Fantastic Gradients and Where to Find Them: Improving Multi-attribute Text Style Transfer by Quadratic Program

Qian Qu, Jian Wang, Kexin Yang, Hang Zhang, and Jiancheng Lv(✉)

No.24, South Section of 1st Ring Road, Chengdu, China
kylinaive77@gmail.com

Abstract. Unsupervised text style transfer (TST) is an important task with extensive implications in natural language generation (NLG). A prevalent approach involves editing the latent representations of text, guided by gradients from an attribute classifier. However, in multi-attribute TST, the simultaneous satisfaction of all required attributes remains challenging. In this paper, we unveil that the gradient direction during editing might conflict with certain attribute representations through empirical analysis. To tackle this problem, we introduce a mathematical programming method to impose constraints on the editing direction of multiple attributes, effectively mitigating potential attribute conflicts during the inference stage. Our proposed method considers the potential conflict between different attributes for the first time. Experimental results from the YELP benchmark showcase that our method can effectively improve the multi-attribute-transfer accuracy and quality without compromising single attribute performance. Moreover, our method can be readily integrated with pre-trained auto-encoders, providing an effective and scalable solution for multi-attribute scenarios.

Keywords: Text Style Transfer · Multiple-Attribute Text Generation · Auto-Encoder · Quadratic Program

1 Introduction

Text style transfer (TST) seeks to rephrase the source text in a language style specific to certain attributes while preserving content that is independent of these style attributes. As depicted in Table 1, the *style* in TST could be any attribute requiring modification, such as sentiment, topic, gender, writing style, or a combination thereof [7,23]. Although substantial strides have been made in multi-attribute text generation [6,9], much of the existing research concentrates on achieving a balance between content preservation and style manipulation constraints, often ignoring the challenges inherent in satisfying multiple attributes simultaneously. In light of this, we turn to an approach that modifies the latent representations of texts without disentanglement, guided by classifier gradients [12,26]. This method offers both training efficiency and flexibility in text transformation [12,19], making it an attractive choice for our exploration of multi-attribute TST. Our empirical study reveals that, during the editing

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14304, pp. 3–14, 2023.
https://doi.org/10.1007/978-3-031-44699-3_1

Table 1. Several common TST tasks with example sentences.

Task	Attribute	Example
Sentiment	Positive	This movie is really meaningful and I learned a lot from it
	Negative	This movie is really meaningless and I don't get anything from it
Gender	Male	My wife likes the fried chicken here
	Female	My husband likes the fried chicken here
Formality	Formal	I can't eat another bite. I proceed to chew and explode
	Informal	Ooh, I can't eat another bite (munch munch, explode)
Author styles	Shakespearean	I saw thee late at the Count Orsino's
	Modern	I saw you at Count Orsino's recently

process, the gradient direction may contradict the representations of certain attributes. This could potentially lead to the generation of attribute-incomplete text.

To address this, we introduce a novel method grounded in mathematical programming for constrained multi-attribute text style transfer, based on a latent representation editing model. Specifically, our approach starts with an autoencoder for sentence self-representation, which could also be a pre-trained model. It then establishes constraints on the editing direction of multiple attributes by guiding modifications to the latent representation via a classifier. Our method, for the first time, considers potential conflicts between different attributes, ensuring that generated text satisfies the required attributes to the greatest extent possible during the editing process. Experimental results on the widely-recognized YELP benchmark [10] demonstrate the efficacy of our method in enhancing text-transfer accuracy across multiple attributes. Our main contributions include:

- We identify the limitations and possible reasons for the suboptimal performance of existing latent representation editing methods in multi-attribute scenarios.
- We propose a novel, flexible multi-attribute TST model based on the latent variable editing method, which first takes into account the conflict and satisfaction between multiple attributes of generated text.
- By utilizing Quadratic Programming (QP) with inequality constraints, we can modify the gradient while preserving its key attributes, resulting in improved overall accuracy for multiple attributes.

2 Related Work

For the disentanglement-based methods [5,15,22], its main idea is to separate the style and content of the original text, then generate the new text with the target style and content representation. The key lies include adversarial learning methods [5,17,22], attention mechanism [27,29], or other method such as Levenshtein Editing [11,20]. Instead of performing a disentanglement of content and style, entanglement-based methods rewrite the entangled representations directly in a specific manner, such as reinforcement learning [14], back-translation technique [1,2,21], and latent vector editing method [12,19,26].

Multi-attribute style transfer is an extension of single attribute but is more difficult. There are also studies that focus specifically on multiple attributes. [9] proposed an adversarial training model using word-level conditional architecture and a two-stage training program for multi-attribute generation. [10] implemented multi-attribute style transfer by adjusting the average embedding of each target attribute and using a combination of DAE and back translation techniques. [6] use multiple style-aware language models as discriminators in combination with transformer-based encoder-decoders to enhance their rewriting capabilities.

3 Methodology

We start with a dataset $\mathcal{D} = (x^i, s^i)_{i=1}^n$, wherein each unit (x, s) denotes a sentence x together with its corresponding attribute vector s. This attribute vector might cover multiple attributes, such as *sentiment* and *gender*, which can be represented as $s = \{s_{sent}, s_{gend}\}$. The main aim is to transform a given sentence x^i, accompanied by its associated attribute s^i, into a new sentence $\hat{x}$ that aligns with a target attribute $\hat{s}$.

3.1 Preliminary

In pursuing this aim, we turn our attention to the latent representation revision method. The fundamental concept here involves fine-tuning the entangled latent representation of the input sentence to align with the desired attribute. As depicted in Fig. 1(A), the model typically integrates three core components: an encoder G_{enc}, a decoder G_{dec}, and an attribute classifier C. It's noteworthy that some studies [12] prefer to utilize multiple classifiers. The process begins with the encoder, which translates a given input sentence x into a latent representation z. This representation integrates both the attribute and the content in a tangled manner. Following this, z is modified to match the target attribute, under the guidance of the classifier. Ultimately, the decoder converts z back into a sentence $\hat{x}$, embodying the desired attribute.

The modification of z using the gradient provided by C [12,26] is merely one among a host of potential strategies. Another option suggested by [19] involves steering z directly across the surface of the decision boundary. In this work, we have chosen to adhere to the gradient modification approach to execute multi-attribute style transfer.

3.2 Problem Analysis

Our approach modifies the latent representation z of the input sentence to incorporate the target attribute by using the gradient from the attribute classifier C.

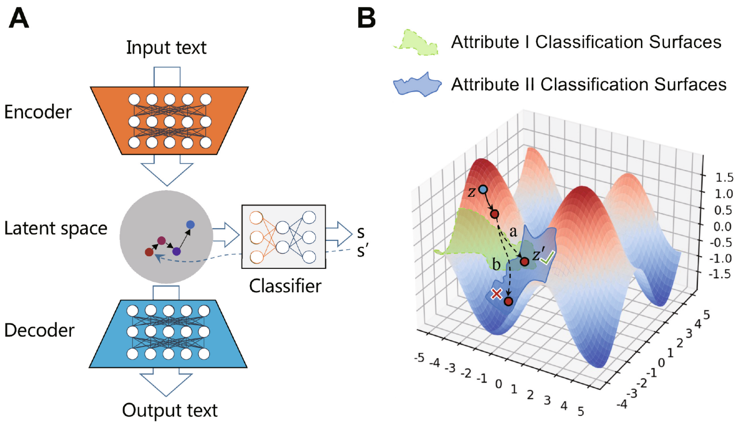

Fig. 1. Overview of Latent representation revision method. (A) Model architecture. (B) Latent representation editing for multi-attribute TST. Given a sentence, the objective of the model is to rephrase it such that it incorporates both Attribute I and Attribute II.

This gradient guides the search for a new representation z' that satisfies the desired attribute s' while staying close to the original sentence:

$$z' = z - \omega_i \nabla_z \mathcal{L}_C(C_{\theta_C}(z), s'), \tag{1}$$

where θ_C and ω_i are the parameters of the classifier C and the adjustment factor, respectively. This process is repeated until the classifier C confirms that z' matches the target style.

However, this method may compromise the generation quality when transferring multiple attributes. The attribute classifier provides a joint gradient on all labels s' to update the latent representation, i.e.,

$$\sum_{s'} \nabla_z \mathcal{L}_C(C_{\theta_C}(z), s'). \tag{2}$$

Given that the decision surfaces for each attribute in the classification may not completely overlap under multi-attribute style transfer, conflicts might arise between the gradient orientations of different attributes when modifying the latent representation along a specific gradient path. For example, certain steps might draw z' nearer to attribute s_1 while distancing it from attribute s_2. As depicted in Fig. 1 (B), adjusting the latent representation along the gradient direction of path b, leading to a lower final loss, only meets the target attribute s_2 criteria. Conversely, path a, despite a higher loss value, accommodates two attributes and thus produces a more desirable outcome. Our experiments confirmed this phenomenon by identifying instances of conflict between the editing gradient orientation and the single-attribute gradient orientation during the transfer process, as discussed in Sect. 4.4.

3.3 Model Architecture

Our framework is designed to be compatible with any auto-encoder (AE) and multi-attribute classifier, making it agnostic to the specific neural network architecture employed. In this work, we employed a transformer-based auto-encoder [24] in conjunction with an MLP-based classifier.

Transformer-Based Auto-Encoder G. Given an input sentence $x = \{x_1, x_2, ..., x_m\}$, the encoder $G_{enc}(\theta_{enc}; x)$ transforms it into a continuous latent representation: $z \sim G_{enc}(\theta_{enc}; x) = q(z|x)$, while the decoder $G_{dec}(\theta_{dec}; z)$ maps the latent representation z back to the sentence, reconstructing it: $x \sim G_{dec}(\theta_{dec}; z) = p(x|z)$. During training, the objective of G is to minimize the reconstruction error. The reconstruction loss is defined as:

$$\mathcal{L}_G(\theta_{enc}, \theta_{dec}; x) = -\frac{1}{|s|} \sum_{i=1}^{|s|} q(z|x) \log p(x|z), \tag{3}$$

where $|s|$ denotes the number of attributes.

Multiple-Attribute Classifier C. Our classifier is implemented as an MLP consisting of two linear layers and a sigmoid activation function. Specifically, it is defined as $C(z) = MLP(z) = p(s|z)$. The attribute classification loss is:

$$\mathcal{L}_C(\theta_C; z, s) = -\frac{1}{|s|} \sum_{i=1}^{|s|} [s_i \log(p(s_i|z)) + (1 - s_i) \log(1 - p(s_i|z))],$$

where $|s|$ is the number of attributes and s_i is the ground truth label for the i-th attribute.

3.4 Multiple-Attributes Gradient Iterative Modification

Conflict Resolution in Modification. To align z' with all target attributes, we adopt a gradient direction detection and conflict resolution strategy inspired by GME [13] when modifying z. We consider not only the gradient conflict of classifiers over z, but also the conflict that arises during the intermediate gradient propagation in C. Therefore, during the inference stage, we detect the gradient direction by computing the inner product of the gradient vectors in each linear layer as the gradient backpropagates in C. This enables us to identify potential directional conflicts between the gradient of any single attribute g_i and the overall gradient g[1]. The constraint is satisfied when the gradient g agrees with all desired attribute directions, expressed as follows:

$$\langle g_i, g \rangle := \langle \nabla_z \mathcal{L}_C(C_{\theta_C}(z), {s_i}'), \nabla_z \mathcal{L}_C(C_{\theta_C}(z), s') \rangle \geq 0. \tag{4}$$

In cases where a conflict arises, modifying the gradient in question could potentially cause z to deviate from the target property associated with the conflicting

[1] For simplicity, we denote all the gradients to be detected in C by $\nabla_z \mathcal{L}_C(C_{\theta_C}(z), s')$.

direction. To tackle this, we project the gradient g onto the nearest gradient $\tilde{g}$ that fulfills all attributes:

$$\text{minimize}_{\tilde{g}} \frac{1}{2} \|g - \tilde{g}\|_2^2 \quad s.t. \langle g_i, \tilde{g} \rangle \geq 0. \tag{5}$$

To address Eq. 5, which presents a Quadratic Program (QP) with inequality constraints [4, 13], it is useful to return to the primal form:

$$\text{minimize}_r \frac{1}{2} r^\top H r + p^\top r \quad s.t.\ Ar \geq b, \tag{6}$$

where $H \in \mathbb{R}^{p \times p}$ is a symmetric, positive semi-definite matrix, $p \in \mathbb{R}^p$, $A \in \mathbb{R}^{|s| \times p}$, $b \in \mathbb{R}^{|s|}$. The dual problem of inequality [3] (Eq. 6) is:

$$\text{minimize}_{u,v} \frac{1}{2} u^\top H u - b^\top v \quad s.t.\ A^\top v - Hu = p, v \geq 0. \tag{7}$$

Drawing from Dorn's duality theorem [3], if a solution u^* and v^* is obtained from Eq. 7, then there exists a solution r^* to Eq. 6, which satisfies $Hr^* = Hr^*$.

On this basis, the original QP (Eq. 5) can be expressed as:

$$\text{minimize}_w \frac{1}{2} r^\top r - g^\top r + \frac{1}{2} g^\top g \quad s.t.\ Gr \geq 0, \tag{8}$$

where $G = (g, g_1, ..., g_{|s|})$. The dual problem of (Eq. 8) is:

$$\text{minimize}_v \frac{1}{2} v^\top G G^\top v + g^\top G^\top v \quad s.t.\ v \geq 0, \tag{9}$$

where $u = G^\top v + g$ and $g^\top g$ is the constant term. This is a QP on $|s|$ attributes. Then once we solve the problem (9) for v^*, we can get the adjusted new gradient $\tilde{g} = G^\top v^* + g$.

Following the resolution of conflicts, the adjusted gradient $\tilde{g}$ is propagated to the subsequent layer of the network. This gradient then steers the modifications applied to the latent representation z'. The detail of this process is outlined in Algorithm 1.

To preserve the attribute-independent content and linguistic integrity of the latent representation, we confine gradient modifications to large-step gradients. This approach mitigates potential negative impacts on the linguistic fluency and coherence of the decoded text that may result from insignificant style category changes induced by small gradients. It is crucial to note that this procedure is strictly implemented during the inference stage and does not come into play during training.

4 Experiments

4.1 Dataset

We evaluated our approach on the Yelp Review Dataset (YELP) [10], which contains complete reviews along with review sentiment, gender and restaurant category information. We conducted multi-attribute style transfer experiments on the three attributes of sentiment, gender, and restaurant categories. The restaurants here we choose three types: Asian, American, and Mexican.

Algorithm 1: Multiple-Attributes Fast Gradient Iterative Modification Algorithm.

```
Input: Auto-encoder latent representation $z$; Target attribute $s' = ({s_i}', ..., {s_k}')$;
       Well-trained attribute classifier $C_\theta$; Weights $\omega = \{\omega_i\}$
Output: A modified latent representation $z'$
$g \leftarrow \nabla_z \mathcal{L}_C(C_{\theta_C}(z), s')$;
$g_i \leftarrow \nabla_z \mathcal{L}_C(C_{\theta_C}(z), {s_i}')$ for all $i = 1, ..., k$;
if $|s' - C_{\theta_C}(z')| > t$ then
    for each linear_layers $\in C_\theta$ do
        if $\langle g, g_i \rangle \geq 0$ for all $i = 1, ..., k$; then
            $\tilde{g} \leftarrow g$;
        else
            $\tilde{g} \leftarrow \mathrm{PROJECT}(g, g_1, ..., g_k)$, see (9) ;
        end
    end
    $z' = z - \omega_i \tilde{g}$;
else
    reture $z'$;
end
```

4.2 Baselines

We compare our model with the most relevant and state-of-the-art models as follows: 1) **StyIns** [28]. encodes sentences with a certain style to vectors as the style instances and uses the generative flow technique to construct style latent representation based on it, then decodes the input sentence along with the style representation to generate desired text. 2) **ControllableAttrTransfer (CAT)** [26]. edits the sentence latent representations guided by an attribute classifier until it is evaluated as the target style. 3) **MultipleAttrTransfer (MAT)** [10]. is based on the Denoising auto-encoding (DAE) [25] model and back translation strategy. 4) **MUCOCO** [8]. conducts controlled inference from the pre-trained model and formulates the decoding process as a multi-optimization problem. It then generates the target sentences using Lagrange multipliers and gradient-descent based techniques.

4.3 Evaluations Metrics

Automatic Evaluation. Following previous works [7,17,22,23], we use the automatic metrics as follows: 1) **Style transfer Accuracy.** We train an external classifier to measure the accuracy of the transferred sentences related to the required attribute. Here, we have trained a GPT-based [18] classifier on each attribute (sentiment, gender, category) using the training data. 2) **Content preservation.** We calculate the BLEU [16] score between the transferred sentence and the original input sentence (self-BLEU), with higher scores meaning more content retention. 3) **Fluency.** We calculate the perplexity of transferred

Table 2. Automatic and human evaluation results for multi-attribute transfer tasks on YELP. Notice that since there is a multi-attribute task, the accuracy here does not simply refer to the correct rate of one attribute, but to the overall attribute, that is, the generated sentence that satisfies all the target attributes.

	Automatic			Human			
	Acc	BLEU	PPL	Sty	Con	Flu	Avg
StyIns [28]	33.7	23.75	75.25	2.88	3.84	3.82	3.66
CAT [26]	37.5	20.53	52.77	3.21	3.92	4.15	3.76
MAT [10]	34.1	25.34	55.34	3.06	4.12	4.22	3.81
MUCOCO [8]	28.9	28.45	51.68	2.78	3.98	4.01	3.51
Our model	39.8	26.23	49.89	3.35	4.05	4.27	3.89

sentences by a Transformer-Based language model, which is trained with the Training data (the lower the better).

Human Evaluation. We further conduct the human evaluation for transfer results. Following some previous works [6,12,20], evaluators are asked to rate sentences according to the three criteria described above with each aspect rated on a 5-point Likert scale. Especially, for Style transfer strength, a score of five is given when the sentence satisfies all the attributes and makes sense, with an equal proportional reduction for missing attributes or not reasonable enough.

4.4 Main Results

Gradient Conflict Detection. Here, we verify our claim that editing z with a gradient direction may conflict with the gradient direction of some attributes. For the well-trained model, we randomly select 200 data from the test set to perform multi-attribute TST and detect the situation between the edit gradient direction and the single-attribute direction during the transfer process. The experimental results show that the gradient direction conflict occurred in 100% of the 200 texts. Notably, not every conflict will lead to an attribute-incomplete generation text, but increasing the corresponding possibility (we verify such a situation in Sect. 4.5).

Compare with Baselines. In Table 2, we present the automatic and human evaluation results of both our model and the baseline model. The results indicate that our model outperforms the baseline model in terms of style transfer accuracy, achieving the highest score with a significant improvement compared with baseline models (t-test, $p < 0.05$).

Notably, the automatic accuracy of all models is relatively low as it requires all target attributes to be satisfied in a single transferred sentence. In reality, the accuracy of satisfying just one of the attributes would be much higher, as will be demonstrated in detail in the ablation study below. Furthermore, the BLEU and PPL scores are within a normal range. Our model achieves the best results

Table 3. Ablation study results on YELP dataset, comparing the performance of our model without (w_o) and with ($w\_$) the implementation of the gradient conflict adjustment strategy. The accuracy for each target attribute, as well as the overall accuracy, is provided for both scenarios. The 'overall accuracy' refers to the percentage of the generated text that conforms to all of the requisite target attributes simultaneously. Best viewed in **bold**.

	Sentiment	Gender	Category	Overall
	$w\_o/w\_$	$w\_o/w\_$	$w\_o/w\_$	$w\_o/w\_$
Accuracy	0.98/0.98	0.58/0.58	0.69/**0.72**	0.38/**0.40**
F1-score	0.98/0.98	0.68/0.68	0.82/**0.84**	-/-
PPL	-/-	-/-	-/-	50.54/**49.90**

on PPL and the second-best score on BLEU, which could be due to the fact that slightly more of the original text was modified to satisfy additional properties.

In human evaluation, we selected ten sentences from each model for each multi-attribute task and asked five evaluators to rate each comparison sample. In total, we evaluated 70 sentences for each model, taking into account the transfer of each attribute to the other (e.g., positive → negative with any other gender and category transfer, positive → negative with any gender and category transfer, ...). Our model proved to be the unequivocal leader, surpassing all others in both accuracy scores and average scores. Furthermore, we observed that the accuracy of human ratings significantly exceeded that of automatic evaluations. This can be attributed to the fact that human ratings consider sentences that satisfy one or two target attributes, while automatic evaluations only account for the generation that fulfills all attributes when calculating accuracy. This fills a missing in the perspective of automated evaluation, as transfer results that satisfy two attributes are considered superior to results that satisfy only one or fewer attributes.

Moreover, to observe the characteristics of each model under the multi-attribute task more intuitively, we randomly sampled a set of output sentences and showed them in Table 4.

4.5 Ablation Study

To further validate the reliability of our approach, we conduct an extensive analysis of the key components of our model in this section. In Table 3 we show the comparison in performance of our model without and with implementing the gradient conflict adjustment strategy on the YELP dataset. It can be seen that after applying the gradient programming strategy, the overall accuracy improves by 2.3% points with a statistically significant (t-test, $p < 0.05$). And for every single attribute, the correct rate is equal to or greater than before. In particular, the category accuracy experiences a significant improvement (t-test, $p < 0.05$) as it is a multi-attribute scenario with three sub-attributes, therefore, our method can also be effective here This finding underscores the effectiveness

Table 4. Case study of generated text by all models. The blue word indicates relevant text in the output that contain the target sentiment attribute, the red words indicate the target gender attribute and green words indicate category attribute.

Negative to Positive, Female to Male, Mexican to Others	
Original text	awful! all i can say. horrible service for 1 and the food is nothing special and it's over priced !
Our model	wow! i can say that a good place. great service and the food. on top, my wife and i both chose the cheeseburger with turkey burgers and it is well made no longer like home fries. wouldn't be exceptional!
StyIns	wow! i can say that great service and all the food. we are allergic to the menu and we get the greenspsmyhummus without chips, while we ate there to enjoy our dinner, after we are served with several things that is
CAT	wow! i can say that all the good reviews. excellent food and service. this is a great place for sushi, and over spiced with clients the Hunan beef is amazing! they just feels like it's one of her favorite sushi restaurant.
MAT	wow ! i can say food was okay and service was awesome. i must say we really enjoyed it ! took my kids and wife, they were cooking the waffles with spices and had the best seasoning in what you can describe the place, had a great comfort food
MUCOCO	wow! i can say that's good. nice service and the food. i was excited to eat some Chinese dishes but just try to pick up our order and give it a long day of sitting down. obviously, we have been so far because the pork and avocado came delicious!

of our method in improving style transfer accuracy for transferred text in multi-attribute scenarios. Sentiment and gender are binary classes and just a transfer from one class to another, so there is no problem of multiple directions and thereby no improvement by our method. Moreover, the full model also achieves better scores on PPL. This result confirms that our method improves attribute accuracy without sacrificing sentence fluency and, in some cases, even leads to better outcomes.

In addition, we can see that even if we detect conflicts in almost every transfer process, there are still 38% of transferred sentences that satisfy all attributes. After conflict resolution, the overall accuracy has improved. This confirms that the conflicts in the directions of gradients indeed affect the satisfaction of different attributes, but not every conflict will lead to an attribute-incomplete generation text, it increases the corresponding possibility of such occurrences.

5 Conclusions

In this study, we presented a novel mathematical programming approach for coordinating and controlling multi-attribute style transfer, which we evaluated on the YELP dataset. Our experiments demonstrated that this method can effectively enhance the accuracy of multi-attribute transfer, while maintaining the accuracy of each individual attribute. Furthermore, this method allows pre-trained auto-encoders to efficiently transmit language attributes, eliminating the need for additional tuning and enabling faster and more scalable learning.

Moving forward, we plan to extend our approach to cross-lingual style transfer tasks and explore ways to optimize the algorithm's time efficiency.

References

1. Cheng, Y., Gan, Z., Zhang, Y., Elachqar, O., Li, D., Liu, J.: Contextual text style transfer. In: Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020, vol. EMNLP 2020, pp. 2915–2924 (2020)
2. Dai, N., Liang, J., Qiu, X., Huang, X.: Style transformer: unpaired text style transfer without disentangled latent representation. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 5997–6007 (2019)
3. Dorn, W.S.: Duality in quadratic programming. Q. Appl. Math. **18**, 155–162 (1960)
4. Frank, M., Wolfe, P., et al.: An algorithm for quadratic programming. Naval Res. Logist. Q. **3**(1–2), 95–110 (1956)
5. Fu, Z., Tan, X., Peng, N., Zhao, D., Yan, R.: Style transfer in text: exploration and evaluation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
6. Goyal, N., Srinivasan, B.V., Natarajan, A., Sancheti, A.: Multi-style transfer with discriminative feedback on disjoint corpus. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, 6–11 June 2021, pp. 3500–3510 (2021)
7. Jin, D., Jin, Z., Hu, Z., Vechtomova, O., Mihalcea, R.: Deep learning for text style transfer: a survey. Comput. Linguist. **48**, 155–205 (2022)
8. Kumar, S., Malmi, E., Severyn, A., Tsvetkov, Y.: Controlled text generation as continuous optimization with multiple constraints. Adv. Neural Inf. Process. Syst. **34**, 14542–14554 (2021)
9. 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)
10. Lample, G., Subramanian, S., Smith, E.M., Denoyer, L., Ranzato, M., Boureau, Y.: Multiple-attribute text rewriting. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019 (2019)
11. Li, J., Jia, R., He, H., Liang, P.: 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, New Orleans, Louisiana, USA, 1–6 June 2018, Volume 1 (Long Papers), pp. 1865–1874 (2018)
12. Liu, D., Fu, J., Zhang, Y., Pal, C., Lv, J.: Revision in continuous space: Unsupervised text style transfer without adversarial learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 8376–8383 (2020)
13. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
14. Luo, F., et al.: A dual reinforcement learning framework for unsupervised text style transfer. In: Proceedings of the Twenty-Eighth International Joint Conference on

Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019, pp. 5116–5122 (2019)

15. Malmi, E., Severyn, A., Rothe, S.: Unsupervised text style transfer with padded masked language models. arXiv preprint arXiv:2010.01054 (2020)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
17. Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., Black, A.W.: Style transfer through back-translation. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, 15–20 July 2018, Volume 1: Long Papers, pp. 866–876 (2018)
18. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**, 9 (2019)
19. Raedt, M.D., Godin, F., Buteneers, P., Develder, C., Demeester, T.: A simple geometric method for cross-lingual linguistic transformations with pre-trained autoencoders. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event/Punta Cana, 7–11 November 2021, pp. 10108–10114 (2021)
20. Reid, M., Zhong, V.: LEWIS: levenshtein editing for unsupervised text style transfer. In: Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, 1–6 August 2021, vol. ACL/IJCNLP 2021, pp. 3932–3944 (2021)
21. dos Santos, C.N., Melnyk, I., Padhi, I.: Fighting offensive language on social media with unsupervised text style transfer. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, 15–20 July 2018, Volume 2: Short Papers, pp. 189–194 (2018)
22. Shen, T., Lei, T., Barzilay, R., Jaakkola, T.: Style transfer from non-parallel text by cross-alignment. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
23. Toshevska, M., Gievska, S.: A review of text style transfer using deep learning. IEEE Trans. Artif. Intell. **3**, 669–684 (2021)
24. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
25. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, pp. 1096–1103 (2008)
26. Wang, K., Hua, H., Wan, X.: Controllable unsupervised text attribute transfer via editing entangled latent representation. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
27. Xu, J., et al.: Unpaired sentiment-to-sentiment translation: a cycled reinforcement learning approach. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, 15–20 July 2018, Volume 1: Long Papers, pp. 979–988 (2018)
28. Yi, X., Liu, Z., Li, W., Sun, M.: Text style transfer via learning style instance supported latent space. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 3801–3807 (2021)
29. Zhang, Y., Xu, J., Yang, P., Sun, X.: Learning sentiment memories for sentiment modification without parallel data. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31–November 4, 2018, pp. 1103–1108 (2018)