Polite Chatbot: A Text Style Transfer Application

Sourabrata Mukherjee and **Vojtěch Hudeček** and **Ondřej Dušek**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
{mukherjee,hudecek,odusek}@ufal.mff.cuni.cz

Abstract

Generating polite responses is essential to build intelligent and engaging dialogue systems. However, this task is far from well-explored due to the difficulties of rendering a particular style in coherent responses, especially when parallel datasets for regular-to-polite pairs are usually unavailable. This paper proposes a polite chatbot that can produce responses that are polite and coherent to the given context. In this study, a politeness transfer model is first used to generate polite synthetic dialogue pairs of contexts and polite utterances. Then, these synthetic pairs are employed to train a dialogue model. Automatic and human evaluations demonstrate that our method outperforms baselines in producing polite dialogue responses while staying competitive in terms of coherent to the given context.[1]

1 Introduction

Building a chatbot agent that produces stylized and coherent responses can yield more engaging conversations (Niederhoffer and Pennebaker, 2002). Generating stylized dialogue responses has been investigated in various studies, with a broad understanding of style covering emotion (Zhou et al., 2018), personality (Li et al., 2016) or politeness (Niu and Bansal, 2018). In most cases, the stylistic features we want to capture are embedded in unpaired texts that cannot be directly utilized by supervised models (Gao et al., 2019). This typically leads stylized chatbot models to employ complex, multi-step setups, potentially involving reinforcement learning (Niu and Bansal, 2018; Sun et al., 2022; Firdaus et al., 2022).

In this paper, we propose a straightforward polite chatbot training procedure that uses a politeness transfer model to create synthetic training instances and results in an end-to-end model. We build upon the work of Madaan et al. (2020), who use a *tagger* and *generator* pipeline to generate polite sentences. However, we make their process more straightforward by merging these two sub-modules into a single step: We finetune the BART model (Lewis et al., 2020) to transfer neutral sentences into polite ones. Using this model, we then prepare synthetic pairs of contexts and polite responses and train a dialogue model on this synthetic data. We evaluate our approach on The DailyDialog dataset (Li et al., 2017). Automatic and human evaluations show that our method outperforms competitive baselines in response politeness while staying competitive in terms of coherence to the given context.

2 Related Work

Politeness Transfer in Text Politeness Transfer is a sub-task of Text Style Transfer (TST) (Madaan et al., 2020). McDonald and Pustejovsky (1985) have defined *style* as a notion that refers to the manner in which semantics is expressed. The aim of TST is to change the style of the text while preserving style-independent content. Politeness is a text style attribute that is closely related to social interactions, which enables smooth communication in conversations (Coppock, 2005), such as emails or memos, and it can be decoupled from content (Kang and Hovy, 2019). The task of politeness transfer (Madaan et al., 2020) aims to control the politeness of a text while preserving the original content. Madaan et al. (2020) use a two-step architecture here: (1) a tagger tags appropriate insertion points, and (2) a generator generates polite phrases to insert instead of the tags.

Polite Chatbot Response Generation Stylized dialogue generation attracted a lot of attention in recent years (Gao et al., 2019; Zheng et al., 2021; Zeng and Nie, 2021). Previous works focus on personalized (Li et al., 2016; Luan et al., 2017; Su et al., 2019), polite Niu and Bansal (2018) or

[1] Our code and related details are available at https://github.com/souro/polite_chatbot.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 87–93

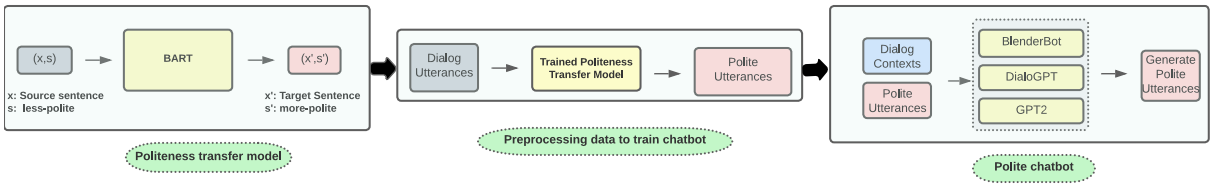

Figure 1: Our method: We (1) train the politeness transfer model; (2) generate synthetic training data by applying the transfer model to neutral utterances; (3) train the dialogue models using the synthetic data.

emotional (Zhou et al., 2018) dialogues.

For politeness, traditionally, polite chatbot responses are accomplished by manual dialogue design, where predefined rules or templates are used to generate responses based on certain keywords or scenarios (André et al., 2004; Gupta et al., 2007; de Jong et al., 2008). This approach has some limitations such as requiring a lot of human effort, being domain-specific, and lacking flexibility or diversity (Firdaus et al., 2022). Alternatively, recent works have used neural language models to generate polite chatbot responses automatically. Niu and Bansal (2018) used a politeness classifier and a language model trained on polite utterances to generate polite dialog responses. Sun et al. (2022) post-process a baseline system response using a two-step phrase replacement trained by reinforcement learning. Firdaus et al. (2022) proposed a two-step decoding approach that first generates a rough response based on the input text and then infuses human-written polite phrases into the response using a separate politeness model.

Perhaps the closest to ours is the work of Silva et al. (2022), who also adapts Niu and Bansal's and Madaan et al.'s models, but their focus is domain transfer, not simplifying the overall architecture.

3 Method

Our method consists of three steps (Figure 1). First, we train a politeness transfer model. Our goal here is to train a model that takes as input a neutral sentence x and outputs a sentence $\hat{x}$ that retains the content while increasing politeness. Second, we apply this politeness transfer model to generate synthetic polite chat data. Finally, we use the corpus $\hat{\mathcal{D}}$ to train a dialogue model.

Politeness Transfer Model Although we do not have parallel corpora available for politeness transfer, our transfer model is trained in a supervised fashion on synthetic input-output pairs. These are obtained following Madaan et al. (2020): polite phrases (politeness markers) are identified using TF-IDF over polite and non-polite texts.[2] The markers are removed from polite texts on the input, and a sequence-to-sequence model is trained to increase sentence politeness by reconstructing the politeness markers on the output. Unlike Madaan et al. (2020), we do not use separate tagging and generation steps here and join the task into a single step. Specifically, we finetune a pre-trained language model for this task using standard cross-entropy loss (see Section 4.2).

Models	PS	BLEU	CS
Madaan et al. (2020)	7.01	60.16	87.86
Ours	**8.68**	**71.65**	**93.25**

Table 1: Evaluation results of politeness transfer on the test set of Madaan et al. (2020)'s data. We measure the Polite Score (PS), BLEU Score, and Content Similarity (CS). Model outputs are predicted based on synthetic sentences where politeness markers have been removed. BLEU and CS compare against original human-written polite sentences.

Creating Synthetic Polite Data We apply our politeness transfer model to a dataset consisting of N dialogues $\mathcal{D} = \{C_1^{k_1}, ..., C_N^{k_N}\}$, where dialogue $C_i^{k_i}$ consists of k_i utterances $\{u_i^1, ..., u_i^{k_i}\}$. We create a corpus of context-utterance pairs $\hat{\mathcal{D}} = \{\langle C_1^1, \hat{u}_1^2\rangle, \langle C_1^2, \hat{u}_1^3\rangle, ..., \langle C_N^{K_N-1}, \hat{u}_N^{K_N}\rangle\}$ where C_1^1 is the first utterance of the first dialogue, C_1^2 are the first two utterances of the first dialogue, etc. In other words, for every partial context, we add a polite version of the next utterance.

Dialogue Model We use a standard dialogue response generation model that produces a dialogue utterance u_i based on context $\mathbf{C} = \{u_1, ..., u_{i-1}\}$, trained using cross-entropy loss. We experiment with multiple pre-trained language models here

[2]In principle, a much higher mean TF-IDF value over polite than non-polite texts means that a phrase is likely to be a politeness marker.

	BlenderBot			DialoGPT			GPT-2			
Finetuned on	PS	BLEU-1,2	CS	PS	BLEU-1,2	CS	PS	BLEU-1,2	CS	
Vanilla (no FT)	7.06	9.80	2.58	20.31	6.31	9.38	1.98	19.33	4.91	0.15	0.09	8.31
DailyDialog (DD)	7.11	17.21	7.25	45.80	6.14	11.72	2.60	38.44	5.08	7.82	2.13	29.72
DD + Madaan et al. (2020)	6.75	17.16	6.73	45.17	6.17	11.47	2.19	35.08	5.99	7.32	1.49	27.42
DD + Ours	**7.65**	17.03	6.85	41.80	**7.75**	11.44	2.57	35.03	**7.20**	5.65	1.03	26.80

Table 2: Evaluation results of polite dialog models. We indicate what version of the DailyDialog dataset (DD) was used for Finetuning (FT) if any. We measure the Polite Score (PS), BLEU score, and Content Similarity (CS). BLEU Score (of n-gram = 1,2) and CS are computed between predicted polite utterances and the original utterances.

Models	PS	BLEU	CS
DailyDialog (DD)	5.41	–	–
DD + Madaan et al. (2020)	6.37	73.34	90.29
DD + Ours	**7.95**	70.21	89.07

Table 3: Evaluation of synthetic data generated using DailyDialogue (DD) to train polite dialog models. We measure the Polite Score (PS), BLEU Score, and Content Similarity (CS) The BLEU and CS are measured between original utterances and polite-transferred utterances.

BlenderBot finetuned on	Pol	CC	Flu
Vanilla (no FT)	3.46	1.16	4.64
DailyDialog (DD)	3.90	3.74	4.54
DD + Madaan et al. (2020)	3.50	3.06	3.98
DD + Ours	**4.26**	2.94	4.30

Table 4: Human Evaluation on BlenderBot outputs. We measured politeness (Pol), coherent to context (CC), and fluency (Flu).

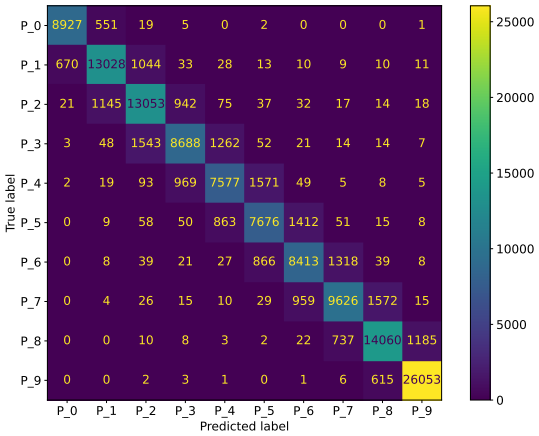

Figure 2: Confusion Matrix of multi-class politeness classification. We can see that the model confuses mainly neighboring buckets and the vast majority is classified correctly.

(see Section 4.2). To achieve politeness in responses, we use the synthetic polite dialogue corpus $\hat{\mathcal{D}}$ obtained using our politeness transfer model.

4 Experiment

4.1 Datasets

Politeness Transfer We use the dataset of Madaan et al. (2020), i.e. preprocessed and filtered sentences from the Enron e-mail dataset (Shetty and Adibi, 2004) into ten buckets (P_0-P_9) based on the score of a politeness classifier by Niu and Bansal (2018). We use Madaan et al. (2020)'s TF-IDF-based approach to remove politeness markers (see Section 3) from the sentences in the most polite P_9 bucket to prepare synthetic parallel data for training our politeness transfer models.

Dialogue To train our response generation models, we use DailyDialog (Li et al., 2017), an open-domain dataset of 13,118 human-human dialogues. DailyDialogue was collected to represent natural day-by-day conversations between human participants. It is constructed mainly from English learner websites and represents rather formal written conversations. Each conversation is focused on certain topics, but the domain is not restricted in general. The average length is 7.9 turns per dialogue. The dataset is split into a training set with 11,118 dialogues and validation and test sets with 1,000 dialogues each.

4.2 Settings

We use BART (Lewis et al., 2020) for politeness transfer. For dialogue modeling, we use multiple pre-trained models: (1) GPT-2 (Radford et al., 2019), which is a Transformer decoder trained for general language modeling (including dialogues), (2) DialoGPT (Zhang et al., 2020), which shares GPT-2's architecture but was pre-trained specifically on dialogue data, (3) BlenderBot (Roller et al., 2021), which is an encoder-decoder Transformer

Context : Nani Book Store. How can I help you? Do you have the 'The man and the Sea' by Ernest Hemingway?	
Ground Truth	**Yes, we have one copy left.**
Vanilla (no FT)	I would love to try it, but I don't think I'll be able to afford it.
DailyDialog (DD)	Yes, it ' s on that shelf. I ' ll get it for you right away.
DD + Madaan et al. (2020)	We do. Would you like a look at it?
DD + Ours	Yes, we do. Would you like to have a look at it? please let me know.

Table 5: A sample output using BlenderBot. Vanilla BlenderBot produces polite but irrelevant responses, and models finetuned on all DailyDialog data versions produce relevant responses, but ours is arguably the most polite.

specifically trained to learn dialogue skills such as empathy or engagement.[3]

4.3 Baselines

Politeness Transfer We compare our system against Madaan et al. (2020). They used 4-layered transformers (Vaswani et al., 2017) to train both tagger and generator modules. Each transformer has 4 attention heads with a 512 dimensional embedding layer and hidden state size.

Dialogue Model We evaluate all dialogue models against three baselines: (1) vanilla version of the model, (2) model fine-tuned on unchanged DailyDialog data, (3) model finetuned on synthetic polite DailyDialog data generated in the same fashion as in our full model, but using Madaan et al. (2020)'s politeness transfer instead of ours.

5 Evaluation

5.1 Metrics

Following prior work (Madaan et al., 2020; Niu and Bansal, 2018), we use automatic metrics for the evaluation of the models along two major dimensions: (1) style transfer and (2) content preservation and relevance. To measure politeness transfer quality, we compute *Polite Score*, which is defined as the average score given to the generated sequences by our politeness classifier, which we created by finetuning BERT (Devlin et al., 2019) on Madaan et al. (2020)'s Enron data (see Section 4.1).[4] Following prior work (Jin et al., 2022; Hu et al., 2022), we evaluate the relevance and content preservation using embedding similarity (Rahutomo et al., 2012) and BLEU score (Papineni et al., 2002). For embedding similarity, we use a pre-trained Sentence-BERT model (Reimers and Gurevych, 2019) and cosine similarity. We use BLEU-1 and BLEU-2 to account for the expected different phrasing in polite outputs and the high output variance common to open-domain dialogue response generation. As automated metrics for language generation do not correlate well with human judgments (Novikova et al., 2017), we conduct a small-scale in-house human evaluation with expert annotators (computational linguistics graduate students). We randomly select 50 context-utterance pairs from the DailyDialog test set for all models based on the strongest BlenderBot language model. The annotators rate model outputs using a 5-point Likert scale for politeness, coherence to context, and fluency.

5.2 Results

Politeness classification The accuracy of our BERT-based politeness classification model is 83.27% on the politeness transfer data. More importantly, the confusion matrix in Figure 2 shows that the model confuses mostly adjacent classes; the average error is only 0.98.

Politeness Transfer We compare the politeness transfer models on content preservation and politeness improvement using a test portion of Madaan et al. (2020)'s data used for training, which consists of synthetic non-polite sentences and the corresponding original polite sentences. Models are tasked with producing polite sentences from synthetic non-polite ones; the result is then compared to the original human-written polite sentences. Table 1 shows the results. Our model achieves a higher politeness score than Madaan et al. (2020) while producing sentences more similar to the original human-written ones based on BLEU and sentence similarity scores.

We also evaluate the performance of the politeness transfer models with respect to content preservation and politeness improvement on the synthetic

[3]We use AdamW optimizer with a learning rate of 5e-4 in all cases. The politeness transfer model is trained for 5 epochs using batch size 8. All dialogue models are finetuned for 4 epochs using batch size 3.

[4]Although the scale of politeness classes is not necessarily linear, we believe that this is still a good indicator of the overall politeness of the data.

pairs of contexts and polite utterances from the DailyDialog dataset we prepared. The results are shown in Table 3. Note that unlike in the previous experiment, we measure content preservation against the original (source) utterances. We observe that our model increases politeness over the source data and outperforms Madaan et al. (2020). We can see a slight drop in content preservation metrics against the original utterances, but this is expected as these metrics also reflect changes in phrasing.

Dialogue modeling Results of automatic metrics for dialogue modeling are shown in Table 2. The performance differences between the pre-trained models used are expected given the models' properties and intended use cases. While GPT-2 scores low on politeness, the dialogue-specific models obtain better results. As expected, all models perform much better in terms of content preservation after finetuning. Both ours and Madaan et al.'s politeness transfer result in an increase in politeness, and we can observe that our method consistently outperforms Madaan et al.'s. Moreover, our method is the only one that improves the Polite Score over the vanilla BlenderBot model. Finally, although the application of politeness transfer causes a decrease in content similarity with reference responses from DailyDialog, the drop is marginal, not consistent with all metrics, and could be caused by different phrasing, same as in the case of politeness transfer (cf. Table 3).

Human Evaluation We have evaluated 50 model outputs for each variant of the BlenderBot model (see Table 5 for a sample). The results are presented in Table 4. The human evaluation results mostly agree with our automatic evaluation results: our data preparation method performs better than Madaan et al. (2020)'s transfer in terms of politeness and is able to improve the base BlenderBot model. Both politeness-increasing methods cause a slight degradation in context coherency of the generated utterances; ours performs slightly worse in this aspect. However, our full approach yields more fluent outputs than the model trained on Madaan et al. (2020)'s politeness transfer.

6 Conclusion

We propose an innovative way of increasing dialogue models' politeness. Our method is trained in two steps: the creation of synthetic training corpora with increased politeness and dialogue model training. The resulting dialogue response generation model is end-to-end and does not require postprocessing. Compared against multiple baselines for both politeness transfer and dialogue modeling, our politeness transfer model and dialogue response generation achieve increased politeness while still preserving important content. In future work, we aim to extend our method to other stylized response generation tasks.

Acknowledgments

This research was supported by Charles University projects GAUK 392221, GAUK 302120, and SVV 260575, and by the European Research Council (Grant agreement No. 101039303 NG-NLG). We would like to express our gratitude to our colleague Zdeněk Kasner for his insightful discussions and helpful feedback on this project.

References

Elisabeth André, Matthias Rehm, Wolfgang Minker, and Dirk Bühler. 2004. Endowing spoken language dialogue systems with emotional intelligence. In *Affective Dialogue Systems, Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004, Proceedings*, volume 3068 of *Lecture Notes in Computer Science*, pages 178–187. Springer.

Liz Coppock. 2005. Politeness strategies in conversation closings. *Unpublished manuscript: Stanford University*.

Markus de Jong, Mariët Theune, and Dennis Hofs. 2008. Politeness and alignment in dialogues with a virtual guide. In *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal, May 12-16, 2008, Volume 1*, pages 207–214. IFAAMAS.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Mauajama Firdaus, Arunav Shandilya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Being polite: Modeling politeness variation in a personalized dialog agent. *IEEE Transactions on Computational Social Systems*.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019. Structuring latent spaces for stylized response generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1814–1823. Association for Computational Linguistics.

Swati Gupta, Marilyn A. Walker, and Daniela M. Romano. 2007. How rude are you?: Evaluating politeness and affect in interaction. In *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*, volume 4738 of *Lecture Notes in Computer Science*, pages 203–217. Springer.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *SIGKDD Explor.*, 24(1):14–45.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Comput. Linguistics*, 48(1):155–205.

Dongyeop Kang and Eduard Hovy. 2019. xSLUE: A benchmark and analysis platform for cross-style language understanding and evaluation.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. 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, Berlin, Germany. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabás Póczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1869–1881. Association for Computational Linguistics.

David D. McDonald and James Pustejovsky. 1985. A computational theory of prose style for natural language generation. In *EACL 1985, 2nd Conference of the European Chapter of the Association for Computational Linguistics, March 27-29, 1985, University of Geneva, Geneva, Switzerland*, pages 187–193. The Association for Computer Linguistics.

Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2241–2252. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain

chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Jitesh Shetty and Jafar Adibi. 2004. The enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*, 4(1):120–128.

Diogo Silva, David Semedo, and João Magalhães. 2022. Polite task-oriented dialog agents: To generate or to rewrite? In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 304–314, Dublin, Ireland. Association for Computational Linguistics.

Feng-Guang Su, Aliyah R. Hsu, Yi-Lin Tuan, and Hung-yi Lee. 2019. Personalized dialogue response generation learned from monologues. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 4160–4164. ISCA.

Qingfeng Sun, Can Xu, Huang Hu, Yujing Wang, Jian Miao, Xiubo Geng, Yining Chen, Fei Xu, and Daxin Jiang. 2022. Stylized knowledge-grounded dialogue generation via disentangled template rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3304–3318, Seattle, United States. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yan Zeng and Jian-Yun Nie. 2021. A simple and efficient multi-task learning approach for conditioned dialogue generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4927–4939, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

Yinhe Zheng, Zikai Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, and Minlie Huang. 2021. Stylized dialogue response generation using stylized unpaired texts. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14558–14567. AAAI Press.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press.