Text Style Transfer: A Comprehensive Study on Methodologies and Evaluation

Nirali Parekh$^{[0000-0002-0726-6464]}$, Siddharth Trivedi$^{[0000-0002-1049-621X]}$, and Kriti Srivastava$^{[0000-0001-7262-9201]}$

Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra 400056, India
nirali25parekh@gmail.com
siddharthtrivedi19@gmail.com
kriti.srivastava@djsce.ac.in

Abstract. Text Style Transfer (TST) rewords a sentence from one style (e.g. polite) to another (e.g. impolite) while conserving the meaning and content. This domain has attracted the attention of many researchers as it makes natural language generation (NLG) tasks more user-oriented. TST finds its applications widely in industry such as conversational bots and writing assistance tools. With the success of deep learning, a plethora of research works on style transfer based on Machine Learning have been proposed, developed and tested. This systematic review presents the past work on Text Style Transfer clustered into categories based on Machine Learning and Deep Learning algorithms. It briefly explains the various subtasks within Text Style Transfer and assembles its publicly available datasets. It also summarizes the automatic and manual evaluation practices used for style transfer tasks and finally, sheds some light on current challenges and points towards promising future directions for research in TST domain.

Keywords: Text Style Transfer, Deep Learning, Natural Language Generation, Natural Language Processing, Neural Networks

1 Introduction

In the domains of Deep learning and Artificial intelligence, style transfer has recently been a hot issue of research and development. It deals with transferring the style (or attributes) of a source content into a target style. Substantial research work involving state-of-the-art algorithms [7, 44] developed for image style transfer have demonstrated astounding results. Style transfer techniques have also recently influenced the audio domain establishing methods for Music Style transfer [4]. Such works have proved the potential of deep learning techniques for style transfer in generating artificial style-transferred content. In the domain of text and linguistics, style is a highly subjective phrase that can be interchanged with the term attribute. The style-specific characteristics of the text tend to vary across situations while the style-independent content is maintained. For instance, ‘If you have any further requirements, please do not hesitate to contact me' is used in a formal setting and can be easily paraphrased to 'Let me know if you need anything else' for usage in an informal context. The goal in a Text Style Transfer (TST) problem is to generate a style-controlled text in a target style while preserving the semantics and content of the source text. Text style transfer models can be formulated [12] as: $p\left(x \mid a, x'\right)$, where x' is a source text with attribute value a' and x is the style-transferred text with target attribute a.

TST methods have developed from classic replacement and template-based approaches to neural network-based strategies as deep-learning progresses. TST can also be formulated as a Natural Language Generation (NLG) problem as it extend NLG techniques while manipulating the attributes of the text. A wide range of deep-learning techniques employed for TST tasks like Adversarial Learning, Sequence-to-Sequence Learning and Reinforcement learning-based methods which are covered in detail in section 3. Experimentation on Text Style Transfer techniques is largely classified based on subtasks under the domain. List of some common subtasks in TST is elucidated in the Table 1.

Table 1: Subtasks within Text Style Transfer

Subtask	Attribute transfer	Example
Formality	Formal	Please accept our apologies for any inconvenience
	Informal	We're sorry for the trouble
Sentiment	Positive	I admire my college professors a lot
	Negative	I hate my college professors
Politeness	Polite	Sorry, I'm a bit busy right now
	Impolite	Leave me alone
Simplicity	Complicated	Can I acquire assistance in deciphering this conundrum?
	Simple	Can you help me solve this problem?
Gender	Masculine	My wife went to the mall to buy a skirt
	Feminine	My husband went to the mall to buy a shirt
Authorship	Shakespearean	Hast thou slain tybalt?
	Modern	Have you killed tybalt?

Impactful applications of Text Style Transfer in NLP research, like paraphrasing, as well as commercial uses such as AI-assisted writing tools have driven the burgeoning interest of NLP researchers into this domain. This rapid growth in TST research has produced a variety of datasets, implementation algorithms and evaluation metrics for this task, but also at the same time lacks a sense of standardization for the same. The objective of this review paper is to present an account of various corpora and TST methodologies that could facilitate further research and uniformity in the field of TST. The contributions of this work are:

1. We conduct a comprehensive survey that reviews recent works on TST based on machine learning.
2. We describe the various machine learning architectures and evaluation metrics used in Text Style Transfer.
3. We provide a systematic summary of contributions and evaluation practices in Table 3, Table 4, Table 5.

The organization of this paper starts with a discussion on some of the publicly available datasets is done in section 2. The Machine Learning algorithms used in TST research are explored in section 3. section 4 presents few metrics used for evaluating TST algorithms. Finally, we conclude in section 5 and discuss some open issues and research scope in TST.

2 Datasets

2.1 Parallel and Non-parallel data

The datasets available for Text Style Transfer are classified into two categories widely based on data used for training.

Parallel In parallel data, the texts of both the source and target style are available. Here, simple machine translation techniques such as sequence-to-sequence can be used. The issue with parallel data is that they are not readily available in various sub-styles and data collection is very expensive.

Non-parallel Costly and scarce parallel data has led to researchers to utilize non-parallel data for Text Style Transfer. Non-parallel corpus consists of non-matching texts from source and target styles. This category of data is readily available in various styles and hence a large proportion of works on TST utilizes non-parallel datasets.

2.2 Available benchmark Datasets for TST

Table 2 records various publicly available datasets distinguished by their subtasks, sizes, whether they are parallel or non-parallel, and annotation method incase they are parallel.

3 Methodologies

3.1 Parallel

Style transfer by means of a style parallel corpus is considered a monolingual machine translation task.

Table 2: Publicly available datasets for TST

Subtask	Dataset	Size	Parallel	Domain	Annotation
Formality	GYAFC	52K	✓	Yahoo Answers (online)	Manual
Politeness	Politeness	1.39M	✗	Emails	-
Gender	Yelp	2.5M	✗	Restaurant Reviews	-
Humor & Romance	Flickr style	5K	✓	image captions	Manual
biasedness	Wiki Neutrality	181K	✓	Wikipedia (online)	Automatic
Toxicity	Twitter	58K	✗	tweets (online)	-
Toxicity	Reddit	224K	✗	politics threads	-
Authorship	Shakespeare	18K	✓	literature, SparkNotes	Automatic
Sentiment	Yelp	150K	✗	restaurant reviews	-
sentiment	Amazon	277K	✗	clothing reviews	-
Fluency	SWBD	192K	✓	telephonic conversations	Manual
Politics	Political	540K	✓	facebook posts	-

Sequence To Sequence A seq2seq model known as encoder-decoder architecture converts sequences from one domain to another. Basis to many machine translation algorithms, many research works on TST utilized seq2seq neural networks on parallel datasets. Hence, a seq2seq model is trained such that the encoder is input is the text of source style, and output is the corresponding text of target style. As shown in Figure 1, RNN layers act as encoder process the input sentence of source style and recover the state to serve as context for the decoder. For the next step, another layer of RNNs acting as decoder predicts the output sentence of the target style.

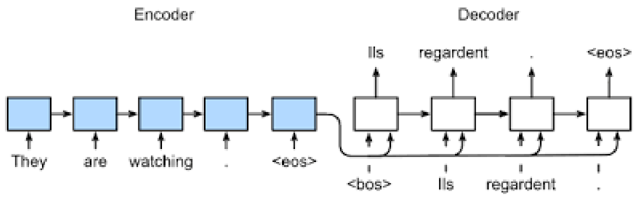

Fig. 1: Sequence-to-sequence attention-based model

In their work, Wei Xu [39] present some early work on the task of rephrasing text in a particular style. H. Jhamtani et al. [11] employ a seq2seq neural network to convert Modern English to Shakespearean text. Using dictionary to map

Table 3: Summary of some previous works in TST - their methodologies, data sources, and evaluation practices - Part 1

Ref	ML Method	Contribution	Subtask	Data	Automatic Evaluation				Human Evaluation
					Overall	Style Transfer Accuracy	Content Preservation	Fluency	
[6]	Adversarial Learning	multi-decoder and style embedding	paper-news title, sentiment transfer	He and McAuley	-	Lstm sigmoid classifier	Cosine distance	-	✓
[14]	Adversarial Learning	incorporates auxiliary and adversarial objectives	sentiment transfer	Yelp reviews, Amazon reviews	Geometric Mean of STA, WO and 1/PPL	CNN classifier	Cosine similarity, Unigram word overlap	Perplexity by trigram language model	✓
[18]	Adversarial Learning	word-level conditional architecture & two-phase training	sentiment transfer, tense transfer	Yelp, Amazon Reviews, Yelp Tense	-	CNN classifier	BLEU	Perplexity by bi-directional LSTM	✓
[43]	Adversarial Learning	adversarially regularized autoencoders (ARAE)	sentiment transfer, topic transfer	SNLI corpus, Yahoo dataset, Yelp Reviews	-	fastText classifier	BLEU	Perplexity	✓
[11]	Sequence-to-Sequence	dictionaries mapping Shakespearean to modern words	Old-Modern English	Shakespeare dataset	BLEU, PINC	-	-	-	✗

Table 4: Summary of some previous works in TST - their methodologies, data sources, and evaluation practices - Part 2

Ref	ML Method	Contribution	Subtask	Data	Automatic Evaluation				Human Evaluation
					Overall	Style Transfer Accuracy	Content Preservation	Fluency	
[39]	Sequence-to-Sequence	translation model with a language model	Paraphrasing	Shakespeare dataset	-	Cosine similarity, Language model, Logistic regression	BLEU	-	✓
[2]	Sequence-to-Sequence	Encoder–decoder recurrent neural networks	Old-Modern English	Various English Bible versions	BLEU, PINC	-	-	-	✗
[20]	Keyword Replacement	Delete, Retrieve, Generate	Normal-Romantic, Sentiment transfer	Yelp Review, Amazon Review, Captions dataset	-	LSTM-based classifier	BLEU	-	✓
[37]	Keyword Replacement	Transformer that leverages DRG framework	sentiment transfer, gender transfer, political slant	Yelp, Amazon Reviews, Captions, Gender, Political dataset	GLEU	FastText style classifier	BLEU	Perplexity by GPT-2	✓
[19]	Unsupervised Learning, Back-translation	TGLS (Text Generation by Learning from Search), framework	paraphrase generation, formality transfer	GYAFC	-	classifier based on RoBERTa features	BLEU, iBLEU	Perplexity by GPT-2	✓

modern and shakespearean English, they utilized a basic encoder-decoder architecture. Carlson et al. [2] utilized a seq2seq model with attention mechanism

Table 5: Summary of some previous works in TST - their methodologies, data sources, and evaluation practices - Part 3

Ref	ML Method	Contribution	Subtask	Data	Automatic Evaluation				Human Evaluation
					Overall	Style Transfer Accuracy	Content Preservation	Fluency	
[38]	Keyword Replacement, Reinforcement Learning	cycled reinforcement learning method	sentiment transfer	Amazon Reviews, Yelp Reviews	G-score: Geometric Mean of ACC and BLEU	CNN classifier	BLEU	-	✓
[10]	Unsupervised Learning	encoder-decoder architecture reinforced through auxiliary modules	formality transfer	emails, english prose essays	-	Encoder-based neural classifier	Cosine similarity	Perplexity by 4-gram back-off model using KenLM	✓
[30]	Back-translation	personalized SMT models in automatic translation	gender transfer	TED talks transcripts Europarliament corpus	-	SVM classifier	-	-	✓
[24]	Adversarial Learning, Keyword Replacement	Focused on 3 models: CAAE, ARAE, DAR models	sentiment transfer	Yelp Reviews	-	Earth Mover's Distance	BLEU, METEOR, Word Mover's distance	Neural logistic regression classifiers	✓
[23]	Adversarial Learning	tag-and-generate - tagger extracts content& generator converts style	politeness, Normal - Romantic, sentiment, gender, Political	Enron Email, Captions, Yelp reviews, Amazon reviews	-	LSTM classifier	BLEU, METEOR, ROUGE	-	✓

for converting a prose to Bible style text. Some later works have experimented techniques like data augmentation along with the sequence-to-sequence models [13,25]. Nikolav et al. [25] work on the task of text simplification from regular Wikipedia to simpler version. They collect two pseudo-parallel corpus pairs from sources like technical articles and news websites and propose an unsupervised technique LHA for extracting pseudo parallel sentences from the sources. However, seq2seq approach requires parallel corpus, and is hence challenging due to its scarcity.

3.2 Non-Parallel

Keyword Replacement Certain keywords in texts often are indicative of sentiments and tone underneath. For instance, words like “wow”, “excellent” have a positive meaning and “awful” and “worst” have a negative connotation. Hence, some works [20,37] introduce the use of machine learning models to replace these words using Natural language generative models.

An initial work of the keyword replacement approach is the Delete-Retrieve-Generate framework [20]. Firstly, the model identifies all the style-connotated words like “bad”, “rude” etc. from the source sentences (eg. negative sentiment). Then it eliminates these words and only the content indicative words are left, such as “hotel” or “shirt”. Next, reference texts similar to the content-related words are retrieved from the target sentiment corpus (here, eg. positive). Then, the model similarly gets all the style-attributed words and combines them with the content words extracted previously. This combining is done with seq2seq models

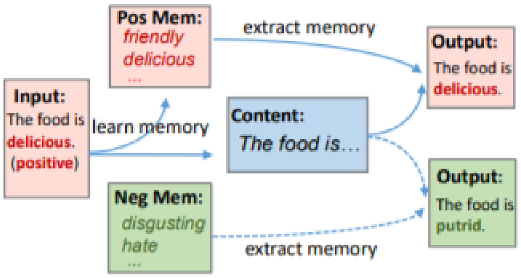

Fig. 2: Overview of the model proposed by [41] using self-attention and keyword replacement

The authors of [41] uses attention mechanisms with Deep Learning architectures for Natural Language Generation tasks as shown in Figure 2. Recent works explored hybrid architectures using keyword replacement methods with cycled reinforcement losses to iteratively transfer the style of the text while maintaining the content. Since the words replaced can be visually inspected by a human, it imparts explainability to the models. The parts of text modified can be examined to understand the performance of such techniques.

Adversarial Learning Another effective method for TST is adversarial learning which separates the text's style and content data for transferring the text to target style. An early work proposed by Fu et al. [6] uses an adversarial framework of two models shown in Figure 3.

The first model consists of a single encoder and multiple decoders. The style of text input is learned by the encoder representations which in turn trains multiple decoders to decipher the representation and output the text in target style. In the second model, encoder behaves similar to the first model, and the decoder outputs the target style by concatenating the embedding of the encoder to parameter representation.

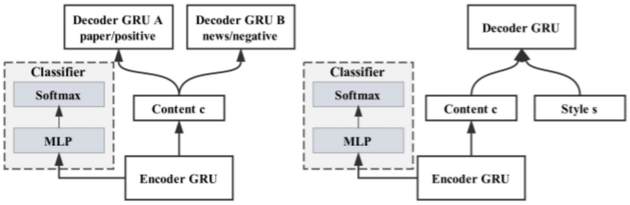

Fig. 3: Two adversarial learning based models proposed by [6]

Similar to the previous methods, a variety of hybrid architectures have been proposed for adversarial-based frameworks. Recently, works on the usage of autoencoders for TST models implementing adversarial learning to refine the performance of style transfer tasks is studied. For example, some works [3,14,18,43] implemented cycled-consistency loss where the generated text is again fed to the model and the output sentence is compared to the original input text. This way, the loss is cycled or transferred to the model to generate more accurate transferred sentences.

Reinforcement Learning Reinforcement learning works on the idea that reward functions guide the decision of deep learning models instead of loss functions. The parameters of a reinforcement-based model change in a way that the estimated reward of the output style-transferred text is maximized. An attention-based model proposed by [8] uses encoder-decoder architecture to perform Text Style Transfer. Its proposed generator and evaluator based method is shown in Figure 4. Here, three rewards to guide the output text of the desired style are introduced. Style classifiers, semantic and language models are employed to impart style, semantic and language feedback to the model respectively.

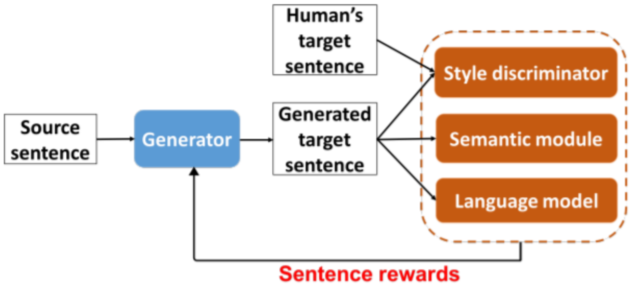

Fig. 4: Overview of the model proposed by [8] where rewards are returned to the generator

In [22], the authors propose a dual reinforcement learning framework where one seq2seq model learns source style to target style embeddings. The second seq2seq learns the target to source style parameters. This dual task is designed to provide style and reconstruction rewards and their average is used to provide feedback to the model. Thus, reinforcement learning without any parallel corpus is used for style transfer tasks.

Backtranslation Back-translation means translating a text of target style to the source style and mixing both original source and back-translated text to train the model. The work by authors of [30] researched the factors like gender that are often obscured in tasks like Machine Language Translation. The back-translation method used for Text Style Transfer by [29] is shown in Figure 5. Their approach was to first rephrase the sentences and retain only the content words thereby eliminating the style information. Using an NMT model, they translated English text into another language and then back into the English language. It is proposed with the understanding that the stylistic properties are lost during the back and forth translation while preserving the content. In [19], the authors propose a for text generation framework with two stages. First stage of Learning simulated annealing search to generate pseudo-input sentences and a generator undergoes training via supervised learning. In the next stage, iterations are performed of beam search process for model improvement.

The authors in [42] employ a two-fold framework where first a pseudo-parallel dataset is formed using word embeddings and latent representations similarities. Thus this unsupervised task of style transfer is now employed to back translation-based deep learning models. A style classifier is utilized to enhance the performance of the model. Like the methods using adversarial models, back-translation models also proves less efficient in style-content agreement.

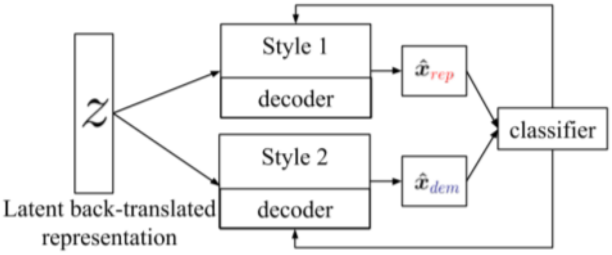

Fig. 5: Back-translation method and style classifier used by [29]

3.3 Unsupervised methods

The previous models using parallel or non-parallel data are proposed for supervised settings for style transfer tasks. Recent works have also explored Machine Learning techniques in purely unsupervised settings where no labelled style-text corpus is provided to the model. There are relatively fewer works proposed to perform TST in such a way [10, 31, 35].

Initial study conducted by Radford et al. [31] made use of the characteristics RNNs and LSTMs wherein training of such models is done on the UTF8 byte of the input text. This preprocessing allowed the researchers to identify and modify the neuron-level embeddings of the Deep Learning models. Text formalization performed by Jain et al. [10] uses an unsupervised method with unlabelled corpus. They make use of external language processing tools called scorers to provide style information. The information learned by the encoder-decoder is backpropagated to the model for computing the loss. The output scores determine the formality level of the text and the resulting output text. Shen et al. in their work [35] proposed adversarial autoencoders called AAE with denoising models for mapping of similar latent representations. These models perform sentiment transfer by computing a vector using these representations. There is still much research potential in natural language generation based unsupervised methods that can be extrapolated to other TST subtasks.

4 Evaluation Techniques

Measuring the true efficacy of the TST models is one of the most challenging tasks in the domain of text style transfer. At present, there are no standard automatic evaluation metrics that are being followed conventionally and still no metric exists that can outperform human evaluation. Since human evaluation can be onerous and expensive, there is a compelling need for appropriate automatic evaluation practices. Various evaluation metrics for TST have been proposed previously which mainly focus on three aspects to measure the effectiveness of a TST model:

1. Style transfer strength: the ability to convert a source style to the desired target style.
2. Content preservation: the extent to which the original content is preserved.
3. Fluency (naturalness): the ability to generate fluent sentences.

A TST model should perform deftly in all these three criteria of evaluation. Underperformance in any of the three aspects would rather deem the model ineffective. For instance, if the algorithm successfully transfers a positive sentiment sentence, "The teachers in this university are excellent" to a negative sentiment sentence, "The waiters in the restaurant were very rude", the algorithm, here, doesn't preserve the original meaning and hence is considered to be inadequate.

4.1 Automatic Evaluation

Style transfer strength This criterion of evaluation deals with how well the style of a given text is transferred to the target style. In most past works, transfer strength is tested by making use of pre-trained classifiers. A lot of previous papers [13, 22, 34] have used TextCNN, a sentence-level text classifier trained over pre-trained word vectors, proposed by [16]. An LSTM classifier, first used for this task in [6], is employed to measure the transfer strength in [8, 9]. Some works [5,21] have made use of fastText [15], which shows identical performance as the deep learning methods, while being faster in speed, for the style classification task. Another alternate metric proposed in [24] for measuring transfer strength is to compute the Earth Mover's Distance [32] between the source text and the style transferred output. This metric can be used to handle even non-binary classification and exhibited a higher correlation with human evaluation than fastText and TextCNN classifiers.

Content Preservation The metrics under this criterion of evaluation are concerned with the measurement of the extent of original content that is preserved after style transfer. The most widely used metric to evaluate content similarity is the BLEU score [26], originally proposed for the evaluation of machine translation tasks. [33, 38] have computed the BLEU score between the source text and the transferred text to measure content preservation in transferred sentences. In [6], the authors have calculated the cosine similarity between the source and the transfer sentence embeddings by leveraging the pre-trained word embeddings by GloVe [27]. Another popular metric for this specific task is the METEOR score [1] used in [23, 36]. [24] proposes to practice style masking or style removal i.e., to remove or mask the style attribute words from the source and transferred sentences, before using the content preservation metrics. The authors, in [40] have done an extensive comparison of fourteen content similarity metrics over two style transfer datasets and suggested that Word Mover's Distance (WMD) [17] and L2 distance based on ELMo [28] are the best performing metrics for measuring content preservation in style transfer tasks.

Fluency (naturalness) For any natural language generation model to be efficient, it must possess the ability to produce fluent human-like text. Most commonly, like in [10,14,37], a language model is trained and employed to compute perplexity score (PPL), where lower scores indicate higher fluency. Although researchers have relied on using the perplexity scores for evaluating fluency of the TST models in the past, [24] showed that perplexity scores exhibited a very low correlation with human-evaluated scores and instead adversarial neural classifiers should be employed for the task of evaluating naturalness.

4.2 Human Evaluation

A lot of previous works [6,8,34,42] have incorporated the use of human evaluation along with the automatic evaluation metrics discussed above. Generally, a common procedure is followed where the evaluators are asked to rate randomly selected style transferred outputs. The scale of rating varies across different works but essentially the approach is similar where higher scores indicate better performance. The human raters can provide overall scores or give scores separately based on the three criteria i.e., style transfer strength, content preservation, and fluency. Human evaluation, apart from being expensive and cumbersome, for the style transfer task can often be subjective in nature depending on how the rater interprets the styles under consideration. Therefore, it is also often not comparable across different methods due to this subjectiveness. Notwithstanding the cons, human evaluation is extremely important to be carried out in the contemporary scenario of the TST domain along with the automatic evaluation, to study the correlation between various automatic metrics and human judgment. It will play an important role in the evolution of automatic metrics for TST tasks and set benchmarks for future comparison.

5 Discussion and Conclusion

Research in the field of TST is challenged by a few obstacles such as scarcity of publicly available, benchmark parallel corpuses and the absence of standardized automatic evaluation metrics. Currently, there are no automatic evaluation methods developed that can surpass human evaluation, which proves to be expensive and cumbersome. Hence, a prospective research area in TST is using unsupervised machine learning methods to surpass the issue of the absence of large parallel data. Also, while a significant portion of the work is applicable to English corpora, TST's potential to extrapolate it to other languages should not be neglected. In this paper, we have provided a comprehensive review of the existing literature and numerous machine learning methods employed for the task. The review also discussed various benchmark datasets and evaluation practices. This paper can serve as a reference for NLP researchers and is aimed to provide an all-inclusive understanding of TST to facilitate and promote further research in this field.

References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005)
2. Carlson, K., Riddell, A., Rockmore, D.: Evaluating prose style transfer with the bible. Royal Society open science **5**(10), 171920 (2018)
3. Chen, L., Dai, S., Tao, C., Shen, D., Gan, Z., Zhang, H., Zhang, Y., Carin, L.: Adversarial text generation via feature-mover's distance. arXiv preprint arXiv:1809.06297 (2018)
4. Cífka, O., Şimşekli, U., Richard, G.: Groove2groove: One-shot music style transfer with supervision from synthetic data. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 2638–2650 (2020)
5. Dai, N., Liang, J., Qiu, X., Huang, X.: Style transformer: Unpaired text style transfer without disentangled latent representation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (2019)
6. Fu, Z., Tan, X., Peng, N., Zhao, D., Yan, R.: Style transfer in text: Exploration and evaluation. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018)
7. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
8. Gong, H., Bhat, S., Wu, L., Xiong, J., mei Hwu, W.: Reinforcement learning based text style transfer without parallel training corpus. In: Proceedings of the 2019 Conference of the North. Association for Computational Linguistics (2019)
9. Gröndahl, T., Asokan, N.: Effective writing style imitation via combinatorial paraphrasing. CoRR **abs/1905.13464** (2019)
10. Jain, P., Mishra, A., Azad, A.P., Sankaranarayanan, K.: Unsupervised controllable text formalization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6554–6561 (2019)
11. Jhamtani, H., Gangal, V., Hovy, E., Nyberg, E.: Shakespearizing modern language using copy-enriched sequence-to-sequence models. arXiv preprint arXiv:1707.01161 (2017)
12. Jin, D., Jin, Z., Hu, Z., Vechtomova, O., Mihalcea, R.: Deep learning for text style transfer: A survey. CoRR **abs/2011.00416** (2020)
13. Jin, Z., Jin, D., Mueller, J., Matthews, N., Santus, E.: IMaT: Unsupervised text attribute transfer via iterative matching and translation. 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). Association for Computational Linguistics (2019)
14. John, V., Mou, L., Bahuleyan, H., Vechtomova, O.: Disentangled representation learning for non-parallel text style transfer. arXiv preprint arXiv:1808.04339 (2018)
15. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics, Valencia, Spain (Apr 2017)

16. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics (2014)
17. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 957–966. PMLR, Lille, France (07–09 Jul 2015)
18. Lai, C.T., Hong, Y.T., Chen, H.Y., Lu, C.J., Lin, S.D.: Multiple text style transfer by using word-level conditional generative adversarial network with two-phase training. 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). pp. 3579–3584 (2019)
19. Li, J., Li, Z., Mou, L., Jiang, X., Lyu, M.R., King, I.: Unsupervised text generation by learning from search. arXiv preprint arXiv:2007.08557 (2020)
20. Li, J., Jia, R., He, H., Liang, P.: Delete, retrieve, generate: A simple approach to sentiment and style transfer. arXiv preprint arXiv:1804.06437 (2018)
21. Liu, Y., Neubig, G., Wieting, J.: On learning text style transfer with direct rewards. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (2021)
22. Luo, F., Li, P., Zhou, J., Yang, P., Chang, B., Sun, X., Sui, Z.: A dual reinforcement learning framework for unsupervised text style transfer. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization (Aug 2019)
23. Madaan, A., Setlur, A., Parekh, T., Poczos, B., Neubig, G., Yang, Y., Salakhutdinov, R., Black, A.W., Prabhumoye, S.: Politeness transfer: A tag and generate approach. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (2020)
24. Mir, R., Felbo, B., Obradovich, N., Rahwan, I.: Evaluating style transfer for text. In: Proceedings of the 2019 Conference of the North. Association for Computational Linguistics (2019)
25. Nikolov, N.I., Hahnloser, R.H.: Large-scale hierarchical alignment for data-driven text rewriting. arXiv preprint arXiv:1810.08237 (2018)
26. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. Association for Computational Linguistics (2001)
27. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics (2014)
28. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics (2018)
29. Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., Black, A.W.: Style transfer through back-translation. arXiv preprint arXiv:1804.09000 (2018)
30. Rabinovich, E., Mirkin, S., Patel, R.N., Specia, L., Wintner, S.: Personalized machine translation: Preserving original author traits. arXiv preprint arXiv:1610.05461 (2016)
31. Radford, A., Jozefowicz, R., Sutskever, I.: Learning to generate reviews and discovering sentiment. arXiv preprint arXiv:1704.01444 (2017)

32. Rubner, Y., Tomasi, C., Guibas, L.J.: A metric for distributions with applications to image databases. In: Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271). pp. 59–66. IEEE (1998)
33. Shang, M., Li, P., Fu, Z., Bing, L., Zhao, D., Shi, S., Yan, R.: Semi-supervised text style transfer: Cross projection in latent space. 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). Association for Computational Linguistics (2019)
34. Shen, T., Lei, T., Barzilay, R., Jaakkola, T.: Style transfer from non-parallel text by cross-alignment. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6833–6844. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
35. Shen, T., Mueller, J., Barzilay, R., Jaakkola, T.: Educating text autoencoders: Latent representation guidance via denoising. In: International Conference on Machine Learning. pp. 8719–8729. PMLR (2020)
36. Shetty, R., Schiele, B., Fritz, M.: A4nt: Author attribute anonymity by adversarial training of neural machine translation. In: 27th USENIX Security Symposium (USENIX Security 18). pp. 1633–1650. USENIX Association, Baltimore, MD (Aug 2018)
37. Sudhakar, A., Upadhyay, B., Maheswaran, A.: Transforming delete, retrieve, generate approach for controlled text style transfer. arXiv preprint arXiv:1908.09368 (2019)
38. Xu, J., Sun, X., Zeng, Q., Zhang, X., Ren, X., Wang, H., Li, W.: Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics (2018)
39. Xu, W., Ritter, A., Dolan, W.B., Grishman, R., Cherry, C.: Paraphrasing for style. In: Proceedings of COLING 2012. pp. 2899–2914 (2012)
40. Yamshchikov, I.P., Shibaev, V., Khlebnikov, N., Tikhonov, A.: Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. ArXiv **abs/2004.05001** (2021)
41. Zhang, Y., Xu, J., Yang, P., Sun, X.: Learning sentiment memories for sentiment modification without parallel data. arXiv preprint arXiv:1808.07311 (2018)
42. Zhang, Z., Ren, S., Liu, S., Wang, J., Chen, P., Li, M., Zhou, M., Chen, E.: Style transfer as unsupervised machine translation. ArXiv **abs/1808.07894** (2018)
43. Zhao, J., Kim, Y., Zhang, K., Rush, A., LeCun, Y.: Adversarially regularized autoencoders. In: International conference on machine learning. pp. 5902–5911. PMLR (2018)
44. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2242–2251 (2017)