Leveraging Low-resource Parallel Data for Text Style Transfer

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

Abstract

Text style transfer (TST) involves transforming a text into a desired style while approximately preserving its content. The biggest challenge in TST in the general lack of parallel data. Many existing approaches rely on complex models using substantial non-parallel data, with mixed results. In this paper, we leverage a pretrained BART language model with minimal parallel data and incorporate low-resource methods such as hyperparameter tuning, data augmentation, and self-training, which have not been explored in TST. We further include novel style-based rewards in the training loss. Through extensive experiments in sentiment transfer, a sub-task of TST, we demonstrate that our simple yet effective approaches achieve well-balanced results, surpassing non-parallel approaches and highlighting the usefulness of parallel data even in small amounts.[1]

1 Introduction

Text style transfer (TST) aims to modify the style of a given text while preserving its underlying content (Shen et al., 2017; Prabhumoye et al., 2018; Li et al., 2018) (see Figure 1). The limited availability of parallel training data is a major obstacle in TST, as acquiring large-scale aligned datasets for specific style pairs is often impractical or unfeasible (Jin et al., 2022; Hu et al., 2022). The only TST study using parallel data and sequence-to-sequence learning known to us by Jhamtani et al. (2017) is a very specific application: converting modern English to Shakespeare's style, where extensive aligned paraphrases happen to exist for the purposes of literature research. Most recent TST research shifted to using non-parallel datasets and unsupervised learning (Hu et al., 2017; Zhao et al., 2018; Li et al., 2018). While it shows promising results, it does suffer a performance penalty and cannot avoid the data problem completely, as large quantities of non-parallel style-specific data are still hard to come by (Li et al., 2022b).

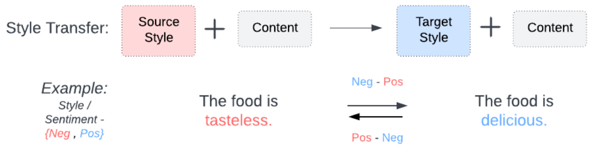

Figure 1: An example of sentiment transfer as a TST task.

In this paper, we address the challenges of TST in low-resource scenarios by proposing methodologies that capitalize on minimal parallel data. Due to parallel data availability, we focus on sentiment transfer, a prominent sub-task within the realm of TST (Jin et al., 2022; Mukherjee et al., 2022; Luo et al., 2019a), in our experiments.[2] However, our model does not rely on a specific kind of textual styles and can be applied to TST in general.

In summary, our contributions are (1) building a TST system with low-resource parallel data, (2) applying multiple low-resource adaptation techniques, (3) and a novel style reward approach. This helps us achieve well-balanced results, surpassing previous non-parallel approaches on both automatic and human evaluation. Our experimental code is available on GitHub.[1]

2 Related Work

TST with Parallel Data TST can be modeled as a sequence-to-sequence task and trained on pairs of texts with similar content but different styles. Here, Jhamtani et al. (2017) used a sequence-to-sequence model with a pointer network to translate modern English into Shakespearean English. However, this

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

[2]The task of sentiment transfer is related to sentence negation (Sarabi et al., 2019; Hosseini et al., 2021; Hossain and Blanco, 2022), but distinct from it, specifically aiming the scope of meaning change to sentiment only and going beyond using simple negation particles (cf. Table 3 in the Appendix).

Proceedings of the 16th International Natural Language Generation Conference, pages 388–395
September 11–15, 2023.

approach to TST is inherently challenging due to the scarcity of parallel data (Hu et al., 2022).

Non-Parallel Approaches to TST Two main strategies were employed to avoid reliance on parallel data: (1) Straightforward text replacement, where style-specific phrases are explicitly identified and replaced (Li et al., 2018), (2) Implicit style-content disentanglement via latent representations through techniques such as backtranslation and autoencoding (Shen et al., 2017; Zhao et al., 2018; Fu et al., 2018; Prabhumoye et al., 2018; Hu et al., 2017), adversarial learning was shown to improve the results of both approaches (Lample et al., 2019; Dai et al., 2019; Li et al., 2019; Luo et al., 2019b). Despite a lot of progress, non-parallel approaches tend to produce mixed results and often require large amounts of non-parallel data, limiting their practical applicability (Li et al., 2022b).

3 Method

Our work sits between the parallel and non-parallel approaches, using parallel data but in very small amounts, in order to maximize performance while minimizing annotation costs. We build on transfer learning by finetuning a pretrained BART model on our task (Lewis et al., 2020). We further explore five techniques aimed at this low-resource scenario:

Hyperparameter tuning: As the effectiveness of Transformer models on low-resource data highly depends on hyperparameters (Araabi and Monz, 2020), we adapt our model, focusing on dropout regularization (Sennrich and Zhang, 2019) and label smoothing (Müller et al., 2019).

Prompt-guided generation: To align the style transfer finetuning with pre-training, we adopt using textual prompts, following Li and Liang (2021) and Li et al. (2022a). By adding prompts like *"POS:"* for positive sentences and *"NEG:"* for negative sentences, we provide explicit guidance to the decoder during fine-tuning.

Data augmentation: We use data augmentation by paraphrasing (see Section 4.2) to generate more training examples and improve data diversity (Shen et al., 2020; Qiu et al., 2020).

Self-training: To further expand our data, we use self-training, i.e., training on synthetic data generated by the model itself (He et al., 2020; Chai et al., 2022). To improve the quality of the synthetic data, we filter them using style classifier accuracy, BLEU, and embedding similarity (cf. Section 5). We use a geometric mean of all three metrics as a sentence score, then choose a portion of the generated data with the top k highest scores.

Style reward: To make our generator better focus on the target style accuracy, we incorporate rewards from a style classifier into the training loss. We use a simple reward R, which is $+1$ for instances where the generated output matches the target style, and -1 where it does not. We then modify the basic cross-entropy generation loss $\mathcal{L}_{\text{CE}}$ in the following way to get the overall loss $\mathcal{L}$:

$$\mathcal{L} = \alpha \cdot \text{norm}(R) + (1 - \alpha) \cdot \mathcal{L}_{\text{CE}} \quad (1)$$

norm denotes normalization (zero mean, unit standard deviation), and α is a weight parameter.

4 Experiments

4.1 Dataset

We experiment on a small parallel sentiment transfer dataset of Yelp reviews by Li et al. (2018), comprising 500 positive-to-negative and 500 negative-to-positive sentences. The data was intended as an evaluation set only, but we repurpose it as a full low-resource set and split it into 400 examples for training, 100 for development, and 500 for testing. For self-training, we additionally use non-parallel sets of 2000+2000 positive and negative sentences from Li et al. (2018)'s development set.

4.2 Settings

We use BART-base (Lewis et al., 2020) from the HuggingFace library (Wolf et al., 2020).

Hyperparameter tuning: We ran three small-scale random searches for optimal values of individual parameters, resulting in the following changes from the defaults based on development set results: (1) We adjusted the learning rate (*LR*) ($5e-5 \rightarrow 1e-5$) and *batch size* ($8 \rightarrow 3$). (2) We increased the *Dropout* rate ($0.1 \rightarrow 0.15$) and introduced additional attention and activation dropout (both 0.1). (3) We introduced *L2* regularization with a value of 0.01 and *label smoothing* with a value of 0.05.

Prompt-guided generation does not have any specific settings; we only add the prompts on the input as described in Section 3.

Data augmentation: We used the following operations from the NLPAug library (Ma, 2019): substitute words with a *Spelling* mistake from a dictionary, *Insert* or *Substitute* words based on BERT embedding similarity, substitute words with a *Synonym* from WordNet, *Swap* or *Delete* words randomly, *Split* words into two tokens randomly. Additionally, we used *Back-translation* (Sennrich et al., 2016; Prabhumoye et al., 2018) via German using the online translation tool of Košarko et al. (2019).

We apply an augmentation to each training data example at random with a 50% probability (i.e., roughly 200 additional instances per augmentation type). We also consider an *"All"* setting where we include all augmented data.

Self-training: We generated parallel synthetic data of various sizes up to 2k examples. We further applied our filtering via automatic metrics (see Section 3) to choose the best 1k out of 2k examples.

Style reward We train a simple BERT-based (Devlin et al., 2019) sentiment classifier for this experiment, only using the same limited training set as for the main task. Its accuracy on our test set is 95.8%. We use this classifier for the style rewards, with a $\alpha = 0.5$, i.e., even split between the base cross-entropy loss and the style rewards.

4.3 External baselines

We compare our approaches to well-performing systems for sentiment transfer using large non-parallel datasets.[3] Our goal is to demonstrate the effectiveness of leveraging low-resource parallel data. We compare to Shen et al. (2017)'s cross-aligned autoencoder with style-specific decoders, Prabhumoye et al. (2018)'s system based on back-translation via French, and Li et al. (2018)'s text-replacement-based approach.

We also compare to state-of-the-art instruction-finetuned large language models: ChatGPT[4] and HuggingFace Chat.[5] We prompt them with a task specification and 10 randomly chosen examples from the training set. We only report results for ChatGPT, as HuggingFace Chat did not adhere to the given task, and its outputs were not parsable with our evaluation scripts.

[3] We faced difficulties when attempting to run some other recent approaches on our data (Xiao et al., 2021; Lee, 2020).

[4] `https://openai.com`, model gpt-3.5-turbo.

[5] `https://huggingface.co/chat/`, model OpenAssistant/oasst-sft-6-llama-30b (Köpf et al., 2023).

5 Evaluation & Results

We evaluate three main dimensions: style transfer accuracy, content preservation, and fluency.

We measure sentiment accuracy using DistilBERT (Sanh et al., 2019) finetuned for sentiment analysis on the SST-2 dataset (Socher et al., 2013).[6] Following prior work (Jin et al., 2022; Hu et al., 2022), we evaluate content preservation using BLEU score (Papineni et al., 2002) and embedding similarity (Rahutomo et al., 2012) against the input sentences. We use Sentence-BERT (Reimers and Gurevych, 2019) and cosine similarity for the embedding similarity. We use GPT-2's (Radford et al., 2019) perplexity to estimate fluency.

We also run a small-scale in-house human evaluation on a random sample of 100 sentences from the test set (50 for each direction – positive-to-negative and negative-to-positive). Outputs are rated on a 5-point Likert scale for style transfer accuracy, content preservation, and fluency.

5.1 Automatic Metrics Results

Table 1 shows automatic metrics results. Our base BART model (experiment 01) performs decently in all metrics, but style accuracy is further improved via hyperparameter tuning (02-04), with a slight drop in BLEU score. Adding prompts (05) further increases style accuracy and makes up for the content similarity drop.

Data augmentation (06-14) leads to further improvements, especially for replacing *Synonym*s from WordNet (09), random word *Deletion* (10), and *Back-translation* (11). The best performance is achieved using *All* (14) data augmentation types (which also means a larger number of augmented examples). Augmentation generally leads to a style accuracy increase; perplexity rises, but BLEU and embedding similarity is preserved, indicative of less frequent expressions, but not much change in content.

Self-training with synthetic data (15-20) maintained the performance across the board with a slight improvement in BLEU score, but synthesizing too many examples does not lead to further improvements (18-19), likely due to an imbalance between original and synthetic data. The best results are achieved using 1k synthesized instances filtered using automatic metrics (20).

[6] `https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english`

ID	Models	ACC	BLEU	CS	PPL
		Baseline			
01	BART-base	55.4 ± 2.6	33.8 ± 0.2	65.5 ± 0.9	127.7 ± 2.4
		Hyperparameter tuning			
02	01 + LR & batch size	61.7 ± 3.1	33.1 ± 0.2	67.6 ± 1.4	126.4 ± 1.6
03	02 + Dropout	61.1 ± 2.7	33.3 ± 0.3	67.4 ± 1.3	126.1 ± 1.2
04	03 + L2 & label smoothing	61.6 ± 3.1	33.2 ± 0.3	67.6 ± 1.4	126.9 ± 1.4
		Prompt-guided generation			
05	04 + Prompt	67.7 ± 2.6	33.3 ± 0.3	70.1 ± 1.0	126.7 ± 1.8
		Data augmentation			
06	05 + Spelling	71.1 ± 2.5	33.6 ± 0.4	70.0 ± 1.2	132.2 ± 2.2
07	05 + Insert	71.6 ± 2.4	33.1 ± 0.4	70.8 ± 1.4	131.5 ± 0.9
08	05 + Substitute	70.9 ± 3.5	33.2 ± 0.6	69.9 ± 1.2	131.9 ± 1.3
09	05 + Synonym	71.5 ± 2.7	33.5 ± 0.5	71.2 ± 2.1	131.9 ± 0.9
10	05 + Delete	72.0 ± 1.9	33.0 ± 0.5	70.7 ± 1.8	132.6 ± 0.8
11	05 + Back-translation	72.7 ± 2.5	32.9 ± 0.7	70.6 ± 1.3	132.7 ± 1.6
12	05 + Swap	71.1 ± 3.3	33.5 ± 0.1	70.1 ± 1.0	131.9 ± 1.4
13	05 + Split	70.8 ± 4.5	33.5 ± 0.4	70.5 ± 1.4	133.5 ± 0.7
14	05 + All	74.2 ± 3.2	33.2 ± 0.7	70.6 ± 2.7	132.5 ± 1.5
		Self-training			
15	05 + 250	68.4 ± 2.5	33.4 ± 0.2	69.4 ± 1.5	132.5 ± 0.4
16	05 + 500	70.5 ± 5.0	33.6 ± 0.5	71.4 ± 2.3	132.3 ± 2.2
17	05 + 1k	71.5 ± 4.8	34.1 ± 0.4	70.5 ± 2.7	131.0 ± 2.8
18	05 + 1.5k	70.1 ± 5.0	34.2 ± 0.2	70.8 ± 2.8	132.4 ± 1.2
19	05 + 2k	70.0 ± 4.6	34.3 ± 0.2	70.2 ± 2.2	132.4 ± 1.6
20	05 + 1k filtered	72.6 ± 4.4	34.2 ± 0.4	71.5 ± 2.3	132.7 ± 1.3
		Style reward			
21	14 + reward	78.8 ± 2.7	33.1 ± 0.7	72.4 ± 2.4	132.8 ± 1.5
22	20 + reward	78.4 ± 2.9	33.9 ± 0.7	72.2 ± 1.9	132.6 ± 1.2
		External baselines			
23	Shen et al.	64.4	6.7	46.0	338.5
24	Li et al.	71.9	11.6	55.3	366.6
25	Prabhumoye et al.	72.4	3.0	41.7	318.8
26	ChatGPT	95.4	19.4	61.4	115.3

Table 1: Automatic evaluation results. We measure the sentiment classifier accuracy (ACC), BLEU score, Content Similarity (CS), and Fluency (PPL), see Section 5. The model names follow a format of experiment ID + Model name, indicating that the current model is built upon a base model from that particular ID. All our models' scores are averages of five runs with different random initializations, with standard deviations shown after "±".

Models	Style	Content	Fluency
Li et al.	2.36	1.57	1.58
ChatGPT	4.48	2.75	4.49
Ours	3.98	3.96	4.45

Table 2: Human evaluation of 100 randomly selected outputs on style transfer accuracy (Style), Content Preservation (Content), and Fluency (see Section 5).

Using style rewards and combining them with data augmentation (21) or self-training (22) brings further improved style accuracy, with other metrics staying approximately the same. Since both experiments 21 and 22 perform very similarly, we choose 22 as the best model for further evaluation because the self-training approach does not require additional tools, unlike the data augmentation toolkit needed for 21.

Compared to unsupervised approaches (23-25), our experiments show similar or better style accuracy while maintaining content preservation and fluency, both of which are very low for unsupervised systems. ChatGPT (26) excelled in style transfer accuracy and fluency, but also lacked in content preservation. Table 3 (see Appendix A) shows a few illustrative examples, comparing our chosen best model (22) with external baselines.

5.2 Human Evaluation

For the human evaluation, we compared our chosen model (experiment 22) with Li et al. (2018)'s work (24) and ChatGPT (26), chosen for their best automatic metrics results of the external models. The results in Table 2 largely confirm the automatic metrics results – the unsupervised system shows relatively poor performance, and while ChatGPT excels in hitting the target style, our approach is best on content preservation.

6 Conclusion

We showed that leveraging minimal parallel data in text style transfer can yield a good balance of style transfer accuracy, content preservation, and fluency. Standard low-resource approaches (hyperparameter tuning, data augmentation, self-training), improve results, while further improvement is achieved by using style classifier rewards. In future research, we plan to extend the range of low-resource techniques used and apply our approach to other style transfer tasks.

Acknowledgments

This research was supported by the European Research Council (Grant agreement No. 101039303 NG-NLG) and by Charles University projects GAUK 392221 and SVV 260575. We acknowledge the use of resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).

References

Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3429–3435.

Junyi Chai, Reid Pryzant, Victor Ye Dong, Konstantin Golobokov, Chenguang Zhu, and Yi Liu. 2022. FAST: improving controllability for text generation with feedback aware self-training. *CoRR*, abs/2210.03167.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pages 5997–6007, Florence, Italy.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. 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 663–670.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

Md Mosharaf Hossain and Eduardo Blanco. 2022. Leveraging affirmative interpretations from negation improves natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 5833–5847, Abu Dhabi.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R. Devon Hjelm, Alessandro Sordoni, and Aaron C. Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, pages 1301–1312, Online.

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.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, Sydney, NSW, Australia.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, Denmark.

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.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. OpenAssistant conversations - democratizing large language model alignment. *CoRR*, abs/2304.07327.

Ondřej Košarko, Dušan Variš, and Martin Popel. 2019. LINDAT translation service.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA, May 6-9, 2019.

Joosung Lee. 2020. Stable style transformer: Delete and generate approach with encoder-decoder for text style transfer. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*, pages 195–204, Dublin, Ireland.

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*, pages 7871–7880, Online.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. 2019. Domain adaptive text style transfer. 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 3302–3311.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana, USA.

Junyi Li, Tianyi Tang, Jian-Yun Nie, Ji-Rong Wen, and Xin Zhao. 2022a. Learning to transfer prompts for text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pages 3506–3518, Seattle, WA, United States.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers)*, pages 4582–4597, Virtual Event.

Xiangyang Li, Xiang Long, Yu Xia, and Sujian Li. 2022b. Low resource style transfer via domain adaptive meta learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pages 3014–3026, s, WA, United States.

Fuli Luo, Peng Li, Pengcheng Yang, Jie Zhou, Yutong Tan, Baobao Chang, Zhifang Sui, and Xu Sun. 2019a. Towards fine-grained text sentiment transfer. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pages 2013–2022, Florence, Italy.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019b. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 5116–5122, Macao.

Edward Ma. 2019. NLP augmentation. `https://github.com/makcedward/nlpaug`.

Sourabrata Mukherjee, Zdenek Kasner, and Ondrej Dusek. 2022. Balancing the style-content trade-off in sentiment transfer using polarity-aware denoising. In *Text, Speech, and Dialogue - 25th International Conference, TSD 2022*, volume 13502 of *Lecture Notes in Computer Science*, pages 172–186.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 4696–4705, Vancouver, BC, Canada.

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.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*, pages 866–876, Melbourne, Australia.

Siyuan Qiu, Binxia Xu, Jie Zhang, Yafang Wang, Xiaoyu Shen, Gerard de Melo, Chong Long, and Xiaolong Li. 2020. EasyAug: An automatic textual data augmentation platform for classification tasks. In *Companion of The 2020 Web Conference 2020*, pages 249–252, Taipei, Taiwan.

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*, pages 3980–3990, Hong Kong.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Zahra Sarabi, Erin Killian, Eduardo Blanco, and Alexis Palmer. 2019. A corpus of negations and their underlying positive interpretations. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2019*, pages 158–167, Minneapolis, MN, USA.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Volume 1: Long Papers*, Berlin, Germany.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pages 211–221, Florence, Italy.

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *CoRR*, abs/2009.13818.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 6830–6841, Long Beach, CA, USA.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1631–1642, Seattle, Washington, USA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Fei Xiao, Liang Pang, Yanyan Lan, Yan Wang, Huawei Shen, and Xueqi Cheng. 2021. Transductive learning for unsupervised text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 2510–2521, Virtual Event / Punta Cana, Dominican Republic.

Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5897–5906, Stockholm, Sweden.

A Appendix: Sample Outputs

	Negative → Positive	Positive → Negative
Source	terrible menu, high prices, bad customer service .	it 's a much better option than the club scene .
Gold	nice menu , good prices , great service - for both dinner and breakfast !	i would rather go to the club than here .
Shen et al.	fantastic selection of great customer !	it 's a good experience for the whole airport i would .
Li et al.	no nonsense in service .	it 's a much better than the club scene .
Prabhumoye et al.	bad customer service with the food of this location .	she did n't go back with this place .
ChatGPT	marvelous entertainment, budget-friendly choices, exceptional atmosphere.	absolutely disastrous, it's worse than the late-night traffic.
Ours (exp. 22)	great menu, high prices, great customer service .	it's a terrible alternative to the club scene .
Source	the bad news that my vision had deteriorated made the visit even worse .	all of my clothes are returned in sparkling condition !
Gold	the good news that my vision had improved made the visit even better .	all of my clothes are returned in terrible condition !
Shen et al.	the good thing i have the whole nails made my whole gem !	all of my car here are nothing in any room .
Li et al.	the problem was the red deal by handles the night my questions did n't .	all of my clothes are returned in my condition !
Prabhumoye et al.	the worst time i have ever had to get a disappointment .	all of the food is not very good in all .
ChatGPT	the remarkable revelation of my surprise birthday party plans made the visit even more special.	The condition of all my belongings is extremely terrible!
Ours (exp. 22)	the good news that my vision had improved made the visit even better .	all of my clothes are returned in terrible condition !
Source	it's located in a slum scottsdale area and isn't accomodating.	my father has decided to upgrade my mothers engagement ring this xmas .
Gold	it 's located in a great part of scottsdale and was really accommodating .	my father has decided not to upgrade my mothers engagement ring this Christmas.
Shen et al.	cute shop in a sunday area and desert !	my son did to have my whole card to celebrate my appointment off .
Li et al.	no bueno in the north nonsense and not acknowledged a word or anything .	my father has decided to upgrade paint now .
Prabhumoye et al.	minutes later for the food and not worth the food .	my husband ordered me to get the worst service in the food .
ChatGPT	this place is family-owned, but it could greatly benefit from improving their staff.	my father has decided to downgrade my mother's engagement ring.
Ours (exp. 22)	it's located in a slum scottsdale area and is accomodating.	my father has decided not to upgrade my mothers engagement ring this xms.

Table 3: Example output comparison on samples from the test set. Sentiment marker words are colored. Note that our model balances well between style transfer accuracy and content preservation, better than others.