StyleFlow: Disentangle Latent Representations via Normalizing Flow for Unsupervised Text Style Transfer

Kangchen Zhu[1], **Zhiliang Tian**[1,†], **Jingyu Wei**[1], **Ruifeng Luo**[1], **Yiping Song**[2], **Xiaoguang Mao**[1]
[1]College of Computer, National University of Defense Technology, Hunan, China
[2]College of Science, National University of Defense Technology, Hunan, China
{zhukangchen18, tianzhiliang, weijingyu, luoruifeng21, songyiping, xgmao}@nudt.edu.cn

Abstract

Unsupervised text style transfer aims to modify the style of a sentence while preserving its content without parallel corpora. Existing approaches attempt to separate content from style, but some words contain both content and style information. It makes them difficult to disentangle, where unsatisfactory disentanglement results in the loss of the content information or the target style. To address this issue, researchers adopted a "cycle reconstruction" mechanism to maintain content information, but it is still hard to achieve satisfactory content preservation due to incomplete disentanglement. In this paper, we propose a new disentanglement-based method, StyleFlow, which effectively avoids the loss of contents through a better cycle reconstruction via a reversible encoder. The reversible encoder is a normalizing flow that can not only produce output given input but also infer the exact input given the output reversely. We design a stack of attention-aware coupling layers, where each layer is reversible and adopts the attention mechanism to improve the content-style disentanglement. Moreover, we propose a data augmentation method based on normalizing flow to enhance the training data. Our experiments on sentiment transfer and formality transfer tasks show that StyleFlow outperforms strong baselines on both content preservation and style transfer.

Keywords: Text Style Transfer, Normalizing Flow, Data Augmentation, Unsupervised Learning

1. Introduction

Text style transfer aims to convert a sentence with a specific style into another style (e.g., sentiment, formality) while retaining the original content (Hu et al., 2022). Due to the lack of large parallel corpora, most researchers focused on unsupervised style transfer (Shen et al., 2017; Dai et al., 2019; Lee et al., 2021; Reif et al., 2021).

On the one hand, some researchers encode the content information of the input sentences and then add the target style on it (Lample et al., 2018; Dai et al., 2019). Those methods maintain the content information well but do not effectively remove the source style information, so the transferred sentence may not express the target style information well. Some recent works utilize large language models (LLMs) (Touvron et al., 2023) that directly feed the target style to LLMs through prompting (Suzgun et al., 2022; Luo et al., 2023). However, researchers (Ji et al., 2023) find that LLMs tend to generate creative content and do not fully maintain the original content in transferred sentences.

On the other hand, researchers attempt to disentangle the style and content information explicitly or implicitly and then replace the style with a new one (Shen et al., 2017; Reid and Zhong, 2021). Explicit disentanglement methods utilize style marker detection at the token level to explicitly replace tokens that carry vital style information. This approach heavily relies on the accuracy of style marker detection and is only effective for sentence types where the style is evident (Hu et al., 2022). The implicit way disentangles word embeddings into style and content representations and combines the source-content and target-style representations to generate target-style sentences. However, some words often carry both style and content information. For example, the term "delicious" conveys not only a positive sentiment (style) but also the taste of food (content). Consequently, these disentanglement-based methods struggle to completely disentangle content and style representations, leading to either the retention of style information or the loss of content information.

To tackle this issue, researchers try to keep the content information via the cycle reconstruction (reconstructing the source sentence with the transferred sentence and source style information) (Lee et al., 2021; Jin et al., 2022; Wen et al., 2023). However, an incomplete disentanglement makes the models inevitably miss some content information, which means they can hardly achieve a complete cycle reconstruction (i.e., the cycle reconstruction loss is always not zero) (Yi et al., 2021) and perform unsatisfactorily in content preservation.

In this paper, we propose StyleFlow, a novel disentanglement-based method, to sharply avoid the loss of content information via a reversible encoder. The reversible encoder is a normalizing flow (Dinh et al., 2014) that can generate outputs

[†] Corresponding Author.

LREC-COLING 2024, pages 15384–15397
20-25 May, 2024.

by input and infer the exact input by the output reversely. It means the output covers the input information in the reversible encoder. Unlike typical encoder-decoder frameworks, StyleFlow has only a reversible encoder to accomplish encoding and decoding. It achieves the encoding in the forward process and the decoding in the reverse process. Specifically, we design a stack of attention-aware coupling layers in the reversible encoder, where each layer is reversible and adopts the attention mechanism (Vaswani et al., 2017) to enhance the effect of content-style disentanglement. They gradually achieve content-style disentanglement layer-by-layer. The structure ensures the reversibility of the whole encoder. Its reversibility reduces the loss of content by bridging a bijection between the input and the output while ensuring the input can be inferred exactly from the output. To further improve the performance, we propose a data augmentation method with the bijection of the normalizing flow. It adds small perturbations to the source sentence representations and reversely encodes them to generate pseudo sentences located in the original sentences' neighborhood.

Our contributions are fourfold: (1) We propose a novel disentanglement-based method with a reversible encoder. Its reversibility sharply avoids content loss during the forward and reverse processes. We also provide theoretical proof of the reversibility. (2) We propose a stack of attention-aware coupling layers to disentangle the content and style information layer by layer gradually. (3) We propose a data augmentation method with the bijection of the normalizing flow. (4) Experiments indicate StyleFlow outperforms other baselines on most metrics.

2. Related Work

2.1. Unsupervised Text Style Transfer

In recent years, there has been a growing interest in unsupervised text style transfer due to the limited availability of large parallel corpora for training. One intuitive idea is to explicitly disentangle the style from the content by removing explicit style markers and infusing target style markers. Then, it trains a generator to produce sentences with the desired style. Different strategies have been proposed to detect attribute markers. Li et al. (2018) compute the relative co-occurrence frequency with style information, while Zhang et al. (2018) employ LSTM and Sudhakar et al. (2019) endorse a BERT-based classifier for marker detection, which has shown superior performance. Wu et al. (2019) combine frequency-ratio methods with attention-driven techniques to prioritize attribute markers. To enhance marker identification, Lee (2020) introduce a word importance scoring method based on attribute probability changes after token removal, and Reid and Zhong (2021) use Levenshtein edit operations to edit the prototype.

Although explicit disentanglement methods have achieved good results to some extent, they heavily depend on the classifier's ability to recognize markers. Researchers have explored implicit approaches to disentangle content and style representations. Shen et al. (2017) use adversarial discriminators to align source and transferred content distributions, Fu et al. (2018) concatenate discerned content distributions, and John et al. (2018) employ multiple loss functions to constrain sentences within a latent domain and separate them into content and style sub-domains. Additionally, Lee et al. (2021) use "reverse attention" to further disentangle content and style representations.

In addition to the disentanglement methods mentioned above, some researchers argue that it is not necessary to perform disentanglement and instead propose directly modulating the stylistic attributes of sentences to achieve style transfer (Dai et al., 2019; Bayer et al., 2022; Diao et al., 2023). Dai et al. (2019) propose StyleTransformer to overlay style attributes onto sentence embeddings from transformer, Luo et al. (2019) utilize reinforcement learning and control conditions for effective style modulation, and Jin et al. (2019) approximate unsupervised training to supervised training using pseudo data to enhance model performance.

Recently, the prompt-based technique has been a promising approach for style transfer. It generates texts in a training-free or exemplar-free manner. Reif et al. (2021) leverage large pre-trained language models (PLMs) to interpret prompts and generate sentences with varied styles. Suzgun et al. (2022) apply prompts to PLMs and incorporate re-ranking mechanisms to select the most suitable style sentence. Besides, Luo et al. (2023) formulates the generation process as a classification problem and employs discrete search techniques to generate sentences in the target style. However, researchers (Ji et al., 2023) find that LLMs tend to generate creative content and do not fully maintain the original content in transferred sentences.

2.2. Normalizing Flow on NLP

Normalizing flow is a generative model class that offers effective sampling and precise density evaluation. This methodology has been applied to natural language processing tasks. Ma et al. (2019) introduce Flowseq to combine non-autoregressive conditional sequence generation with normalizing flow. Li et al. (2020) utilize normalizing flow to determine the cosine similarity of BERT embeddings. Zhao et al. (2021) propose multi-split reversible transformers based on normalizing flow.

Tang et al. (2021) develop a flow-based language model for machine translation and Liu et al. (2022) apply normalizing flow to latent sentence representation optimization. Furthermore, normalizing flow has also been leveraged in addressing text style transfer tasks. Samanta et al. (2021) use normalizing flow to process embeddings from BERT and only take the final dimension as the style representations, which may not effectively capture the characteristics of style information. Yi et al. (2021) use normalizing flow as an assistive technology to sample stylistic features from the latent style space. However, neither uses normalizing flow as the main framework of models. In contrast, StyleFlow innovatively applies normalizing flow as the main framework of the model to achieve text style transfer. It facilitates a more efficient disentanglement of content and style and enhances content preservation in implicit disentanglement methods.

3. Approach

3.1. Coupling Layer

The coupling layer is a type of normalizing flow that allows for efficient computations in both forward and inverse directions by employing linear transformations with special structures, often utilizing triangular Jacobian matrices (Dinh et al., 2014). In the forward process of the coupling layer, the input x is split into two parts: $x_{i\leq d}$ and $x_{i>d}$. The first part $x_{i\leq d}$ is directly copied to obtain $y_{i\leq d}$. The The second part $x_{i>d}$ is updated $x_{i>d}$ by feeding $x_{i\leq d}$ into arbitrary functions F and H to calculate the coefficients β and γ ($\beta_i = F(x_{i\leq d}), \gamma_i = H(x_{i\leq d})$), which are used to calculate $y_{i>d} = \beta_i x_i + \gamma_i$. Finally, we concatenate $y_{i\leq d}$ and $y_{i>d}$ to obtain the output y. The coupling layer is a powerful tool for constructing a tractable and expressive function, enabling efficient computations in both directions. As a result, it has been widely used in many real-world applications such as image generation and density estimation (**?**).

3.2. Reversible Encoder

Unlike the typical encoder-decoder scheme, StyleFlow has only a reversible encoder, consisting of embedding layers and a stack of attention-aware coupling layers, as shown in Fig. 1. **Embedding layers** consist of a word embedding layer and a style embedding layer. The word embedding layer is a neural network that maps the input sentence $\mathbf{x}$ (a sequence of words) into a series of embeddings, denoted as $\mathbf{E} = [\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_T}]$. The style embedding layer is a learnable embedding layer representing the target style information; its input is a style label. **Attention-aware coupling layers** are a series of reversible layers to disentangle the content and style representations. Each attention-aware coupling layer consists of an *attention-aware disentanglement layer* and a *basic coupling layer* as mentioned in Sec. 3.1. Firstly, we consider the attention-aware disentanglement layer as a style classifier to disentangle the content and style representations according to attention scores with two steps: (1) Obtaining the attention scores reflecting the stylistic elements of embeddings. We used a self-attention layer, which is initialized with a self-attention layer in the pre-trained RoBERTa-Large (Liu et al., 2019) classifier, to calculate attention scores about the style of the input embeddings. (2) Splitting style and content representations according to their attention scores. We set the portion of embeddings with attention scores that exceed the mean score as style representations and those with attention scores below the mean score as content representations. Secondly, for the basic coupling layer, we first copy content representations directly and then feed content representations into the F and H (i.e., transformer block) to calculate the coefficients β and γ, which are used to obtain the new style representations. Finally, we concatenate the content and style representations from the previous layer to form a new input fed into the next layer and repeat the above operations.

3.3. Encoding with Disentanglement

In the encoding step, as the red arrows (from left to right) shown in Fig. 1, we first feed the input sentence $\mathbf{x}$ into the word embedding layer to obtain the word embeddings $\mathbf{E_x}$. Next, $\mathbf{E_x}$ is passed to a stack of attention-aware coupling layers. Suppose in the k-th layer ($k \in [1, K]$ and K is the total), the input is $\mathbf{Z}$. We first use the attention-aware disentanglement layer to calculate attention scores of $\mathbf{Z}$ and disentangle $\mathbf{Z}$ into style representations $\mathbf{z_s}$ and content representations $\mathbf{z_c}$ by the scores. Then, the basic coupling layer takes $\mathbf{z}_s$ and $\mathbf{z}_c$ as input to generate new style representations $\mathbf{z}'_\mathbf{s}$ and new content representations $\mathbf{z}'_\mathbf{c}$, which act as $k+1$-th layer's input. A stack of K layers repeatedly conducts the operations one by one. The encoding outputs of the reversible encoder are the outputs of the last attention-aware coupling layer (content representations $\mathbf{Z_c}$ and style representations $\mathbf{Z_s}$.

3.4. Reverse Encoding

In the reverse encoding step, as the blue arrows (from right to left) shown in Fig. 1, we first feed the target style label $\hat{s}$ to the style embedding layer to get a target style embedding $\mathbf{Z_{\hat{s}}}$. Next, we fuse $\mathbf{Z_{\hat{s}}}$ and the content representation $\mathbf{Z_c}$ from the encoding step to get target-style embeddings $\hat{\mathbf{Z}}$ and input $\hat{\mathbf{Z}}$ to attention-aware coupling layers in the reverse step. Specifically, in the k-th layer, we use

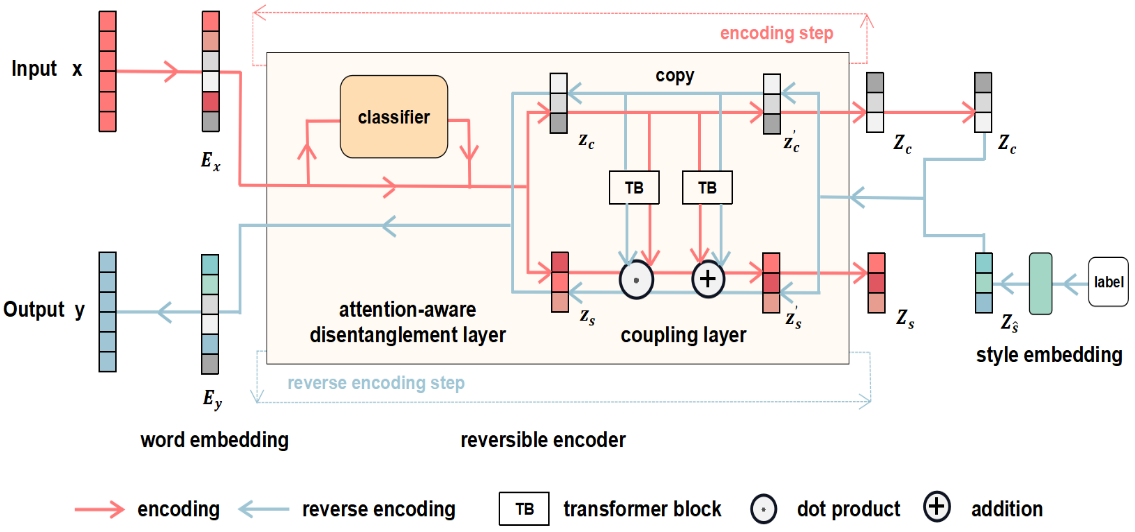

Figure 1: Model overview of StyleFlow. The word embedding input first passes the attention-aware disentanglement layer to compute attention scores of style information for disentanglement. Then, the coupling layer disentangles the word embedding into content and style space layer-by-layer. Lastly, source content space and target-style embeddings are fed into the encoder to generate the target-style sentence by the reverse encoding process.

the reverse operation of attention-aware disentanglement layer to obtain $\hat{\mathbf{z}}_\mathbf{c}$ and $\hat{\mathbf{z}}_\mathbf{s}$ from $\hat{\mathbf{Z}}$ by using the same disentanglement as those in the encoding step. It means that in the inverse process, we get new representations $\hat{\mathbf{z}}_\mathbf{c}$ and $\hat{\mathbf{z}}_\mathbf{s}$ by reversing the operations of disentanglement in the encoding step instead of recalculating the attention scores. By the above process, we guarantee the reversibility of the attention coupling layers in both encoding and reverse encoding steps.

By repeating the above operations for a stack of K layers, we obtain all the words' representations $\mathbf{E}_\mathbf{y}$ of the target-style sentence $\mathbf{y}$ from the first attention-aware coupling layer. At last, we conduct the multiplication between the representation $\mathbf{E}_\mathbf{y}$ and a learnable vector to get scores for every word over the vocabulary. Scores are fed into a softmax layer to reach the target distribution for all words in the sentence. Finally, we generate the target sentence $\mathbf{y}$ according to the distribution.

3.5. Analysis of Reversibility

The reversibility of the reversible encoder comes from the normalizing flow $\mathcal{F}$. Suppose we have observations x from an unknown data distribution $p_\mathcal{X}$ over $\mathcal{X} \subset \mathbb{R}^d$, and a tractable prior probability distribution $p_\mathcal{Z}$ over $\mathcal{Z} \subset \mathbb{R}^k$, from which we sample a latent variable z. Eq. 1 shows two requirements for reversibility: (1) the inverse function is easy to calculate; (2) the determinant of Jacobian is calculable.

$$p_\mathcal{X}(x) = p_\mathcal{Z}(\mathcal{F}(x)) \cdot \left| \det \left(\frac{\partial \mathcal{F}(x)}{\partial x^T} \right) \right| \tag{1}$$

Firstly, the inverse function in the coupling layer can be done directly by the inverse process ($x_{1:d} = z_{1:d}$ and $x_{d+1:D} = \frac{z_{d+1:D} - H(z_{1:d})}{F(z_{1:d})}$). Secondly, the Jacobian determinant is a triangular matrix (the product of its diagonal terms) as Eq. 2, where $diagF(x_{1:d})$ is the diagonal matrix whose diagonal elements correspond to the vector $F(x_{1:d})$.

$$\frac{\partial y}{\partial x}^T = \begin{bmatrix} \mathbb{I}_d & 0 \\ \frac{\partial y_{d+1:D}}{\partial x^T_{1:d}} & diagF(x_{1:d}) \end{bmatrix} \tag{2}$$

Due to the limited expression ability of a single layer, we connect multiple coupling layers to obtain a flow-based encoder. The reverse function f^{-1} exists and is easy to calculate as Eq. 4, where $\circ$ denotes composition, ℓ is the number of layers.

$$f = f_1 \circ \cdots \circ f_{\ell-1} \circ f_\ell, \tag{3}$$

$$f^{-1} = f_\ell^{-1} \circ \cdots \circ f_2^{-1} \circ f_1^{-1}. \tag{4}$$

Similarly, since every single layer's Jacobian is a triangular matrix, the Jacobian of a flow-based encoder is a new triangular matrix composed of small triangular matrices and easy to calculate by the product of its diagonal terms. Note that the attention mechanism only changes the way the coupling layer splits inputs and does not affect the reversibility of the structure, so the whole flow-based encoder is reversible. In addition, we need to claim that our entire style transfer processing is not entirely reversible since (1) the target style directly replaces a part of representations; (2) the randomness makes the processing nonreversible if we apply top-k sampling to generate y.

3.6. Objective Functions

(1) **Self reconstruction loss.** The loss $\mathcal{L}_{self}$ is the average of all words' cross-entropy loss between the reconstructed $\hat{x} = f^{-1}(f(x), s)$ and input x, where f denotes the encoding and f^{-1} means the reverse encoding. (2) **Cycle reconstruction loss.** We transfer the input x with another style $\hat{s}$ to obtain a transferred sentence y by $y = f^{-1}(f(x), \hat{s})$ and transfer y with the original style s to reconstruct the input by $\hat{x}' = f^{-1}(f(y), s)$. The loss $\mathcal{L}_{cycle}$ is the average of all words' cross-entropy loss between $\hat{x}'$ and x. (3) **Content loss.** The content representation $\mathbf{z}_c$ of the input x and the content representation $\mathbf{z}_{\hat{c}}$ of the transferred sentence y should be similar: $\mathcal{L}_c = \mathbb{E}\,\|\mathbf{z}_c - \mathbf{z}_{\hat{c}}\|_2^2$. (4) **Style transfer loss.** The transferred sentence y would be regarded as the target style $\hat{s}$ by the style classifier C: $\mathcal{L}_s = -\mathbb{E}[\log P_C(\hat{s}|y)]$. Our objective is to minimize $\mathcal{L} = \lambda_1\mathcal{L}_{self} + \lambda_2\mathcal{L}_{cycle} + \lambda_3\mathcal{L}_c + \lambda_4\mathcal{L}_s$, where $\lambda_i, i = \{1, 2, 3, 4\}$ are hyper-parameters to balance the losses.

4. Data Augmentation based on Normalizing Flow

Text style transfer corpora usually have limited samples [1], resulting in poor model robustness. We propose a data augmentation technique based on normalizing flow inspired by Yüksel et al. (2021) to address this issue. Specifically, the reversible encoder establishes a bijection ($\mathcal{X} \leftrightarrow \mathcal{Z}$) between the sample space $\mathcal{X}$ and the latent space $\mathcal{Z}$. We add perturbations $\mathcal{P}$ to the latent space $\mathcal{Z}$ to generate pseudo samples via reverse encoding ($\mathcal{F}^{-1}(\mathcal{F}(x) + \mathcal{P}(\mathcal{F}(x), \epsilon_1, \epsilon_2))$. Parameter ϵ_1 and ϵ_2 controls the perturbation size, and the pseudo samples help extend the corpus for further training.

During training, we obtain the latent space $z_i = \mathcal{F}(x_i)$ corresponding to a given sample x_i with its associated label l_i ($1 \leq i \leq N$) and use the trained normalizing flow to search for $\triangle_{z_i} \in \mathcal{Z}$ such that the loss achieved by the generated sample $\hat{x_i} = \mathcal{F}^{-1}(z_i + \triangle_{z_i})$ is maximal. Adversarial perturbations can help the model find samples that are difficult to handle during cycle reconstruction, and $\triangle_{z_i}$ indicates the direction where the current model cannot perform well by searching for the fastest direction of gradient descent.

$$\mathcal{P} = \triangle_{z_i}^{\star} = \underset{\epsilon_1 \leq \|\triangle_{z_i}\|_{\ell_p} \leq \epsilon_2}{\arg\max} \; \mathcal{L}_\theta(\mathcal{F}^{-1}(z_i + \triangle_{z_i}), l_i) \quad (5)$$

where $\mathcal{L}_\theta$ is the loss of function of the classifier, and ℓ_p denotes the normalization method. In practice, we use the number of steps k and step size α to optimize $\triangle_{z_i} \in \mathcal{Z}$ as follows: (1) Initialize a random $\triangle_{z_i}^0$ with $\epsilon_1 \leq \| \triangle_{z_i} \|_{\ell_p} \leq \epsilon_2$. (2) Iteratively update $\triangle_{z_i}^j$ for $j = 1, \ldots, k$ number of steps using the following step size α and choose ℓ_∞ as ℓ_p with the projection operator σ:

$$\triangle_{z_i}^j = \sigma\left(\triangle_{z_i}^{j-1} + \alpha \cdot \frac{\nabla\mathcal{L}_\theta(\mathcal{F}^{-1}(z_i + \triangle_{z_i}^{j-1}), l_i)}{\|\nabla\mathcal{L}_\theta(\mathcal{F}^{-1}(z_i + \triangle_{z_i}^{j-1}), l_i)\|_{\ell_\infty}}\right) \quad (6)$$

where $\sigma(x_i) = \max(\epsilon_1, \min(\epsilon_2, x_i))$ ensures that $\epsilon_1 \leq \| \triangle_{z_i} \|_{\ell_\infty} \leq \epsilon_2$ and gradient is with respect to $\triangle_{z_i}^{j-1}$. (3) Output $\mathcal{P}(z_i, \epsilon) = \triangle_{z_i}^k$. Multi-step search encourages the model to search toward the fastest gradient descent in the cycle reconstruction loss. In the case where $\ell_p = \ell_\infty$, we replace normalization of gradient with the $sign(\cdot)$ operator:

$$\triangle_{z_i}^j = \sigma(\triangle_{z_i}^{j-1} + \alpha \cdot sign(\nabla\mathcal{L}_\theta(\mathcal{F}^{-1}(z_i + \triangle_{z_i}^{j-1}), l_i))) \quad (7)$$

Therefore, we obtain the adversarial perturbation $\triangle_{z_i}$ through the above steps, and obtain the pseudo sample $\hat{x_i}$ by $\hat{x_i} = \mathcal{F}^{-1}(F(x_i) + \triangle_{z_i})$. This data augmentation method utilizes the reversible normalizing flow and adversarial training to obtain pseudo samples closer to real samples.

5. Experiments

5.1. Experimental Settings

Datasets We conducted experiments on two standard style transfer tasks: (1) **Sentiment transfer.** We used the Yelp Review Dataset (Yelp) processed by Li et al. (2018) and IMDb Movie Review Dataset (IMDb) by Dai et al. (2019) for sentiment style. YELP contains restaurant and business reviews, and IMDb consists of movie reviews written by online users. (2) **Formality transfer.** We used Grammarly's Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault, 2018) and chose the "Family & Relationships" domain for formality transfer. GYAFC consists of sentences extracted from a question-answering forum (Yahoo Answers).

Implementation Details All transformer blocks in StyleFlow are implemented as the basic block of a T5-base encoder-decoder model with eight attention heads. The hidden size and positional encoding size are all 256 dimensions. The word embeddings and the target-style embeddings of 300 dimensions are both learned from scratch. The optimizer is Adam (Kingma and Ba, 2014) with

[1] Usually less than 0.3M samples, much fewer than other generation tasks.

Method	Model		Para	Yelp						IMDb		
				ACC ↓	s-sBLEU ↑	r-sBLEU ↑	GM ↑	HM ↑	PPL ↓	ACC ↓	s-sBLEU ↑	PPL ↓
Explicit disentanglement	DeleteAndRetrieve (Li et al., 2018)		<1B	87.7	29.1	10.4	30.2	18.6	60	58.8	55.4	57
	MaskAndInfill (Wu et al., 2019)			**97.3**	32.5	14.4	37.4	14.4	54	93.1	58.1	53
	SST (Lee, 2020)			70.4	49.1	12.7	29.9	21.5	197	76.5	45.5	71
	MaskTransfomer$_C$ (Wu et al., 2020)			91.8	54.6	19.3	42.1	31.9	81	**94.7**	63.9	92
	MaskTransfomer$_M$ (Wu et al., 2020)			88.3	55.4	20.1	42.1	32.8	75	91	66.2	86
	LEWIS (Reid and Zhong, 2021)			93.1	58.5	24	47.3	38.2	66	94.4	68.9	76
Implicit disentanglement	Cross-alignment (Shen et al., 2017)		<1B	76.3	13.2	4.3	19.3	8.1	244	63.9	1.1	**29**
	ControlledGen (Hu et al., 2017)			88.8	45.7	14.3	35.6	24.6	219	94.3	62.2	143
	Disentangled (John et al., 2018)			91.7	16.7	6.7	24.8	12.5	**26**	62.3	14.4	31
	StyIns (Yi et al., 2021)			90.8	53.9	20.3	39.6	29.1	109	83.8	64.6	121
	RACoLN (Lee et al., 2021)			91.3	59.4	20	42.7	32.8	60	83.1	70.9	45
	TSST (Xiao et al., 2021)			91.8	59.3	19.8	42.6	32.6	108	82.7	69.3	101
W/o disentanglement	DualRL (Luo et al., 2019)		<1B	87.9	58.9	25.9	47.8	40	133	79.2	68.5	118
	IMaT (Jin et al., 2019)			94.4	38.5	22.5	46.1	36.3	49	90.2	34.1	62
	StyleTransfomer$_C$ (Dai et al., 2019)			91.8	52.8	22.9	45.9	36.7	223	82.3	67.5	108
	StyleTransfomer$_M$ (Dai et al., 2019)			85.5	63	26.4	47.6	40.4	75	80.3	70.5	105
	M&M LM (Mireshghallah et al., 2022)			90.6	59.2	20.1	42.7	32.9	78	81.4	69.2	85
W/o disentanglement (LLMs-based)	AugZS$_{0\text{-shot}}$ (Reif et al., 2021)	GPT-3	175B	63.2	45.7	19.9	35.5	19.9	55	58.6	58.4	69
	AugZS$_{4\text{-shot}}$ (Reif et al., 2021)	GPT-3	175B	78.5	48.3	23.2	42.7	35.8	77	73.4	61	91
	P&R$_{0\text{-shot}}$ (Suzgun et al., 2022)	GPT2-XL	1.6B	87.2	28.7	14.8	35.9	25.3	65	57.7	54.2	63
	P&R$_{4\text{-shot}}$ (Suzgun et al., 2022)	GPT2-XL	1.6B	87.4	47.5	23	44.8	36.4	80	83.5	68.1	52
	P&R$_{0\text{-shot}}$ (Suzgun et al., 2022)	GPT-J-6B	6B	61.3	34.2	14.3	29.6	23.2	54	57.6	59	47
	P&R$_{4\text{-shot}}$ (Suzgun et al., 2022)	GPT-J-6B	6B	87.9	47.4	23	45	36.5	80	83.1	55.9	69
Ours	StyleFlow w/o DA		<1B	93.7	61.2	28.6	51.8	43.8	42	93.7	72.4	45
	StyleFlow w DA			95.1	**63.8**	**29.2**	**52.7**	**44.7**	49	94.5	**73.6**	58

Table 1: Automatic evaluation results on Yelp and IMDb datasets. #Para: Number of parameters. We took the average value of five runs as experimental results. For MaskTransformer and StyleTransformer, C and M refer to conditional and multi-class settings, respectively. For LLMs-based methods, n-shot means using n exemplars in the few-shot setting. Bolded data indicates optimal performance and underlined data indicates suboptimal performance. Note that all other tables have similar settings.

10^{-4} initial learning rate for training. To match the traditional 16-layer (8 encoder blocks and eight decoder blocks) transformer model parameters, we set the length K of flow to 8. For balancing parameters of the objective functions, we chose the best-performed $1/6$, $1/2$, $1/6$, and $1/6$ for λ_1, λ_2, λ_3 and λ_4 in Eq. 3.6. It shows that the cycle reconstruction loss is the most important loss function. For data augmentation in Eq. 6, we use $k = 10$ and $\alpha = 0.5$. The hyper-parameters ϵ_1 ϵ_2 are set to 0.1 and 1.0, respectively. We conducted our experiments on RTX 3090Ti GPUs.

Model		GYAFC					
		ACC ↑	s-sBLEU ↑	r-sBLEU ↑	GM ↑	HM ↑	PPL ↓
DeleteAndRetrieve		61.1	27.6	21.2	34	31.5	110
MaskAndInfill		75.2	29.5	25.4	43.7	38	105
SST		53.9	46.2	36.9	44.6	43.8	178
MaskTransfomer$_C$		62.7	59.7	42.6	51.7	50.7	129
MaskTransfomer$_M$		58.6	60.2	43.3	50.4	49.8	117
LEWIS		75.2	62.9	51.8	62.4	61.3	82
Cross-alignment		61.6	2.2	3.25	14.1	6.2	37
ControlledGen		68.2	55.9	41.6	53.3	51.7	195
Disentangled		67.5	7.4	8.1	23.4	14.5	**24**
StyIns		67.8	61.9	46.7	56.3	55.3	92
RACoLN		74.7	59.9	50.6	61.5	60.3	78
TSST		74.4	63.7	50.5	60.1	61.2	103
DualRL		71.7	52.8	41.9	54.6	52.7	159
IMaT		72.1	59.4	38.2	52.5	49.9	33
StyleTransfomer$_C$		60.3	61.2	43.9	51.5	50.8	168
StyleTransfomer$_M$		57.6	63.1	46.4	51.7	51.4	126
M&M LM		72.2	60.2	27.7	44.7	40	119
AugZS$_{0\text{-shot}}$	GPT-3	48.3	53.6	46.6	47.4	47.4	54
AugZS$_{4\text{-shot}}$	GPT-3	59.6	58.2	51.2	55.2	55.1	64
P&R$_{0\text{-shot}}$	GPT2-XL	60.9	26.7	25.8	39.6	36.2	122
P&R$_{4\text{-shot}}$	GPT2-XL	82.2	32.7	41.9	58.7	55.5	58
P&R$_{0\text{-shot}}$	GPT-J-6B	58.6	31.3	25.2	42.8	40.8	99
P&R$_{4\text{-shot}}$	GPT-J-6B	68.8	47.9	52.3	60	59.4	49
StyleFlow w/o DA		85.4	65.4	54.1	68	66.2	61
StyleFlow w DA		**85.6**	**67.6**	**56**	**69.2**	**67.7**	67

Table 2: Automatic evaluation results on GYAFC.

Evaluation Metrics (1) Style transfer accuracy (**ACC**) measures whether a generated output is correctly transferred. We used a finetuned RoBERTa-Large (Hartmann et al., 2023) for sentiment classification. This model correctly classifies 98.2% of the sentiment classification test set by Shen et al. (2017). We finetuned another classifier initialized with RoBERTa-Large (Liu et al., 2019) for formality classification with accuracy 93.0%. (2) BLEU is the standard metric for measuring semantic content preservation. It is shown that SacreBLEU is a more reliable and accessible metric than BLEU (Post, 2018; Liu et al., 2020; Suzgun et al., 2022). We used SacreBLEU (sBLEU) implementation to compute both self-sBLEU (**s-sBLEU**) and reference-BLEU (**r-sBLEU**) scores. Whereas s-sBLEU indicates the degree to which the model directly copies the source, r-sBLEU helps measure the distance of generated sentences from the ground-truth references [2]. (3) Geometric mean and harmonic mean. We use the accuracy and r-sBLEU averages to evaluate the overall performance of text style transfer following Luo et al. (2019). For brevity, **GM** and **HM** represent the geometric mean and harmonic mean, respectively. (4) **PPL.** For fluency, we used GPT-2 (117M) (Radford et al., 2019) to measure the perplexity (PPL) of transferred sentences. The sentences with smaller PPL scores are considered

[2] Since the IMDb dataset lacks human references as ground truth, we do not use r-sBLEU metric on the IMDb dataset. Similarly, we do not evaluate the scores of human references on IMDb in human evaluation (Sec. 5.2).

	Yelp			IMDb			GYAFC		
	Style ↑	Content ↑	Fluency ↑	Style ↓	Content ↑	Fluency ↑	Style ↑	Content ↑	Fluency ↑
Human (Upper Bound)	**4.8**	**4.7**	**4.9**	-	-	-	**4.8**	**4.7**	**4.8**
LEWIS (Reid and Zhong, 2021)	4.5	4.3	4.4	**4.6**	4.3	4.1	4.4	4.2	4.1
RACoLN (Hu et al., 2017)	4.2	3.9	4.4	4.2	4.5	**4.4**	4.3	4.0	4.3
TSST (Xiao et al., 2021)	4.1	3.8	3.9	4.1	4.4	3.8	4.0	4.3	3.9
DualRL (Wu et al., 2020)	3.9	4.4	4.0	3.8	4.3	3.3	3.8	3.9	3.7
P&R$_{4\text{-shot}}$ (Lee et al., 2021)	3.9	4.2	4.2	4.2	3.8	4.3	3.9	3.8	4.6
StyleFlow	4.6	4.6	4.5	**4.6**	**4.6**	**4.4**	4.6	4.5	4.5

Table 3: Human evaluation results on three datasets. For a fair comparison, our model does not use data augmentation. For the LLMs-based methods, we selected P&R$_{4\text{-shot}}$ with GPT-J-6B as the baseline. Krippendorff's alpha of human rating on three datasets is 0.77, 0.74, and 0.78, respectively, indicating acceptable inter-annotator agreement.

more fluent. We also performed a human evaluation to evaluate transferred sentences on three aspects: **Style** (style transfer accuracy, **Content** (content preservation), and **Fluency** (see Sec. 5.2 for more details).

Baselines As shown in Table 1, our baselines cover different approaches: explicit disentanglement, implicit disentanglement, and without disentanglement (including LLMs-based methods). For a fair comparison, we select Augmented zero-shot learning (**AugZS**) (Reif et al., 2021), Prompt and Rerank (**P&R**) (Suzgun et al., 2022) in different few-shot settings as baselines of LLMs-based methods.

5.2. Main Results

Automatic Evaluation Results. From Tables 1 and 2, it is evident that our model consistently outperforms other baselines across all three datasets, even without data augmentation. Specifically, MaskAndInfill (Wu et al., 2019) achieves the highest accuracy on the Yelp dataset, while MaskTransformer (Wu et al., 2020) performs best on IMDb. However, both models exhibit poor content preservation and low GM and HM scores. It highlights the challenge of effectively balancing accuracy and content preservation in explicit disentanglement methods. On the other hand, Cross-alignment (Shen et al., 2017) and Disentangled (John et al., 2018) consistently obtain the lowest PPL scores, indicating their struggles in generating meaningful tokens and preserving content, thus leading to inferior GM and HM scores. It demonstrates the limitations of incomplete disentanglement approaches in achieving successful style transfer. In contrast, our model significantly improves GM and HM scores, showcasing its proficiency in style transfer and preserving content. Moreover, the substantial increases in s-sBLEU and r-sBLEU further validate StyleFlow's expertise in content preservation. Our model also achieves highly competitive results in terms of accuracy and fluency. Furthermore, we observe further improvements in most metrics by employing data augmentation techniques (see Sec. 5.3 for more details).

Human Evaluation Results. We selected five representative methods from different categories for human evaluation, as shown in Table 3. We randomly selected 100 outputs from five baselines for each dataset and our approach without data augmentation. It can be seen that human evaluation results are consistent with automatic evaluation results, and our model performs significantly better than other methods in style transfer accuracy and content preservation, as well as fluency of the target sentence.

Model	ACC	s-sBLEU	r-sBLEU	GM	HM	PPL
StyleFlow	93.7	61.2	28.6	51.8	43.8	42
(-) reversible	87.5	52.9	25.4	47.1	39.4	78
(-) attention	89.6	50.1	23.2	45.6	36.9	95
(-) $\mathcal{L}_{\text{self}}$	51.3	0.4	0.4	4.5	0.8	N/A
(-) $\mathcal{L}_{\text{cycle}}$	89.4	51.5	24.9	47.2	39	86
(-) $\mathcal{L}_{\text{c}}$	90.2	53.3	25.8	48.2	40.1	97
(-) $\mathcal{L}_{\text{s}}$	3.4	100	22.7	8.8	5.9	18

Table 4: Ablation study of the proposed model on Yelp datasets, where (-) indicates removing the corresponding component from the model or the loss terms in the objective functions, and N/A is a tremendous value. "Reversible" means the reversible encoder, and "attention" refers to attention-aware coupling layers. StyleFlow does not use data augmentation.

5.3. Detailed Analysis

Ablation Study To delve into the influence of various components and loss functions on the overall performance, we conducted an ablation study in Table 4. Our observations indicate a notable decline in performance concerning content preservation (evidenced by an 11.1-point decrease in s-sBLEU and a 5.4-point drop in r-sBLEU) when substituting the attention-aware coupling layer with its vanilla counterpart, which segregates inputs without considering style attention scores. This setback stems from the model's impaired capability to disentangle content and style vectors, a consequence of omitting the attention mechanism's guidance.

Turning to the loss functions, the empirical data reaffirms the critical role $\mathcal{L}_{cycle}$ holds in fortifying content preservation. To further demonstrate our model's competency in reducing $\mathcal{L}_{cycle}$, we show the fluctuations of $\mathcal{L}_{cycle}$ for both StyleFlow and StyleTransformer during various training and testing phases in Fig.2. Intriguingly, StyleFlow and StyleTransformer demonstrate convergence at approximately 2k iterations in the training phase. However, $\mathcal{L}_{cycle}$ of StyleFlow remarkably tapers to an absolute zero, a feat attributable to the integration of the reversible encoder. This trend persists in the testing phase, with $\mathcal{L}_{cycle}$ of StyleFlow converging at a markedly lower threshold than StyleTransformer. These findings compellingly attest to StyleFlow's superior capacity in curtailing $\mathcal{L}_{cycle}$.

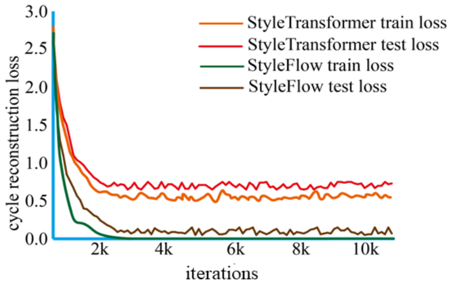

Figure 2: Cycle reconstruction loss of different models.

Comparison of Different Data Augmentation Methods Table 5 shows the effects of different data augmentation methods. We selected three data augmentation methods for comparison and results show that these three methods showed no significant improvement: (1) word-level mix-up interpolates word embeddings without considering stylistic words carrying both content and style information (John et al., 2018; Lee et al., 2021); (2) sentence-level mix-up interpolates the sentence-level representations (Kwon and Lee, 2022; Zheng et al., 2023) without considering semantic information leading to a large gap between pseudo samples and original samples; (3) latent perturbation adds noise perturbations on the latent representations from VAE (Yoo et al., 2018) but the randomness of perturbations makes the pseudo samples of low quality (Yüksel et al., 2021). Our data augmentation method based on normalizing flow has significantly improved the model's performance because it generates more diverse pseudo samples and enhances content preservation through adversarial training.

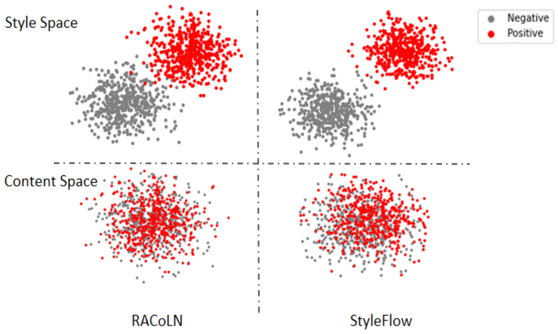

Figure 3: Visualizations with t-SNE on Yelp.

Comparison of Disentanglement of Style and Content Space We utilize t-SNE for visualizing sentences from Yelp, projected into content and style spaces by both RACoLN (Lee et al., 2021) and StyleFlow, as depicted in Fig. 3. Compared to RACoLN, StyleFlow distinguishes sentences of different sentiments more clearly in the style space. This discriminative distribution underscores the robust style transfer capabilities of the attention-aware coupling layer. Furthermore, the distribution of StyleFlow in the content space is more concentrated, indicating its superior content preservation abilities. Consequently, these visual results intuitively demonstrate StyleFlow's enhanced capacity for disentangling style and content.

	ACC	s-sBLEU	r-sBLEU	GM	HM	PPL
StyleFlow w/o DA	93.7	61.2	28.6	51.8	43.8	**42**
+ word-level mix-up	93.1	60.8	27.7	50.8	42.7	52
+ sentence-level mix-up	94	61.6	28.7	51.9	44	59
+ latent perturbation	94.2	61.4	28.9	52.2	44.2	55
+ DA-normalizing flow	**95.1**	**63.8**	**29.2**	**52.7**	**44.7**	49

Table 5: Results of data augmentation methods on Yelp. We augmented the data by doubling the amount of available data.

6. Conclusion

This paper proposes a new disentanglement-based model that leverages a reversible encoder to improve content preservation in unsupervised text style transfer. We design attention-aware coupling layers to gradually disentangle content and style information for a better cycle reconstruction, which can sharply reduce the cycle loss and preserve the content information. Moreover, we augment training data using data augmentation, which adds perturbation to latent space based on normalizing flow. Results from sentiment and formality transfer experiments show that StyleFlow outperforms several strong baselines, achieving better content preservation and accuracy with a lower cycle reconstruction loss.

7. Acknowledgments

This work is supported by the following fundings: Young Elite Scientist Sponsorship Program by CAST (2023QNRC001) under Grant No. YESS20230367, the National Natural Science Foundation of China under Grant No. 62306330, No. 62106275, Natural Science Foundation of Hunan Province under Grant No. 2022JJ40558, and the Grant of No. WDZC20235250103. We would like to thank Weiguo Chen, Zixuan Dong, Zhihua Wen and Yuxuan Gu for their helpful discussions. Moreover, we thank anonymous reviewers for their valuable suggestions.

8. Ethical Considerations

There is excellent potential for style transfer to be a powerful tool for editing and regulating text content. However, this technology has ethical implications that must be considered. It is crucial to recognize the potential for misuse and abuse of text style transfer when using it for specific styles from text, such as hate speech or offensive language. Text style transfer should not be used to propagate hate or harmful messages. Furthermore, text style transfer can be beneficial for reducing hate and promoting more inclusive and positive messages. As researchers and users of text style transfer, we are responsible for considering these ethical issues and using this technology responsibly and respectfully.

9. Limitations

Some of the limitations of StyleFlow are the following. **First**, although normalizing flow has an advantage in the interpretability of the neural networks (Rezende and Mohamed, 2015; Ho et al., 2019; Liu et al., 2022), we pay less attention to use it to improve the interpretability of our model, which will be explored in depth in future work. **Second**, there are only two different types of styles in our experiment setting, such as sentiment and formality. However, there are often more than two styles in the natural language environment. The multiple styles transfer will be feasible in future works. **Thirdly**, our model only applies to a single language and will be extended to multilingual text style transfer tasks in future work.

10. Bibliographical References

Abdelrahman Abdelhamed, Marcus A Brubaker, and Michael S Brown. 2019. Noise flow: Noise modeling with conditional normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3165–3173.

Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. 2021. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 862–871.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xinyu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. *arXiv preprint arXiv:1907.05789*.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Eleftheria Briakou, Sweta Agrawal, Ke Zhang, Joel Tetreault, and Marine Carpuat. 2021a. A review of human evaluation for style transfer. *arXiv preprint arXiv:2106.04747*.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021b. Xformal: A benchmark for multilingual formality style transfer. *arXiv preprint arXiv:2104.04108*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. *arXiv preprint arXiv:2005.00701*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuan-Jing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shizhe Diao, Yongyu Lei, Liangming Pan, Tianqing Fang, Wangchunshu Zhou, Sedrick Keh, Min-Yen Kan, and Tong Zhang. 2023. Doolittle: Benchmarks and corpora for academic writing formalization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13093–13111.

Laurent Dinh, David Krueger, and Yoshua Bengio. 2014. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Cicero dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. 2019. Neural spline flows. *Advances in neural information processing systems*, 32.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*.

Yao Fu, Hao Zhou, Jiaze Chen, and Lei Li. 2019. Rethinking text attribute transfer: A lexical analysis. *arXiv preprint arXiv:1909.12335*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. *arXiv preprint arXiv:1903.10671*.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. 2019. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, pages 2722–2730. PMLR.

Emiel Hoogeboom, Victor Garcia Satorras, Jakub Tomczak, and Max Welling. 2020. The convolution exponential and generalized sylvester flows. *Advances in Neural Information Processing Systems*, 33:18249–18260.

Nabil Hossain, Marjan Ghazvininejad, and Luke Zettlemoyer. 2020. Simple and effective retrieve-edit-rerank text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2532–2538.

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

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International* *conference on machine learning*, pages 1587–1596. PMLR.

Yufang Huang, Wentao Zhu, Deyi Xiong, Yiye Zhang, Changjian Hu, and Feiyu Xu. 2020. Cycle-consistent adversarial autoencoders for unsupervised text style transfer. *arXiv preprint arXiv:2010.00735*.

Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson. 2020. Semi-supervised learning with normalizing flows. In *International Conference on Machine Learning*, pages 4615–4630. PMLR.

Philip TG Jackson, Amir Atapour Abarghouei, Stephen Bonner, Toby P Breckon, and Boguslaw Obara. 2019. Style augmentation: data augmentation via style randomization. In *CVPR workshops*, volume 6, pages 10–11.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. *arXiv preprint arXiv:2004.01980*.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. Imat: Unsupervised text attribute transfer via iterative matching and translation. *arXiv preprint arXiv:1901.11333*.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2018. Disentangled representation learning for non-parallel text style transfer. *arXiv preprint arXiv:1808.04339*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29.

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. 2020. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. *arXiv preprint arXiv:2010.05700*.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. *arXiv preprint arXiv:2006.09639*.

Soonki Kwon and Younghoon Lee. 2022. Explainability-based mix-up approach for text data augmentation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you bart! rewarding pre-trained models improves formality style transfer. *arXiv preprint arXiv:2105.06947*.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Dongkyu Lee, Zhiliang Tian, Lanqing Xue, and Nevin L Zhang. 2021. Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 93–102.

Joosung Lee. 2020. Stable style transformer: Delete and generate approach with encoder-decoder for text style transfer. *arXiv preprint arXiv:2005.12086*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2019. Domain adaptive text style transfer. *arXiv preprint arXiv:1908.09395*.

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

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2201.05273*.

Yihong Liu, Haris Jabbar, and Hinrich Schütze. 2022. Flow-adapter architecture for unsupervised machine translation. *arXiv preprint arXiv:2204.12225*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. *Advances in Neural Information Processing Systems*, 31.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. *arXiv preprint arXiv:1905.10060*.

Guoqing Luo, Yu Tong Han, Lili Mou, and Mauajama Firdaus. 2023. Prompt-based editing for text style transfer. *arXiv preprint arXiv:2301.11997*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. Powertransformer: Unsupervised controllable revision for biased language correction. *arXiv preprint arXiv:2010.13816*.

Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. 2019. Flowseq: Non-autoregressive conditional sequence generation with generative flow. *arXiv preprint arXiv:1909.02480*.

Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. Mix and match: Learning-free controllable text generation using energy language models. *arXiv preprint arXiv:2203.13299*.

Sandeep Nagar, Marius Dufraisse, and Girish Varma. 2021. Cinc flow: Characterizable invertible 3x3 convolution. *arXiv preprint arXiv:2107.01358*.

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

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Machel Reid and Victor Zhong. 2021. Lewis: Levenshtein editing for unsupervised text style transfer. *arXiv preprint arXiv:2105.08206*.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2021. A recipe for arbitrary text style transfer with large language models. *arXiv preprint arXiv:2109.03910*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Danilo Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR.

Bidisha Samanta, Mohit Agrawal, and Niloy Ganguly. 2021. A hierarchical vae for calibrating attributes while generating text using normalizing flow. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2405–2415.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. *arXiv preprint arXiv:1805.07685*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*.

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. *arXiv preprint arXiv:1909.11493*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. Transforming delete, retrieve, generate approach for controlled text style transfer. *arXiv preprint arXiv:1908.09368*.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip S Yu, and Lifang He. 2020. Mixup-transformer: dynamic data augmentation for nlp tasks. *arXiv preprint arXiv:2010.02394*.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. *arXiv preprint arXiv:2205.11503*.

Zineng Tang, Shiyue Zhang, Hyounghun Kim, and Mohit Bansal. 2021. Continuous language generative flow. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4609–4622.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Enrica Troiano, Aswathy Velutharambath, and Roman Klinger. 2023. From theories on styles to their transfer in text: Bridging the gap with a hierarchical survey. *Natural Language Engineering*, 29(4):849–908.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12).

Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation. *Advances in Neural Information Processing Systems*, 32.

Zhihua Wen, Zhiliang Tian, Zhen Huang, Yuxin Yang, Zexin Jian, Changjian Wang, and Dongsheng Li. 2023. Grace: Gradient-guided controllable retrieval for augmenting attribute-based text generation. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8377–8398.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420.

Chunhua Wu, Xiaolong Chen, and Xingbiao Li. 2020. Mask transformer: Unpaired text style transfer based on masked language. *Applied Sciences*, 10(18):6196.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. " mask and infill": Applying masked language model to sentiment transfer. *arXiv preprint arXiv:1908.08039*.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970.

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

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. 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)*, pages 979–988.

Min Yang, Zhou Zhao, Wei Zhao, Xiaojun Chen, Jia Zhu, Lianqiang Zhou, and Zigang Cao. 2017. Personalized response generation via domain adaptation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1021–1024.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. *Advances in Neural Information Processing Systems*, 31.

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2021. 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*, pages 3801–3807.

Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. 2018. Data augmentation for spoken language understanding via joint variational generation. *arXiv preprint arXiv:1809.02305*.

Oguz Kaan Yüksel, Sebastian U Stich, Martin Jaggi, and Tatjana Chavdarova. 2021. Semantic perturbations with normalizing flows for improved generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6619–6629.

Jiaao Zhan, Yang Gao, Yu Bai, and Qianhui Liu. 2022. Stage-wise stylistic headline generation: Style generation and summarized content insertion. In *International Joint Conference on Artificial Intelligence*. IJCAI.

Yi Zhang, Jingjing Xu, Pengcheng Yang, and Xu Sun. 2018. Learning sentiment memories for sentiment modification without parallel data. *arXiv preprint arXiv:1808.07311*.

Yuekai Zhao, Shuchang Zhou, and Zhihua Zhang. 2021. Multi-split reversible transformers can enhance neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 244–254.

Haoqi Zheng, Qihuang Zhong, Liang Ding, Zhiliang Tian, Xin Niu, Changjian Wang, Dongsheng Li, and Dacheng Tao. 2023. Self-evolution learning for mixup: Enhance data augmentation on few-shot text classification tasks. *abs/2305.13547*, pages 8964–8974.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.