removing stop words (using the list of English stop words from `scikit-learn`), obtaining almost identical results: 22% F1 score on *Friends* and 18% on *The Big Bang Theory*.

6 Related Work and Conclusion

In recent years, there has been a surge of work on modelling different stylistic aspects, such as politeness and formality, in Natural Language Generation with deep learning methods (among others, Sennrich et al., 2016; Hu et al., 2017; Ficler and Goldberg, 2017; Niu and Bansal, 2018). Regarding generation in dialogue systems, besides the two response generation models we have tested, other recent approaches to open-domain dialogue have considered stylistic aspects. For example, Yang et al. (2017) leverage metadata about speakers' personal information, such as age and gender, to condition generation using domain adaptation methods; while Luan et al. (2017) use multi-task learning to incorporate an autoencoder that learns the speaker's language style from non-conversational data such as blog posts. The output of these models could also be assessed for personality differences using our method.

More recently, Oraby et al. (2018) have used the statistical rule-based generator PERSONAGE (Mairesse and Walker, 2010) to create a synthetic corpus with personality variation within the restaurant domain. They use the data to train and evaluate neural generation models that produce linguistic output given a dialogue act and a set of semantic slots, plus different degrees of personality information, and show that the generated output correlates reasonably well with the synthetic data generated by PERSONAGE. Our work differs from Oraby et al. (2018) in several respects: We focus on open-domain chit-chat dialogue, where the input to the model is surface text (rather than semantic representations such as dialogue acts) from naturally occurring dialogue data. Rather than relying on parallel data with systematic personality variation, we exploit a personality recogniser. In this respect, our approach has some similarities to Niu and Bansal (2018), who use a politeness classifier for stylistic dialogue generation. Here we have used the personality recogniser by Mairesse et al. (2007), which may not be ideal as it was originally trained on snippets of conversations combined with stream of consciousness essays. Our method, however, is not tied to this particular recogniser—any other personality recogniser that produces numerical scores may be used instead.

We think that the automatic evaluation method we have proposed can be a useful complement to qualitative human evaluation of chatbot models.

Our study shows that the models under investigation produce output that retains some stylistic features related to personality, and can learn surface patterns that generalise beyond the training data.

Acknowledgements

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104. Association for Computational Linguistics.

Jonathan Herzig, Michal Shmueli-Scheuer, Tommy Sandbank, and David Konopnicki. 2017. Neural response generation for customer service based on personality traits. In *Proceedings of The 10th International Natural Language Generation conference*, pages 252–256.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Controllabale text generation. In *Proceedings of ICML*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. 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*, pages 110–119, San Diego, California. Association for Computational Linguistics.

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

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Data distillation for controlling specificity in dialogue generation. *arXiv preprint arXiv:1702.06703*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 2017. Multi-task learning for speaker-role adaptation in neural conversation models. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 605–614.

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

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:273–389.

Warren T Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, 66(6):574.

Shereen Oraby, Lena Reed, Shubhangi Tandon, TS Sharath, Stephanie Lukin, and Marilyn Walker. 2018. Controlling personality-based stylistic variation with neural natural language generators. In *Proceedings of SIGdial*.

James W Pennebaker and Laura A King. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40.

Iulian V. Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016a. Generative deep neural networks for dialogue: A short review. In *30th Conference on Neural Information Processing Systems (NIPS 2016), Workshop on Learning Methods for Dialogue*.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016b. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3776–3783. AAAI Press.

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. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Jörg Tiedemann. 2009. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248.

Min Yang, Zhou Zhao, Wei Zhao, Xiaojun Chen, Jia Zhu, Lianqiang Zhou, and Zigang Cao. 2017. Personalized response generation via domain adaptation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1021–1024, New York, NY, USA. ACM.

Kaisheng Yao, Geoffrey Zweig, and Baolin Peng. 2015. Attention with intention for a neural network conversation model. *arXiv preprint arXiv:1510.08565*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? *ArXiv e-prints*.