INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Personalized Dialogue Response Generation Learned from Monologues

Feng-Guang Su[1*], *Aliyah R. Hsu*[2*], *Yi-Lin Tuan*[3], *Hung-Yi Lee*[4]

Department of Electrical Engineering, National Taiwan University

{[1]b04901070,[2]b04705036}@ntu.edu.tw, {[3]pascaltuan,[4]tlkagkb93901106)@gmail.com

Abstract

Personalized responses are essential for having an informative and human-like conversation. Because it is difficult to collect a large amount of dialogues involved with specific speakers, it is desirable that chatbot can learn to generate personalized responses simply from monologues of individuals. In this paper, we propose a novel personalized dialogue generation method which reduces the training data requirement to dialogues without speaker information and monologues of every target speaker. In the proposed approach, a generative adversarial network ensures the responses containing recognizable personal characteristics of the target speaker, and a backward SEQ2SEQ model reconstructs the input message for keeping the coherence of the generated responses. The proposed model demonstrates its flexibility to respond to open-domain conversations, and the experimental results show that the proposed method performs favorably against prior work in coherence, personality classification, and human evaluation.

Index Terms: personalized dialogue generation, generative adversarial network, sequence-to-sequence model, monologues

1. Introduction

A common problem in dialogue generation is that the produced responses often lack speaker information and diversity. As one way to reduce the generic responses, a persona model [1] was first proposed. It incorporates persona embeddings in a SEQ2SEQ model to capture representative terms of speakers such as their birthplaces, names, and occupations. Similar models [2, 3] also require much background knowledge of the speakers and a sufficient amount of dialogues with speaker labels (denoted as personalized dialogues here). To tackle the difficulty of collecting such data, recent work [4, 5] has attempted to use a large unlabeled open-domain dialogue corpus and speaker monologues as a substitution. The monologues are leveraged to restrict or fine-tune the general SEQ2SEQ model for imposing recognizable styles of target speakers on the generated outputs [4]. In another approach [5], an auto-encoder is used to incorporate non-conversational persona data of target speakers, and the decoder parameters are shared between the auto-encoder and a SEQ2SEQ model for multi-task learning. Although the incorporation of speaker information can disentangle parts of the diverse responses, the sole reliance on cross-entropy loss can still lead to the generation of generic responses [6, 7, 8].

In this paper, we proposed a novel personalized dialogue response generation model which can be learned without personalized dialogues. An illustration of the concept is shown in Fig. 1. The proposed model is composed of a generator, a discriminator and a reconstructor. Given an input message, the generator generates a response. The discriminator pushes the model to generate personalized responses with a persona adversarial classifier, and ensures the grammar and the logic of the responses with a real/fake score. The reconstructor reconstructs the input message from the generated response, which ensures the mutual information between the generated responses and the input messages. Through jointly trained from general open-domain dialogues and monologues of target speakers, our model is able to respond to open-domain conversations with an assigned personality rather than limiting to speaker-specific conversations. The proposed model, shown in Fig. 2, is inspired by the work of [9] and [10] on the multi-domain translation in image processing. However, in our work, the distinctive personas are considered as the different domains. Different from previous work [4, 5] which simply transforms the styles of the generated responses, the coherence between input messages and responses are further considered in our model.

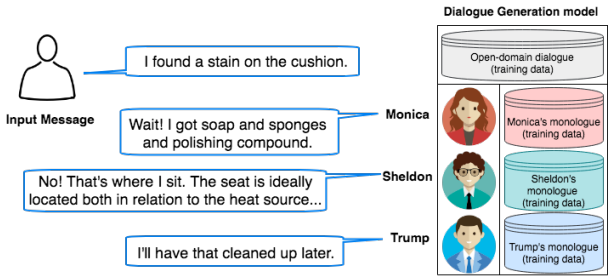

Figure 1: *The concept of personalized dialogue generation learned from Monologues. Given an input message, the network outputs personalized responses conditioned on different target speakers.*

In the experiment, we compared our model with the following baselines: standard SEQ2SEQ [11], persona [1], overtrain [4], mtask-M [5] and sentence style transfer (SST) [12, 13]. Overtrain, mtask-M and SST are all able to generate personalized response with speaker monologues and open-domain dialogues, while the other baselines require personalized dialogues. To quantify the performance, we used four automatic evaluation metrics: MaxBLEU [14] (for personality similarity), personality classification accuracy, BLEU [15] (for reasonability of responses) and a proposed evaluation measure for the coherence of the input message and response. Since automatic metrics are not purely reliable [16, 17, 18], we further conducted human evaluation on sentence appropriateness and persona resemblance. Our model significantly outperforms the state-of-the-art methods in both the automatic measures and human evaluation.

2. Model Description

In our setting, we have a general open-domain dialogue corpus containing pairs of unlabeled input message and response (x, y^*). For each target persona v, we have a set of sentences

[*] indicates the authors contribute equally to this work.

http://dx.doi.org/10.21437/Interspeech.2019-1696

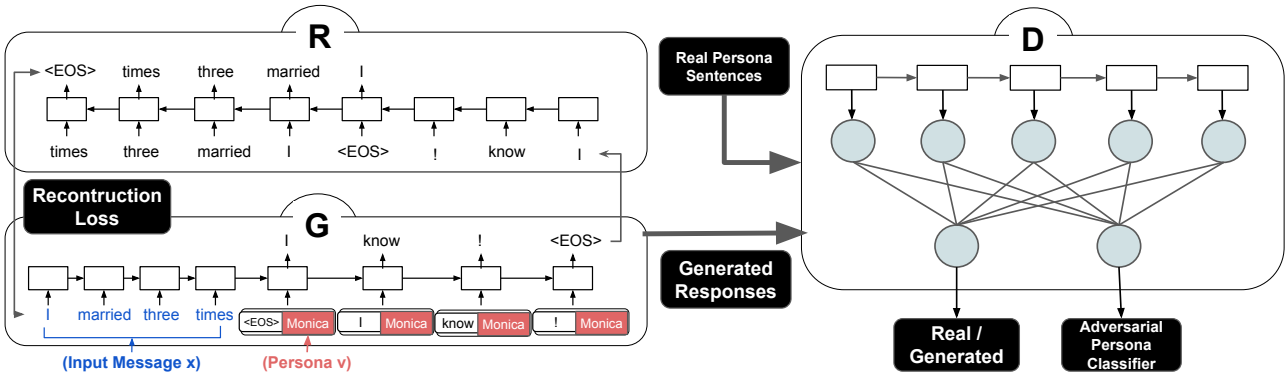

Figure 2: *The model is composed of three systems, a generator(G) a reconstructor(R), and a discriminator(D). The generator G takes in the input message x and target persona v to generate a personalized response $\hat{y}_v$. The reconstructor R then takes in $\hat{y}_v$ to reconstruct the input sequence x. D discriminates between the real and generated responses, and classifies their personas.*

y_v^* in the monologues uttered by the persona v.

2.1. Models

2.1.1. Generator and Reconstructor

The generator G is a SEQ2SEQ model for conditional text generation. The generator takes a sentence $x = \{x_1, x_2, ..., x_T\}$ and a randomly sampled target persona v as inputs, where x_t represents a word. The one-hot vector v is fed into each decoding time step of the generator to generate a personalized response $\hat{y}_v = G(x, v)$

The goal of the reconstructor R is to ensure that the generated responses are related to the input messages. Here the reconstructor R is also a SEQ2SEQ model, which reconstructs the original input sentence x from $\hat{y}_v$.

2.1.2. Discriminator

The discriminator D consists of a recurrent layer and two sets of fully connected feedforward layers. The recurrent layer takes a response y as input. The average outputs [19] from the recurrent layer throughout all time steps are averaged as the input of the feedforward layers. The two sets of feedforward layers produce two scores, $D_{src}(y)$ and $D_{cls}(y)$, respectively. $D_{src}(y)$ distinguishes between the real response y^* and the generated (fake) response $\hat{y}$; $D_{cls}(y)$ predicts the persona of the response y, where each dimension corresponds to the probability of y being classified to a specific persona. We use $D_{cls}(v|y)$ to denote the probability of y being classified to the target persona v.

2.2. Loss Function

2.2.1. Adversarial Loss

We impose the adversarial loss L_{adv} [20] as below.

$$L_{adv} = \mathbb{E}_{y^*}\left[log\left(D_{src}\left(y^*\right)\right)\right] + \mathbb{E}_{x,v}\left[log\left(1 - D_{src}\left(G\left(x, v\right)\right)\right)\right] \quad (1)$$

The discriminator learns to maximize L_{adv}. By maximizing L_{adv}, the discriminator learns to assign a higher score $D_{src}(y)$ to real sentence y^* and a lower score to the generated one $\hat{y}_v = G(x, v)$. The real response y^* is sampled from the monologues of the target persona, while x is sampled from open-domain dialogues. v is uniformly sampled from all available personas. The generator G learns to fool the discriminator by minimizing L_{adv}. The adversarial loss L_{adv} helps G to generate naturalistic responses.

2.2.2. Persona Classification Loss

The discriminator learns to minimize L_{cls}^r to correctly predict the persona of a given sentence. L_{cls}^r is defined as below.

$$L_{cls}^r = \mathbb{E}_{y_v^*}\left[-logD_{cls}\left(v|y_v^*\right)\right]. \quad (2)$$

v is the persona of the sentence y_v^*, and $D_{cls}\left(v|y_v^*\right)$ represents the predicted probability for y_v^* to be uttered by the persona v.

The generator G learns to minimize L_{cls}^f to make the imposed style on the generated response as evident as possible.

$$L_{cls}^f = \mathbb{E}_{x,v}\left[-logD_{cls}\left(v|G\left(x, v\right)\right)\right]. \quad (3)$$

$D_{cls}(v|G(x, v))$ is the predicted probability for $\hat{y}_v = G(x, v)$ to be uttered by persona v. L_{cls}^f is used to push G to generate personalized responses that most resemble the target persona v.

2.2.3. Reconstruction Loss

Contextual coherence is also a concern when evaluating the quality of a generated response. To reinforce the learning of the mutual information with the input message, the reconstructor R is used to reconstruct the input sentence x from the generated response $\hat{y}_v$. The reconstruction loss L_{rec} is defined as the following.

$$L_{rec} = \mathbb{E}_{x,v}\left[-logP_R(x|G\left(x, v\right))\right], \quad (4)$$

L_{rec} is the expected negative log likelihood of the probability (cross-entropy here) to reconstruct the context x given the generated response $\hat{y}_v = G(x, v)$. $P_R(x|y)$ is the probability that the reconstructor R generates the context x given the generated sentence y. By minimizing L_{rec}, the generator learns to generate outputs coherent with input messages.

2.2.4. Loss for General Responses

To ensure the coherence of the generated sentences and to make the model able to handle not just speaker-specific conversation but the general open-domain conversation, we adopted a joint training method. We sampled input-response pairs (x, y^*) from the open-domain dialogue corpus and trained the model on the

Table 1: *Evaluation results of The Big Bang Theory (TBBT) models and Friends models tested on the input messages from Opensubtitles (denoted as "open") and the scripts of the corresponding TV series (denoted as "in"). Persona-cls indicates personality classification accuracy, and Coh indicates the coherence score.*

train set	**Friends**						**TBBT**					
model	**MaxBLEU**		**Persona-cls**		**Coh**	**BLEU**	**MaxBLEU**		**Persona-cls**		**Coh**	**BLEU**
	open	in	open	in			open	in	open	in		
SEQ2SEQ [11]	0.141	0.203	0.172	0.195	-2.90	0.211	0.127	0.184	0.181	0.149	-2.84	0.209
persona [1]	0.271	0.302	0.198	**0.309**	-2.98	0.218	0.233	0.309	0.187	0.256	-2.99	0.209
overtrain [4]	0.312	0.347	0.294	0.305	-2.40	0.215	0.328	0.325	0.250	0.246	-2.25	0.279
mtask-M [5]	0.229	0.233	0.252	0.241	-2.77	0.297	0.213	0.225	0.199	0.201	-2.53	**0.301**
SST [12, 13]	0.301	0.381	0.227	0.257	-2.68	**0.314**	0.290	0.301	0.248	0.218	-2.63	0.288
proposed	**0.379**	**0.390**	**0.324**	0.308	**-2.26**	0.306	**0.346**	**0.335**	**0.306**	**0.272**	**-2.12**	**0.301**
-rf	0.288	0.333	0.265	0.246	-2.38	0.288	0.289	0.269	0.247	0.228	-2.25	0.260
-recon	0.310	0.373	0.264	0.213	-2.57	0.281	0.321	0.297	0.244	0.236	-2.29	0.261

paired samples with teacher forcing. The generator G learns to minimize L_{gr}:

$$L_{gr} = \mathbb{E}_{x,y^*}\left[-logP_G(y^*|x)\right]. \quad (5)$$

$P_G(y^*|x)$ is the probability that G generates the response y^* given the input message x. (x, y^*) is from the general open-domain dialogues. Since the persona of y^* is not available, v is set to be a zero vector here. With L_{gr}, our model is able to perform well without pre-training.

2.3. Training Algorithm

The discriminator D learns to minimize L_D consisting of L_{adv} and L^r_{cls} by applying gradient descent.

$$L_D = -\lambda_{adv}L_{adv} + \lambda_{cls}L^r_{cls} \quad (6)$$

The reconstructor R is trained to minimize L_{rec} in (4). The generator G learns to minimize L_{gr} in (5) and L_F:

$$L_F = \lambda_{adv}L_{adv} + \lambda_{cls}L^f_{cls} + \lambda_{rec}L_{rec}. \quad (7)$$

The hyperparameters $(\lambda_{adv}, \lambda_{cls}, \lambda_{rec})$ determine the weights we place on each loss. Since L_{adv}, L^f_{cls}, and L_{rec} are all not differentiable, we minimize L_F by policy gradient [21].

3. Data

In the experiments, we used three types of datasets: (unlabeled) open-domain dialogues, (labeled) personalized dialogues and persona monologues.

Open-domain Dialogues. OpenSubtitles dataset[1] [22] consists of movie subtitles and is used to provide a general conversational knowledge and to keep the contextual coherence in our experiment. Since OpenSubtitles is a large and noisy corpus [23] containing a good amount of generic sentences like 'I don't know' and 'I'm sorry', we filtered out those input-response pairs containing the frequent generic responses with a defined threshold. The pruned training set has in total $453,106$ lines.

Personalized Dialogues. To compare with prior persona models [1], we used dialogues in American TV series *Friends* and *The Big Bang Theory* (TBBT). Since the speaker of each line is labeled and the dialogues are properly split, these datasets are cleaner than OpenSubtitles. However, such data is more difficult to collect and has insufficient lines to train a reliable conversation model. Noted that the proposed approach does not require personalized dialogues and that the data here is not overlapped with the open-domain dialogue corpus.

Persona Monologues. We collected the monologues of main characters in the TV scripts mentioned in Personalized Dialogues. There are six main characters and a total $101,435$ lines in the *Friends* training set; seven main characters and a total $66,880$ lines in the TBBT training set. We also collected the speech of Donald Trump[2] in his 2016 campaign for the U.S. President. The dataset consists of $29,480$ lines. Trump's monologue is used in human evaluation because political speech is more distinguishable from the chit-chat in Friends or TBBT.

4. Experiments

4.1. Implementation Details

The recurrent neural networks of the discriminator, generator and reconstructor are all two-layered LSTM with 512 hidden cells for each layer. For encoder, word representations are initialized with 512-dimensional GloVe vectors [24]. For decoder, 384-dimensional GloVe vectors concatenated with 128-dimensional persona vector representation are used, where the 128-dimensional persona vector is generated by a fully connected network given the one-hot persona representation. D_{steps} and G_{steps} are set to be 5 and 1 to make sure training stability [21]. The hyperparameters $(\lambda_{adv}, \lambda_{cls}, \lambda_{rec})$ are determined after individually experimented with the integers in the range of 1 to 5, and are set to be (1,5,1) for optimal performance. We found they could significantly affect the level of personalization and coherence of the generated responses.

4.2. Evaluation Metrics

To quantify the performance, we used four automatic evaluation metrics: MaxBLEU, personality classification accuracy, coherence score from SEQ2SEQ model, and BLEU. MaxBLEU [14] is an accuracy measure quantifying the similarity between the generated response and persona monologues. We also trained a persona classifier with monologues to evaluate the generated persona accuracy. The personality classification accuracy on real monologues is around 0.731 in *Friends* and TBBT.

[1] http://www.opensubtitles.org/

[2] https://github.com/ryanmcdermott/trump-speeches

Table 2: *Human evaluation scores regarding Appropriateness and Resemblance on models trained with Trump's monologue.*

Human Preference Score	Appropriateness	Resemblance
proposed v.s. SEQ2SEQ	0.889	0.815
proposed v.s. overtrain	0.710	0.481
proposed v.s. mtask-M	0.630	0.693
proposed v.s. SST	0.592	0.714

Besides evaluating the personalization capability of our model, we have to ensure the coherence of the generated responses. We trained a SEQ2SEQ model with OpenSubtitles, and used it to compute the log likelihood of responses conditioned on input messages. The log likelihood is denoted as the coherence score in the paper. By considering the real responses in the corpora as ground truth, we computed the BLEU score [15] of the generated responses, which is our another measure for evaluating the contextual coherence.

4.3. Results

We compared our model to the five baselines on *Friends* and *TBBT*. The baselines include those requiring personalized dialogues: SEQ2SEQ [11], persona-based model [1], and those using a general corpus and monologues: overtrain [4], mtask-M [5] and sentence style transfer (SST) [12, 13].

Overtrain [4] includes a forward and a backward SEQ2SEQs. This model requires parallel dialogue training. Specifically, after the forward and backward SEQ2SEQ are trained with the open-domain dialogues, persona monologues are fed into the pre-trained backward model to generate pseudo-context. The pseudo personalized dialogue pairs are then overtrained on the forward SEQ2SEQ and thus the model is named "overtrain". **Mtask-M [5]** consists of an auto-encoder to encode persona monologues and a persona-based model to take in the encoded vector for personalized response generation.

As shown in Table 1, for MaxBLEU and personality classification accuracy (Persona-cls), we tested input messages from the validation set of OpenSubtitles (open-domain) and *Friends* and TBBT (in-domain) to evaluate if the model can generalize to usual open-domain dialogues. We observe that our model outperforms the baselines in both metrics no matter the testing data is in-domain or open-domain. Although persona-based model [1] performs almost comparable to our model when tested with in-domain data, it obtains lower accuracy in the two metrics when tested with open-domain dialogues. Since persona-based model is trained purely with personalized dialogues, it easily overfits and fails to generalize.

We also tested the two additional ablation models, *proposed -rf* and *proposed -recon*, which have the same structure as the proposed model but the real/fake loss and the reconstruction loss are respectively removed. According to the results, the two ablation models have similar performance decays in MaxBLEU and Persona-cls, which demonstrates the validity and equal importance of the discriminator D and the reconstructor R.

The coherence score (Coh) and BLEU in Table 1 also show that our model outperforms previous work. This means that the proposed model does not sacrifice coherence to generate personalized responses. We suspect the reason for the baselines to perform poorer should be the lack of a mechanism to maintain the mutual information of the generated responses and the input messages. The ablation tests further support the proposed reason in that *proposed -recon* did perform worse than *proposed -rf* in coherence score due to the lack of reconstruction loss.

Table 3: *Different personalized responses generated by our model for: Friends and the Big Bang Theory (TBBT). Generated responses from the standard SEQ2SEQ model is also provided for reference.*

Friends- message	I saw him talking to her!
(SEQ2SEQ)	Have you even seen her?
Monica	I told her he was killed ...
Chandler	I know what the truth is.
TBBT- message	Do you like him?
(SEQ2SEQ)	No, just stop.
Sheldon	He's a really good friend.
Penny	He's a cool kid.

4.4. Human Evaluation

We conducted a human evaluation of the generated responses of our model and the four baselines: SEQ2SEQ, overtrain, mtask-M and SST on Trump's monologue. We chose to evaluate on Trump's monologue because its political characteristic is more apparent than the daily conversation in Friends and TBBT. The SEQ2SEQ model here is trained with OpenSubtitles. The persona-based model is not evaluated here since the model requires personalized dialogues for training while the that for Trump is not available. The evaluation form consists of 81 input messages. For each input message, we asked the 43 judges to choose blindly between the outputs generated by either a baseline model or our proposed model based on which performed better in terms of appropriateness[3] and resemblance [4]. The chosen model would then be credited with one score, and we averaged the final scores obtained by each model based on the total amount of input messages. The statistics are reported in Table 2. The result is aligned with what we obtained from automatic metrics in that our model outperforms the baselines.

4.5. Qualitative Analysis

Here we analyze the sentence examples generated by our model[5]. In Table 3, we report four sets of generated samples imitating some main characters from *Friends* and TBBT. The results of standard SEQ2SEQ are provided as the neutral responses without any persona style for comparison. It is evident that different conditional characters lead to diverse responses in *Friends* and TBBT datasets.

5. Conclusion

We have introduced a dialogue response generation model to produce stylistic responses for multiple speakers. The proposed model can be utilized to build a personal avatar-like agent to imitate a target speaker with simply a collection of his/her speech. In this paper, we show that our model outperforms the state-of-the-art methods in MaxBLEU, personality classification accuracy, coherence score, and BLEU. The human evaluation results are also aligned with that obtained from the automatic metrics. Our model is subjectively considered better in capturing personal characteristics and keeping coherence.

[3]Appropriateness: Which output answers the input message best?

[4]Resemblance: Which output resembles the persona best?.

[5]https://adelaidehsu.github.io/Personalized-Dialogue-Response-Generation-learned-from-Monologues-demo/

6. References

[1] 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: Association for Computational Linguistics, Aug. 2016, pp. 994–1003. [Online]. Available: https://www.aclweb.org/anthology/P16-1094

[2] 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: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2775–2779. [Online]. Available: https://www.aclweb.org/anthology/D18-1298

[3] 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: Association for Computational Linguistics, Jul. 2018, pp. 2204–2213. [Online]. Available: https://www.aclweb.org/anthology/P18-1205

[4] D. Wang, N. Jojic, C. Brockett, and E. Nyberg, "Steering output style and topic in neural response generation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2140–2150. [Online]. Available: https://www.aclweb.org/anthology/D17-1228

[5] Y. Luan, C. Brockett, B. Dolan, J. Gao, and M. Galley, "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)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 605–614. [Online]. Available: https://www.aclweb.org/anthology/I17-1061

[6] S. Jiang, P. Ren, C. Monz, and M. de Rijke, "Improving neural response diversity with frequency-aware cross-entropy loss," in *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA: ACM, 2019, pp. 2879–2885. [Online]. Available: http://doi.acm.org/10.1145/3308558.3313415

[7] R. Nakamura, K. Sudoh, K. Yoshino, and S. Nakamura, "Another diversity-promoting objective function for neural dialogue generation," *CoRR*, vol. abs/1811.08100, 2018. [Online]. Available: http://arxiv.org/abs/1811.08100

[8] J. Xu, X. Ren, J. Lin, and X. Sun, "Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3940–3949.

[9] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," *arXiv preprint*, vol. 1711, 2017.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

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

[12] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, "Style transfer from non-parallel text by cross-alignment," in *Advances in neural information processing systems*, 2017, pp. 6830–6841.

[13] C.-W. Lee, Y.-S. Wang, T.-Y. Hsu, K.-Y. Chen, H.-Y. Lee, and L.-s. Lee, "Scalable sentiment for sequence-to-sequence chatbot response with performance analysis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6164–6168.

[14] Z. Xu, N. Jiang, B. Liu, W. Rong, B. Wu, B. Wang, Z. Wang, and X. Wang, "Lsdscc: A large scale domain-specific conversational corpus for response generation with diversity oriented evaluation metrics," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, vol. 1, 2018, pp. 2070–2080.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[16] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing statistical machine translation for text simplification," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 401–415, 2016.

[17] E. Sulem, O. Abend, and A. Rappoport, "BLEU is not suitable for the evaluation of text simplification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 738–744. [Online]. Available: https://www.aclweb.org/anthology/D18-1081

[18] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "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*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2122–2132. [Online]. Available: https://www.aclweb.org/anthology/D16-1230

[19] Y.-L. Tuan and H.-Y. Lee, "Improving conditional sequence generative adversarial networks by stepwise evaluation," *arXiv preprint arXiv:1808.05599*, 2018.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[21] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient." in *AAAI*, 2017, pp. 2852–2858.

[22] J. Tiedemann, "News from opus-a collection of multilingual parallel corpora with tools and interfaces," in *Recent advances in natural language processing*, vol. 5, 2009, pp. 237–248.

[23] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[24] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.