Chapter 4

Text Style Transfer: An Introductory Overview

Sourabrata Mukherjee
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Charles University, Czechia
mukherjee@ufal.mff.cuni.cz
ORCID: 0000-0002-1713-2769

Ondřej Dušek
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Charles University, Czechia
odusek@ufal.mff.cuni.cz
ORCID: 0000-0002-1415-1702

DOI: 10.54103/milanoup.282.c636

4.1 Abstract

Text Style Transfer (TST) is a pivotal task in natural language generation to manipulate text style attributes while preserving style-independent content. The attributes targeted in TST can vary widely, including politeness, authorship, mitigation of offensive language, modification of feelings, and adjustment of text formality. TST has become a widely researched topic with substantial advancements in recent years. This paper provides an introductory overview of TST, addressing its challenges, existing approaches, datasets, evaluation measures, subtasks, and applications. This fundamental overview improves understanding of the background and fundamentals of text style transfer.

4.2 Introduction

Natural Language Generation

Natural Language Generation (NLG) is the process of producing meaningful phrases and sentences in natural language. The main goal of NLG is to automatically produce narratives that describe, summarize, and explain the input data in a human-like manner. In other words, it generates fluent texts with minimal grammatical errors and retains the specific intended content.

Some of the popular NLG tasks include machine translation [1], dialogue systems [2], and text summarization [3]. Through these tasks, the generated text has shown to be more coherent, logical, and emotionally rich, especially with the latest approaches based on neural language models.

Controllable NLG

Most of the built NLG systems target text fluency and grammatical correctness, and do not consider any specific control over text style. This is a motivation for research on controllable text generation [4]. The aspects of text generation that are commonly controlled include topic [5–8], style [9–12], emotion [13–16], and user preferences [17–20]. Some of the applications of controllable text generation are context-based text generation [21], topic-aware text generation, [22], knowledge-enhanced text generation [23] and text style transfer [24].

Control can be applied at various stages of the neural generation process, such as input, hidden states, and decoding [25]. The Plug and Play language model (PPLM) that was proposed by [26] takes an external input, performs computations on hidden states, and then combines a pre-trained language model with one or more simple attribute classifiers that guide text generation toward the desired topic or sentiment. Another model by [27] describes a training method based on diverse ensembling that would lead models to learn distinct text styles. It can thus be inferred that end-to-end models can be equipped with the ability to control style and length. More details on how NLG can be controlled using various control strategies in the state-of-the-art models can be found in [4].

Style-Controlled Text Generation

In recent research, more attention has been paid to a subtask of controllable text generation dubbed *style-controlled text generation*, i.e., modeling and manipulating the style of the generated text [28]. The goal of this approach is to model the content of a text along with controlling its style. For example, the persona of a speaker in dialogue [29] or the sentiment of product reviews [30]. Understanding and dealing with style in text proves to be very complex [24], but recent advances in deep lear-

ning techniques are helping stylized text generation tasks in various ways [31]. For example, embedding learning techniques are used to represent style [13], and then adversarial learning is used to match content but to distinguish between different styles [30, 32, 33].

Text Style Transfer

In this paper, we will focus on Text Style Transfer (TST). TST is a task closely related to Style-Controlled Text Generation. *Style-Controlled Text Generation* aims to generate new text in a specific style. In contrast, *Text Style Transfer* is an existing text written in source style, aiming to change the text style, i.e. a text retaining most of the content but conforming to the target style. Our aim is to give a very basic introduction to the TST task. All of the sections are presented in a brief and simple manner with an illustrative number of examples. A more detailed overview can be found in [24, 25, 28, 31, 34].

The paper is organized as follows. After the introductory section, Section 4.3 provides an overview of text style transfer. Section 4.4 reflects on the challenges facing the TST task. The discussion of the existing data sets, approaches, evaluation measures and applications is presented in Sections 4.5-4.8. A short overview of the related ethical considerations is given in Section 4.9. Section 4.10 concludes the paper.

4.3 The Task

Text style transfer (TST) is an NLG task that aims to automatically control the style attributes of a text while preserving the style-independent content. Some of the attributes that TST aims to control are politeness, formality, sentiment, and many others. Table 4.1 shows some basic examples of TST. TST implies the need to understand the difference between the style and content of a text.

Table 4.1 *TST examples regarding sentiment, polarity, and formality.*

	Source Style	**Target Style**
Impolite → Polite:	Shut up! the video is starting!	Please be quiet, the video will begin shortly.
Negative → Positive:	The food is tasteless.	The food is delicious.
Informal → Formal:	The kid is freaking out.	That child is distressed.

4.3.1 Understanding Style and Content

[35] define style as a notion that refers to the manner in which semantics is expressed. Individualistic styles such as choice of words, sentence structures, metaphors, sentence arrangement, etc., vary from person to person. These variations are

shaped by the speakers's personality – everyone has a distinctive set of techniques for using the language to express and achieve their independent goals [36]. This individualistic nature also determines how a person perceives events, describes ideas, or provides additional information about them [24]. Style extends beyond individual sentences to the broader discourse level. This includes elements such as paragraph organization, theme progression, and use of cohesive devices that bind the text together. These stylistic features at the discourse level play a crucial role in ensuring that the text is coherent and engaging, thereby enhancing its ability to convey the intended message and intrigue the reader. Taking these aspects into account, the text can offer a richer and more nuanced understanding of its content.

Style has also been defined by [36] by its pragmatic aspects. Beyond these personal styles of expression, there are certain styles that are used as protocols to regularize the manner of communication. For example, in the case of academic writing, using formal expressions is the regularized protocol.

TST studies adopt a more data-driven approach to define text style in contrast to the theoretical definition used in linguistic studies [31]. We can define style in TST as the text style attributes or labels that are dependent on style-specific corpora [24]. For example, datasets are manually annotated with linguistic style definitions, such as formality [37] or sentiment [38–40]. Unfortunately, not all possible styles have very well-matched corpora, and many recent dataset collection works are looking for meta-information that would automatically link a corpus to a certain style. Some of the TST tasks are built upon the assumption that style is localized to certain tokens in a text, and a token has either content or style information, but not both [41].

In opposition to style, content can be understood as the subject matter, theme, or topics the author writes about.

4.3.2 Problem Formulation

Given a text x, with an original style S, our goal is to rephrase x into a new text $\hat{x}$ with a target style S' ($S' \neq S$) while preserving its content that is independent of style.

Suppose that we have a dataset $X_S = x_1^{(S)}, \ldots, x_m^{(S)}$ representing texts in style S. The task is to transform texts in style S to the target style S' while maintaining the original meaning. We denote the output of this transformation by $X_{S \to S'} = \hat{x}1^{(S')}, \ldots, \hat{x}m^{(S')}$. Similarly, for the inverse transformation from style S' to style S, we denote the output as $X_{S' \to S} = \hat{x}1^{(S)}, \ldots, \hat{x}n^{(S)}$.

4.4 Challenges

Modeling the style of text comes with a lot of challenges in practice, which are discussed in this section.

No Parallel Data

TST models could be trained with respect to parallel text from a given style or on non-parallel corpora. Parallel datasets are those which consist of pairs of texts (i.e. sentences, paragraphs) where each text in the pair expresses the same meaning, but in a different style. Non-parallel datasets, on the other hand, have no paired examples to learn from, and simply exist as mono-style corpora. For parallel datasets, TST can be formulated in such a way that instead of translating between languages, one can translate between styles following machine translation. However, obtaining suitable, sufficient parallel data for each desired style attribute is the biggest challenge.

Style and content are hard to separate

Style transfer text generation implies the need to distinguish content from style. In some scenarios, the line between content and style can be blurry. This is since the subject on which an author is writing can also influence their choice of words and style. This interweaving of the style and semantics makes TST challenging.

No Standard Evaluation Measures

Evaluating the quality of the style-transferred text is hard. Human evaluation is regarded as the best indicator of quality, but unfortunately, it is expensive, slow, and hard to reproduce [42], making it an infeasible approach to use on a daily basis to validate model performance. For this reason, we often rely on automated evaluation metrics to serve as a cheap and quick proxy for human judgment.

In the case of automatic evaluation of TST, it has been noticed that when style transfer accuracy increases, the content preservation scores decrease, and vice versa [43]. The main reason behind this is the entanglement between the content and style (see above). This trade-off between style transfer accuracy and content preservation poses a very big challenge for evaluating TST tasks.

In order to effectively evaluate a TST output, one must pay attention to how semantically accurate the output text is and how fluent it is. The comprehensive TST evaluation also considers three criteria: transferred style accuracy, semantic preservation, and fluency, which often require human evaluation as automated metrics alone do not adequately identify these complex properties. Further discussion on evaluation measures is in Section 4.7.

4.5 Datasets and Benchmarks

To evaluate TST models, many datasets have been proposed over the years. We discuss a few popular datasets by individual subtasks as follows:

Politeness Transfer

Politeness transfer aims to control the politeness of a text [44, 45]. A compiled dataset with automatically labeled instances from the raw Enron e-mail corpus [46] was presented by [45]. This dataset mainly focuses on politeness in North American English.

Sentiment Transfer

Another common task in TST is sentiment transfer (transferring text's polarity from positive to negative or vice-versa) [43, 47]. There are three popular datasets proposed for this task.

- Yelp – This is a corpus consisting of restaurant reviews from Yelp collected by [38].
- Amazon – This is Amazon's product reviews that were collected by [39].
- IMDb – This is a movie review dataset constructed by [40].

Formality Transfer

Formality transfer is yet another task in TST which is not only complex but also involves multiple attributes that affect text formality. Grammarlys Yahoo Answers Formality Corpus (GYAFC) is the largest human-labeled parallel dataset that was proposed for formality transfer tasks by [37]. The authors extracted informal sentences from the Entertainment & Music and Family & Relationship domains of the Yahoo Answers L6 corpus for preparing the dataset.

Author's Style Re-writing

The task of paraphrasing a sentence to match a specific author's style is called author imitation. To tackle such tasks, [48] collected a parallel dataset that captured line-by-line modern interpretations of 16 Shakespeare's plays, with the help of the educational site Sparknotes.[1] The objective behind collecting the dataset was to imitate Shakespeares text style by transferring modern English sentences into Shakespearean-style sentences. This dataset has been used in other TST studies as well [49, 50].

Image Captions Transfer

The task of transferring image captions from factual formal ones to romantic and humorous styles was proposed by [9]. Following this, a caption dataset was

[1] `https://www.sparknotes.com`

collected by the authors where each sentence was labeled as factual, romantic, or humorous.

Text Simplification

Another important use of TST is to lower the language barrier for readers, which includes tasks like converting general English into Simple English, based on a dataset collected from Wikipedia [51]. Another task is to simplify medical descriptions to patient-friendly text [52].

Political-slant Transfer

Political slant transfer is a task that modifies a writer's political affiliation writing style while preserving the content. Comments from Facebook posts from 412 members of the United States Senate and House who have public Facebook pages were collected by [11] and further annotated with each congresspersons political party affiliation, i.e., Democrat or Republican.

Fixing offensive texts

Correcting offensive and abusive language [53] is another important task of TST which is a major problem in today's world, due to the prevalence of abusive comments on social media. Posts from Twitter and Reddit were collected by [54] and then classified into *offensive* and *non-offensive* classes using a classifier pre-trained on an annotated offensive language dataset.

4.6 Text Style Transfer Approaches

Standard data-driven TST approaches can be classified based on the data used for training (parallel vs. non-parallel). Recently, new approaches using large language models (LLMs) emerged that do not specifically need in-domain training data.

4.6.1 Supervised Training on Parallel Data

For situations where style-parallel data is available, like most supervised methods, a standard sequence-to-sequence model [47, 55, 56] with the encoder-decoder structure is typically used [24]. This process is similar to machine translation and text summarization. The encoder-decoder architecture can be implemented by either LSTM [57] or the Transformer [58] architecture. For example, [49] trained a sequence-to-sequence model on a parallel corpus and then applied the model to translate modern English phrases to Shakespearean English. However, the

application of basic sequence-to-sequence approaches is quite limited due to the lack of parallel data (see Section 4.4).

4.6.2 Non-parallel Approaches

Methods applicable to non-parallel data can broadly be divided into three unsupervised approaches:

Prototype Editing

This process works by deleting only the parts of the sentences which represent the source style and replacing them with words with the target style while making sure that the resulting text is still fluent. The advantage of this approach is its simplicity and explainability. For example, [9, 59] found that parts of a text that are associated with the original style can be replaced with new phrases associated with the target style. The text was then fed into a sequence-to-sequence model to generate a fluent text sequence in the target style. However, these approaches are not suitable for TST applications where simple phrase replacement is not enough or a correct way to transfer style. The style marker retrieval might not work if the datasets have confounded style and contents. This is because they may lead to the incorrect extraction of style markers, affecting some content words.

Disentanglement

This approach aims at disentangling the text into its content and style in an embedding latent space, then applies generative modeling. TST models first learn the latent representations of the content and style of the given text. The latent representation of the original content is then combined with the latent representation of the desired target style to generate text in the target style. Techniques such as back-translation [11, 43, 60] and adversarial learning [13, 38, 61] have been proposed to disentangle latent representations into content and style. In general, total disentanglement is impossible without inductive biases or some other forms of supervision [62].

Pseudo-Parallel Corpus Creation

This process is used to train the model in a supervised way by generating pseudo-parallel data. One way of constructing pseudo-parallel data is through retrieval, i.e., extracting aligned sentence pairs from two mono-style corpora. [63] constructed pseudo-parallel corpora by matching sentence pairs in two style-specific corpora according to cosine similarity over pre-trained sentence embeddings. The

constructed pseudo-parallel corpora must reach a certain level of quality to be useful for TST.

4.6.3 Using Large Language Models

LLMs have revolutionized the field of natural language processing by generating coherent and contextually relevant text e.g., [64, 65]. By learning from vast amounts of text data, LLMs capture various linguistic styles and nuances. This capability is particularly beneficial for TST tasks.

A distinctive feature of LLMs is their ability to perform valuable tasks without fine-tuning, showcasing zero- and few-shot capabilities [66]. Style transfer has been framed as a sentence rewriting task, enhancing LLMs' zero-shot performance for arbitrary TST by using task-related exemplars [67]. A reranking method has been proposed to select high-quality outputs from multiple candidates generated by the LLM, thereby improving performance [68]. Additionally, dynamic prompt generation has been introduced to guide the language model in producing text in the desired style [69].

While prompt engineering is the prevalent approach [70, 71], LLMs are highly sensitive to prompts [72, 73] and may not always guarantee optimal performance [69]. Despite good results for prompting, finetuning the LLMs still leads to significant performance improvements [74].

4.7 Evaluation Measures

A successful style transfer output is one that portrays the correct target style along with preserving the original semantics of the text and maintaining natural language fluency.

4.7.1 Automatic Evaluation

Automatic evaluation metrics provide an economic, reproducible, and scalable way to assess the quality of generation results. There are several automated evaluation metrics that have been proposed to measure the effectiveness of TST models [75–78]. They can be divided into three different categories based on the aspect of TST they focus on:

Style Transfer Strength

The ability to transfer the text style or the transfer strength of a TST model is measured using Style Transfer Accuracy [13, 30, 38, 79, 80]. Mostly, a binary style classifier [81] is pre-trained separately to predict the style label of the input sentence and

is then used to estimate the style transfer accuracy of the transferred style sentence. This is done by considering the target style as the ground truth.

Content Preservation

In order to measure the amount of original content preserved after the style transfer procedure, some automated evaluation metrics from other NLG tasks have been adopted for TST. For instance, the BLEU word-n-gram-overlap metric [82] is computed similarly as with machine translation. Match against a target-style sentence can be computed when parallel TST datasets or target-style human references are available. Since most of the TST tasks assume a non-parallel setting and matching references of style transferred sentences are not always a feasible option, evaluation using *source-BLEU (sBLEU)* is adopted. In this method, a transferred sentence is compared to its source. The overlap with the source is considered a proxy for content preservation. Cosine Similarity [83] can also be calculated between the original sentence embeddings and the transfer sentence embeddings [13]. This methodology follows the idea that the embeddings of the two sentences should be close if most of the semantics are preserved.

Fluency

One of the most common goals for all NLG tasks is producing fluent outputs. A common approach to measuring the fluency of a sentence is using a language model [84]: A pre-trained language model is used to compute the perplexity scores of the style-transferred sentences to evaluate the sentences' fluency.

4.7.2 Human Evaluation

Human evaluation stands out from automatic evaluation due to its flexibility and comprehensiveness. However, this evaluation approach is very challenging since the interpretation of text style can be subjective and vary from individual to individual [75, 76, 78]. In spite of this shortcoming, human evaluations still offer valuable insights into how well the TST algorithms can transfer style and generate sentences that are acceptable according to human standards.

In terms of evaluation types, there is point-wise scoring, wherein humans are asked to provide absolute scores of the model outputs (e.g. on a 1-5 Likert scale), and pairwise comparison, wherein they are asked to judge which of two outputs is better, or by providing a ranking for multiple outputs.

4.8 Applications

TST has a wide range of downstream applications in various NLP fields that include stylized chatbots [85], stylized writing assistants, automatic text simplification,

debiasing online text and even fighting against offensive language. A few very popular examples are discussed below.

[86] carried out a study that showcased the impact of chatbot's conversational style on users. [87] encoded personas of individuals in contextualized embeddings that helped in capturing the background information and style to maintain consistency in the generated responses. [88] focused on generating polite personalized dialog responses in agreement with the user's profile and consistent with their conversational history.

Another important application of TST is enhancing the human writing experience [89–91]. This application aids in people restyling their content to appeal to a variety of audiences, i.e., making a text polite, humorous, professional, or even Shakespearean.

Another inspiring application of TST is automatically simplifying content for better communication between experts and non-expert individuals in certain knowledge domains, thus lowering language barriers. For example, complicated legal, medical, or technical jargon is transferred into simple terms that a layman can comprehend [92].

TST can also offer a means to neutralize subjective attitudes for certain texts where objectivity is strongly needed. For example, in the domains of news, encyclopedia, and textbooks. Such applications can help in reshaping gender roles that are portrayed in writing [93]. TST can also help in transforming hateful sentences into non-hateful ones. For instance, [94] propose an extension of a basic encoder-decoder architecture by including a collaborative classifier to deal with abusive language redaction.

4.9 Ethical Concerns

An essential part of research is to consider the ethical implications of the project through its potential benefits and risks.

For example, TST has the potential to reduce toxicity, hate speech, sexist and racist language, aggression, harassment, trolling, and cyberbullying [95]. This task is beneficial for modeling non-offensive text to help reduce toxicity on social media platforms [54,96]. It can also be used on social chatbots to make sure there is no bad content in the generated text [97]. TST is also able to neutralize subjective-toned language, which can be helpful for certain types of publications such as textbooks [98].

However, the same technology can also be misused to purposely generate the opposite attribute, i.e., generating hateful, offensive text, that counters any intended social benefit [99]. Furthermore, as TST is now generally performed using trained language models, these inherit all the potential risks associated with this technology in general, such as reflecting unjust, toxic, or oppressive speech present in the training data [100].

The goal of a discussion on ethics is to take into account various concerns like how a system should be built, who it is intended for, and how to assess its societal impact [99] [101]. Instead of abandoning the whole idea of building such tools, one must explore the concerns and find ways to deal with them [102]. This should be viewed as an opportunity to increase transparency by surfacing the risks and finding the best ways to its strategy into practice.

4.10 Conclusion

The main goal of this work is to offer an introductory overview of Text Style Transfer (TST), highlighting key components such as subtasks, datasets, evaluation methods, and the challenges inherent to TST. Additionally, we discussed the ethical considerations surrounding this area of research. We aim for this overview to serve as a useful guide for those new to the field.

Acknowledgment

This research was funded by the European Union (ERC, NG-NLG, 101039303) and by Charles University projects GAUK 392221 and SVV 260698.

Bibliography

[1] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[2] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," *arXiv preprint arXiv:1503.02364*, 2015.

[3] A. M. Rush, S. Harvard, S. Chopra, and J. Weston, "A neural attention model for sentence summarization," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 2017.

[4] Y. Len, F. Portet, C. Labbé, and R. Qader, "Controllable neural natural language generation: comparison of state-of-the-art control strategies," in *Proc. of the 3rd Workshop on Natural Language Generation from the Semantic Web*, 2020.

[5] N. Dziri, E. Kamalloo, K. Mathewson, and O. Zaiane, "Augmenting neural response generation with context-aware topical attention," in *Proc. of the First Workshop on NLP for Conversational AI.* Association for Computational Linguistics, Aug. 2019, pp. 18–31.

[6] X. Feng, M. Liu, J. Liu, B. Qin, Y. Sun, and T. Liu, "Topic-to-essay generation with neural networks." in *IJCAI*, 2018, pp. 4078–4084.

[7] L. Wang, J. Yao, Y. Tao, L. Zhong, W. Liu, and Q. Du, "A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization," in *Proc. of the 27 Int. Joint Conf. on Artificial Intelligence*, 7 2018, pp. 4453–4460.

[8] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W.-Y. Ma, "Topic aware neural response generation," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 31, no. 1, 2017.

[9] J. Li, R. Jia, H. He, and P. Liang, "Delete, retrieve, generate: a simple approach to sentiment and style transfer," in *Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1865–1874.

[10] A. Sudhakar, B. Upadhyay, and A. Maheswaran, "transforming delete, retrieve, generate approach for controlled text style transfer," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3260–3270.

[11] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black, "Style transfer through back-translation," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 866–876.

[12] L. Chen, S. Dai, C. Tao, H. Zhang, Z. Gan, D. Shen, Y. Zhang, G. Wang, R. Zhang, and L. Carin, "Adversarial text generation via feature-mover's distance," in *Advances in Neural Information Processing Systems*, 2018, pp. 4666–4677.

[13] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan, "Style transfer in text: Exploration and evaluation," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 32, no. 1, 2018.

[14] X. Kong, B. Li, G. Neubig, E. Hovy, and Y. Yang, "An adversarial approach to high-quality, sentiment-controlled neural dialogue generation," *arXiv preprint arXiv:1901.07129*, 2019.

[15] X. Sun, J. Li, X. Wei, C. Li, and J. Tao, "Emotional editing constraint conversation content generation based on reinforcement learning," *Inf. Fusion*, vol. 56, pp. 70–80, 2020.

[16] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 32, no. 1, 2018.

[17] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Aug.

2016, pp. 994–1003.

[18] Y. Luan, C. Brockett, B. Dolan, J. Gao, and M. Galley, "Multi-task learning for speaker-role adaptation in neural conversation models," in *Proc. of the 8 Int. Joint Conf. on Natural Language Processing (Volume 1: Long Papers)*, Nov. 2017, pp. 605–614.

[19] M. Yang, Q. Qu, K. Lei, J. Zhu, Z. Zhao, X. Chen, and J. Z. Huang, "Investigating deep reinforcement learning techniques in personalized dialogue generation," in *Proc. of the SIAM Int. Conf. on Data Mining*, 2018, pp. 630–638.

[20] M. Yang, Z. Zhao, W. Zhao, X. Chen, J. Zhu, L. Zhou, and Z. Cao, "Personalized response generation via domain adaptation," in *Proc. of the 40th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2017, pp. 1021–1024.

[21] A. Jaech, and M. Ostendorf, "Low-rank RNN adaptation for context-aware language modeling," *Trans. of the Association for Computational Linguistics*, vol. 6, pp. 497–510, 2018.

[22] L. Wang, J. Yao, Y. Tao, L. Zhong, W. Liu, and Q. Du, "A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization," *arXiv preprint arXiv:1805.03616*, 2018.

[23] T. Young, E. Cambria, I. Chaturvedi, H. Zhou and S. Biswas, and M. Huang, "Augmenting end-to-end dialogue systems with commonsense knowledge," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 32, no. 1, 2018.

[24] Z. Hu, R. K.-W. Lee, C. C. Aggarwal, and A. Zhang, "Text style transfer: A review and experimental evaluation," *ACM SIGKDD Explorations Newsletter*, vol. 24, no. 1, pp. 14–45, 2022.

[25] S. Prabhumoye, A. W. Black, and R. Salakhutdinov, "Exploring controllable text generation techniques," *arXiv preprint arXiv:2005.01822*, 2020.

[26] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, "Plug and play language models: A simple approach to controlled text generation," *arXiv preprint arXiv:1912.02164*, 2019.

[27] S. Gehrmann, F. Z. Dai, H. Elder, and A. M. Rush, "End-to-end content and plan selection for data-to-text generation," *arXiv preprint arXiv:1810.04700*, 2018.

[28] L. Mou and O. Vechtomova, "Stylized text generation: Approaches and applications," in *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2020, pp. 19–22.

[29] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model," *arXiv preprint arXiv:1603.06155*, 2016.

[30] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *Proc. of the Int. Conf. on Machine Learning*. PMLR, 2017, pp. 1587–1596.

[31] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, and R. Mihalcea, "Deep learning for text style transfer: A survey," *Computational Linguistics*, vol. 48, no. 1, pp. 155–205, 2022.

[32] J. Xu, X. Sun, Q. Zeng, X. Ren, X. Zhang, H. Wang, and W. Li, "Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach," *arXiv preprint arXiv:1805.05181*, 2018.

[33] V. John, L. Mou, H. Bahuleyan, and O. Vechtomova, "Disentangled representation learning for non-parallel text style transfer," *arXiv preprint arXiv:1808.04339*, 2018.

[34] M. Toshevska and S. Gievska, "A review of text style transfer using deep learning," *IEEE Trans. on Artificial Intelligence*, 2021.

[35] D. D. McDonald and J. Pustejovsky, "A computational theory of prose style for natural language generation," in *Proc. of the Conf. of the European Chapter of the Association for Computational Linguistics*, 1985.

[36] E. Hovy, "Generating natural language under pragmatic constraints," *Journal of Pragmatics*, vol. 11, no. 6, pp. 689–719, 1987.

[37] S. Rao and J. Tetreault, "Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer," *arXiv preprint arXiv:1803.06535*, 2018.

[38] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, "Style transfer from non-parallel text by cross-alignment," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[39] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. of the 25th Int. Conf. on World Wide Web*, 2016, pp. 507–517.

[40] N. Dai, J. Liang, X. Qiu, and X.-J. Huang, "Style transformer: Unpaired text style transfer without disentangled latent representation," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5997–6007.

[41] D. Lee, Z. Tian, L. Xue, and N. L. Zhang, "Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization," *arXiv preprint arXiv:2108.00449*, 2021.

[42] A. Belz, S. Agarwal, E. Reiter, and A. Shimorina, "Reprogen: Proposal for a shared task on reproducibility of human evaluations in NLG," 2020.

[43] S. Mukherjee, Z. Kasner, and O. Dušek, "Balancing the style-content trade-off in sentiment transfer using polarity-aware denoising," in *Proc. of the Int. Conf. on Text, Speech, and Dialogue*, 2022, pp. 172–186.

[44] S. Mukherjee, V. Hudeček, and O. Dušek, "Polite chatbot: A text style transfer application," in *Proc. of the 17th Conf. of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, May 2023, pp. 87–93.

[45] A. Madaan, A. Setlur, T. Parekh, B. Poczos, G. Neubig, Y. Yang, R. Salakhutdinov, A. W. Black, and S. Prabhumoye, "Politeness transfer: A tag and generate approach," *arXiv preprint arXiv:2004.14257*, 2020.

[46] J. Shetty and J. Adibi, "The enron email dataset database schema and brief statistical report," *Information sciences institute technical report, University of Southern California*, vol. 4, no. 1, pp. 120–128, 2004.

[47] S. Mukherjee, A. Bansal, P. Majumdar, A. K. Ojha, and O. Dušek, "Low-resource text style transfer for Bangla: Data & models," in *Proc. of the First Workshop on Bangla Language Processing*, Dec. 2023, pp. 34–47.

[48] W. Xu, A. Ritter, B. Dolan, R. Grishman, and C. Cherry, "Paraphrasing for style," in *Proc. of COLING 2012*, 2012, pp. 2899–2914.

[49] H. Jhamtani, V. Gangal, E. Hovy, and E. Nyberg, "Shakespearizing modern language using copy-enriched sequence to sequence models," in *Proc. of the Workshop on Stylistic Variation*, 2017, pp. 10–19.

[50] J. He, X. Wang, G. Neubig, and T. Berg-Kirkpatrick, "A probabilistic formulation of unsupervised text style transfer," in *Proc. of the Int. Conf. on Learning Representations*, 2020.

[51] Z. Zhu, D. Bernhard, and I. Gurevych, "A monolingual tree-based translation model for sentence simplification," in *Proc. of the 23rd Int. Conf. on Computational Linguistics*, 2010, pp. 1353–1361.

[52] L. Van den Bercken, R.-J. Sips, and C. Lofi, "Evaluating neural text simplification in the medical domain," in *Proc. of the World Wide Web Conf.*, 2019, pp. 3286–3292.

[53] M. Sourabrata, B. Akanksha, K. O. Atul, P. M. John, and D. Ondrej, "Text detoxification as style transfer in English and Hindi," in *Proc. of the 20th Int. Conf. on Natural Language Processing*, Dec. 2023, pp. 133–144.

[54] C. dos Santos, I. Melnyk, and I. Padhi, "Fighting offensive language on social media with unsupervised text style transfer," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 189–194.

[55] 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.

[56] S. Mukherjee, A. K. Ojha, A. Bansal, D. Alok, J. P. McCrae, and O. Dušek, "Multilingual text style transfer: Datasets & models for Indian languages," *arXiv preprint arXiv:2405.20805*, 2024.

[57] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[59] S. Mukherjee, A. Bansal, P. Majumdar, A. K. Ojha, and O. Dušek, "Low-resource text style transfer for bangla: Data & models," in *Proc. of the First Workshop on Bangla Language Processing*, 2023, pp. 34–47.

[60] Z. Zhang, S. Ren, S. Liu, J. Wang, P. Chen, M. Li, M. Zhou, and E. Chen, "Style transfer as unsupervised machine translation," *CoRR*, vol. abs/1808.07894, 2018.

[61] J. Zhao, Y. Kim, K. Zhang, A. M. Rush, and Y. LeCun, "Adversarially regularized autoencoders," in *Proc. of the 35th Int. Conf. on Machine Learning*, 2018, pp. 9405–9420.

[62] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *international Conf. on machine learning*, 2019, pp. 4114–4124.

[63] Z. Jin, D. Jin, J. Mueller, N. Matthews, and E. Santus, "IMaT: Unsupervised text attribute transfer via iterative matching and translation," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing*, 2019, pp. 3088–3100.

[64] H. Touvron, T. Lavril, G. Izacard, *et al.*, "LLaMA: Open and Efficient Foundation Language Models, CoRR, vol. abs/2302.13971, 2023.

[65] H. Touvron, L. Martin, K. Stone, *et al.*, "Llama 2: Open Foundation and Fine-tuned Chat Models, CoRR, vol. abs/2307.09288, 2023.

[66] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 195:1–195:35, 2023.

[67] E. Reif, D. Ippolito, A. Yuan, A. Coenen, C. Callison-Burch, and J. Wei, "A recipe for arbitrary text style transfer with large language models," *arXiv preprint arXiv:2109.03910*, 2021.

[68] M. Suzgun, L. Melas-Kyriazi, and D. Jurafsky, "Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models," *arXiv preprint arXiv:2205.11503*, 2022.

[69] Q. Liu, J. Qin, W. Ye, H. Mou, Y. He, and K. Wang, "Adaptive prompt routing for arbitrary text style transfer with pre-trained language models," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 18 689–18 697.

[70] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[71] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?" *Trans. of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020.

[72] S. Mishra, D. Khashabi, C. Baral, Y. Choi, and H. Hajishirzi, "Reframing instructional prompts to gptk's language," *arXiv preprint arXiv:2109.07830*, 2021.

[73] K. Zhu, J. Wang, J. Zhou, Z. Wang, H. Chen, Y. Wang, L. Yang, W. Ye, N. Z. Gong, Y. Zhang *et al.*, "Promptbench: Towards evaluating the robustness of large language models on adversarial prompts," *arXiv preprint arXiv:2306.04528*, 2023.

[74] S. Mukherjee, A. K. Ojha, and O. Dušek, "Are large language models actually good at text style transfer?" *arXiv preprint arXiv:2406.05885*, 2024.

[75] R. Y. Pang, "Towards actual (not operational) textual style transfer auto-evaluation," in *Proc. of the 5th Workshop on Noisy User-generated Text*, 2019, pp. 444–445.

[76] ——, "The daunting task of real-world textual style transfer auto-evaluation," *CoRR*, vol. abs/1910.03747, 2019.

[77] R. Y. Pang and K. Gimpel, "Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer," in *Proc. of the 3rd Workshop on Neural Generation and Translation*, 2019, pp. 138–147.

[78] R. Mir, B. Felbo, N. Obradovich, and I. Rahwan, "Evaluating style transfer for text," in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies)*, 2019, pp. 495–504.

[79] F. Luo, P. Li, J. Zhou, P. Yang, B. Chang, X. Sun, and Z. Sui, "A dual reinforcement learning framework for unsupervised text style transfer," in *Proc. of the 28th Int. Joint Conf. on Artificial Intelligence*. AAAI Press, 2019, pp. 5116–5122.

[80] V. John, L. Mou, H. Bahuleyan, and O. Vechtomova, "Disentangled representation learning for non-parallel text style transfer," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 424–434.

[81] A. Moschitti, B. Pang, and W. Daelemans, Eds., *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 2014.

[82] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 311–318.

[83] F. Rahutomo, T. Kitasuka, and M. Aritsugi, "Semantic cosine similarity," in *The 7th Int. Student Conf. on Advanced Science and Technology*, vol. 4, no. 1, 2012, p. 1.

[84] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, 1995, pp. 181–184 vol.1.

[85] S. Mukherjee, "Stylized dialog response generation," in *Proc. of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue*

Systems, 2023, pp. 44–46.

[86] S. Kim, J. Lee, and G. Gweon, "Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality," in *Proc. of the CHI Conf. on Human Factors in Computing Systems*, 2019, pp. 1–12.

[87] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and W. B. Dolan, "A persona-based neural conversation model," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

[88] M. Firdaus, A. Shandilya, A. Ekbal, and P. Bhattacharyya, "Being polite: Modeling politeness variation in a personalized dialog agent," *IEEE Trans. on Computational Social Systems*, 2022.

[89] F. Can and J. M. Patton, "Change of writing style with time," *Computers and the Humanities*, vol. 38, no. 1, pp. 61–82, 2004.

[90] B. Johnstone, "Stance, style, and the linguistic individual," *Stance: Sociolinguistic Perspectives*, pp. 29–52, 2009.

[91] V. G. Ashok, S. Feng, and Y. Choi, "Success with style: Using writing style to predict the success of novels," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 2013, pp. 1753–1764.

[92] Y. Cao, R. Shui, L. Pan, M.-Y. Kan, Z. Liu, and T.-S. Chua, "Expertise style transfer: A new task towards better communication between experts and laymen," in *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[93] E. Clark, A. S. Ross, C. Tan, Y. Ji, and N. A. Smith, "Creative writing with a machine in the loop: Case studies on slogans and stories," in *Proc. of the Int. Conf. on Intelligent User Interfaces*, 2018, pp. 329–340.

[94] C. N. d. Santos, I. Melnyk, and I. Padhi, "Fighting offensive language on social media with unsupervised text style transfer," *arXiv preprint arXiv:1805.07685*, 2018.

[95] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding abuse: A typology of abusive language detection subtasks," *arXiv preprint arXiv:1705.09899*, 2017.

[96] S. Harrison. (2019) Twitter and Instagram unveil new ways to combat hate–again. [Online]. Available: https://www.wired.com/story/twitter-instagram-unveil-new-ways-combat-hate-again/

[97] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y. Boureau, and J. Weston, "Recipes for building an open-domain chatbot," in *Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics*, 2021, pp. 300–325.

[98] R. Pryzant, R. D. Martinez, N. Dass, S. Kurohashi, D. Jurafsky, and D. Yang, "Automatically neutralizing subjective bias in text," in *Proc. of the aaai Conf. on artificial intelligence*, vol. 34, no. 01, 2020, pp. 480–489.

[99] D. Hovy and S. L. Spruit, "The social impact of natural language processing," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 591–598.

[100] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh *et al.*, "Taxonomy of risks posed by language models," in *Proc. of the ACM Conf. on Fairness, Accountability, and Transparency*, 2022, pp. 214–229.

[101] T. L. Beauchamp and J. F. Childress, "Respect for autonomy," *Principles of biomedical ethics*, vol. 5, pp. 57–112, 2001.

[102] J. L. Leidner and V. Plachouras, "Ethical by design: Ethics best practices for natural language processing," in *Proc. of the First ACL Workshop on Ethics in Natural Language Processing*, 2017, pp. 30–40.