Rethinking Style Transformer with Energy-based Interpretation: Adversarial Unsupervised Style Transfer using a Pretrained Model

Hojun Cho[1], **Dohee Kim**[1], **Seungwoo Ryu**[1], **ChaeHun Park**[1], **Hyungjong Noh**[2], **Jeong-in Hwang**[2], **Minseok Choi**[1], **Edward Choi**[1], and **Jaegul Choo**[1]
[1]KAIST AI [2]NCSOFT Corporation
{hojun.cho,dohee1121,swryu,ddehun,minseok.choi,edwardchoi,jchoo}@kaist.ac.kr
{nohhj0209,jihwang}@ncsoft.com

Abstract

Style control, content preservation, and fluency determine the quality of text style transfer models. To train on a nonparallel corpus, several existing approaches aim to deceive the style discriminator with an adversarial loss. However, adversarial training significantly degrades fluency compared to the other two metrics. In this work, we explain this phenomenon using energy-based interpretation, and leverage a pretrained language model to improve fluency. Specifically, we propose a novel approach which applies the pretrained language model to the text style transfer framework by restructuring the discriminator and the model itself, allowing the generator and the discriminator to also take advantage of the power of the pretrained model. We evaluated our model on three public benchmarks GYAFC, Amazon, and Yelp and achieved state-of-the-art performance on the overall metrics.

1 Introduction

Text style transfer is the task of converting a sentence from one style to another while preserving style-agnostic semantics. In solving the text style transfer task, three criteria must be considered: 1) *style control*, how well a style is transferred from the original sentence to the generated one, 2) *content preservation*, how well the generated sentence has retained the semantics of the original, and 3) *fluency*, how natural the generated sentence is.

Text style transfer is challenging, since fluently converting the style of a sentence often conflicts with content preservation (Prabhumoye et al., 2018; John et al., 2019; Gong et al., 2019). To address this challenge, several supervised text style transfer methods have been attempted (Jhamtani et al., 2017; Al Nahas et al., 2019; Lai et al., 2021b). But style-labeled sentence pairs are often not available, making that approach less practical in a real-world setting. Various unsupervised text style transfer approaches have become popular, including those using an autoencoder (Hu et al., 2017; Huang et al., 2020), back-translation (Prabhumoye et al., 2018; Lample et al., 2019), and reinforcement learning (Xu et al., 2018; Luo et al., 2019). Among previous studies, Style Transformer (Dai et al., 2019) achieved fine-grained style control by deceiving the style discriminator through adversarial training. Aside from their strengths, however, adversarial models including Style Transformer degrades the fluency of generated sentences.

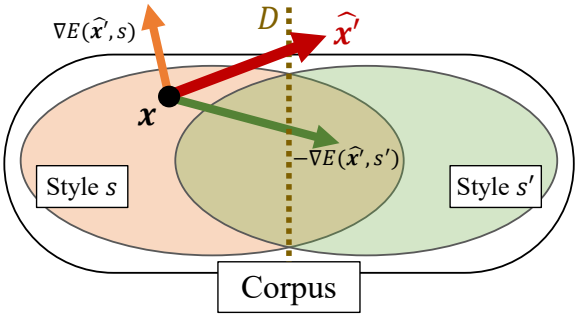

Figure 1: Energy-based interpretation for fluency degradation. Deceiving energy-based discriminator D requires 1) minimizing the energy E between the transferred sentence $\widehat{\mathbf{x}}'$ and the target style s', and 2) maximizing the energy between the sentence $\widehat{\mathbf{x}}'$ and the original style s. However, the style s and s' are originated from the same corpus, so maximizing $E(\widehat{\mathbf{x}}', s)$ degrades the overall fluency. It is an interpretation of Eq. 6.

In this paper, we review Style Transformer to investigate the reason behind the fluency degradation in adversarial models. To more precisely interpret what fluency is, we introduce the notion of *energy* (Hinton, 2002; Lecun et al., 2006), which is the entropy of variables. The energy function, which measures the energy of input variables with respect to a particular style, outputs low energy if the inputs are common in that style and outputs high energy otherwise. For example, formal/informal sentences would likely have low energy in the formality corpus, while sentences that

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5452 - 5467
December 7-11, 2022

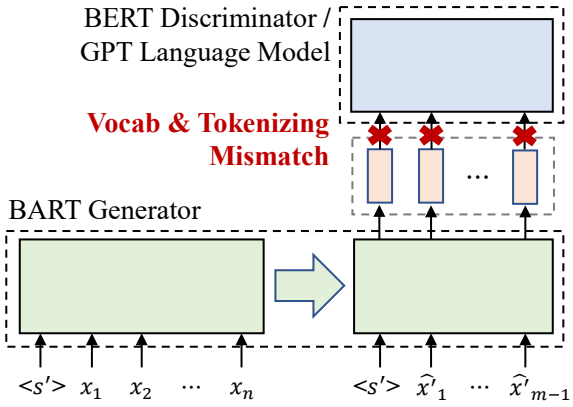

Figure 2: A structural dilemma when applying pretrained models to adversarial learning. To propagate gradients to generated tokens, the generator, discriminator, and LM should have the same vocabulary and tokenizer. However, publicly available pretrained models (e.g., BERT and GPT) use their own tokenizers, so the discriminator and LM need to be trained from scratch if we apply the pretrained model to the generator.

have nothing to do with formality (e.g., political expressions) would have high energy in the corpus. Accordingly, we define *fluency* as having low energy in a particular corpus, in which the fluent sentences express one of the styles in the corpus. As illustrated in Figure 1, fluency degrades while deceiving the discriminator, since adversarial learning maximizes the energy to the source style and drives the generated sentence far away from the distribution of the corpus. To counter fluency degradation, we introduce a regularizer using a language model (LM) to keep the generated sentences in the distribution of the corpus. This LM-based regularizer keeps the generated sentences in the corpus by pulling the sentence to the target corpus.

To apply the LM-based regularizers, we can leverage pretrained models such as GPT-2 to generate fluent sentences. Moreover, fluency is expected to further improve when the generator and the discriminator are also replaced with a pretrained model. However, as shown in Figure 2, the generator, discriminator, and LM must share the same vocabulary and tokenizer in order to propagate gradients successfully. Thus, inefficiency can arise, in that two of the three modules may need to be re-trained from scratch because the existing pretrained models are based on different tokenizers. To address this issue, we restructured the discriminator and LM such that a single pretrained model is applied to all three modules: the generator, discriminator, and LM. By using a single pretrained model, our method not only solves the inconsistent vocabulary problem but also has advantages when applying further pretraining for domain adaptation (Gururangan et al., 2020) or additional dataset (Lai et al., 2021a). This is because additional pretraining is necessary when only a single model is used to improve style transfer performance.

Our contributions can be summarized as follows:

- We analyze the fluency degradation in adversarial training with an energy-based interpretation, and propose a regularizer leveraging a language model to prevent fluency degradation.
- We reconstruct the discriminator and language model such that the single pretrained language model can be employed in the text style transfer framework.
- We achieve new state-of-the-art results on GYAFC, Amazon, and Yelp datasets and carefully analyze the contribution of each component of our model.

2 Related Work

2.1 Unsupervised style transfer

Many of the previous studies have attempted to learn disentangled representations of text by separating representations of content and style in a latent space. For instance, Shen et al. (2017) trained a cross-aligned autoencoder to learn a shared latent space for contents, while learning a separate representation for styles using adversarial learning. Yang et al. (2018) further extended this cross-aligned approach by leveraging LM as a discriminator to enhance the informativeness and stability of adversarial training. Yi et al. (2020) leveraged multiple instances of the same style to model the latent space of underlying stylistic characteristics, and samples extracted from this space were fed into the decoder to balance with contents. These works using disentangled representations exhibited reasonable performance with high interpretability, but disentangled content representations can still contain style-relevant information, as pointed out by Lample et al. (2019). In addition, there is a limitation in that the meaning of the input sentence must be expressed in a fixed-size vector with a limited capacity (Dai et al., 2019).

In contrast, there are methods without disentangled representations that do not explicitly disentan-

gle the content and style of text using reinforcement learning (Xu et al., 2018; Luo et al., 2019) and back-translation (Lample et al., 2019; Prabhumoye et al., 2018). Dai et al. (2019) proposed a novel style transfer model based on the transformer architecture without disentangled representations, and Wang et al. (2019) also utilized a transformer for an unsupervised framework by editing entangled latent representations. These models are novel in their model architecture or training strategy, but they do not utilize pretraining models, which results in lower performance than the state-of-the-art methods. On the other hand, our work proposes a novel approach to effectively leverage a pretrained LM in an unsupervised text style transfer task.

2.2 Style transfer with pretrained models

Recently, pretrained models have achieved great success on various NLP tasks such as machine translation (Chronopoulou et al., 2020; CONNEAU and Lample, 2019) and text summarization (Liu and Lapata, 2019). The pretrained models are also being used for text style transfer tasks. Sudhakar et al. (2019) used two variants of 'decoder-only' transformer to generate sentences in a target style and leveraged the power of GPT (Radford et al., 2018). Malmi et al. (2020) used a padded masked language model (Mallinson et al., 2020) variant, whose architecture and the trained corpus were identical to those of BERT (Devlin et al., 2019). Although these studies exploited the power of pretrained models, our approach differs in that we train our model adversarially in an end-to-end manner.

In addition, several works have focused on style transfer in a specific domain, or for leveraging an additional corpus. To transfer writing styles between authors, Syed et al. (2020) pretrained LM on the author corpus from scratch using masked language modeling. Laugier et al. (2021) detoxified toxic texts by fine tuning a pretrained T5 (Raffel et al., 2020) using additional denoising and cycle-consistency objectives. However, these studies only focused on a specific domain. Lai et al. (2021a) built a pseudo-parallel dataset by leveraging generic resources including WordNet (Baccianella et al., 2010) and Parabank (Hu et al., 2019) to fine tune BART on style transfer tasks. Lai et al. (2021b) fine tuned BART (Lewis et al., 2020) using parallel data with a policy gradient (Sutton et al., 1999) which maximized the style classifier reward and the BLEU score reward. Our work incorporates pretrained models with adversarial training on various domains, while trained only on a non-parallel corpus.

2.3 Energy-based model

The conventional probabilistic model outputs the normalized probability $p(x)$ for input variable x. In contrast, the energy-based model outputs the non-normalized scalar value $E(x)$ denoted as *energy* (Hinton, 2002; Lecun et al., 2006) Using the energy-based model, we can classify x by comparing the energy of each label, or generate x by optimizing $\arg\min_x E(x)$.

Several works have leveraged the energy-based model for image generation (Ngiam et al., 2011; Zhao et al., 2017), text generation (Deng et al., 2020; Bakhtin et al., 2021), and reinforcement learning (Haarnoja et al., 2017). We borrow the main idea of the energy-based model, which expresses the classifier in the form of an energy function. In the work of Che et al. (2020), the authors interpreted the GAN discriminator (Goodfellow et al., 2014) using the energy-based model, but we apply this interpretation to the style transfer task. We show that Style Transformer can be interpreted as an energy-based model by decomposing the discriminator, and provide the reason why fluency degradation occurs when we try to deceive the style discriminator.

3 Method

In an unsupervised setting, we assume the non-parallel corpus $\mathbf{X} = \{\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}\}$ and $\mathbf{X}' = \{\mathbf{x}'^{(0)}, \mathbf{x}'^{(1)}, \cdots, \mathbf{x}'^{(n)}\}$, and denote each style of the corpus as s and s'. The objective is to train a style transfer model G in an unsupervised way such that a sentence $\mathbf{x}$ is turned into a sentence $\widehat{\mathbf{x}}'$ having similar content but the other style.

3.1 Preliminaries

Style Transformer Dai et al. (2019) proposed an unsupervised style transfer model based on the transformer architecture (Vaswani et al., 2017). In their work, the self loss $\mathcal{L}_{\text{self}}$ and cycle loss $\mathcal{L}_{\text{cycle}}$ were used to preserve content, while the style loss $\mathcal{L}_{\text{style}}$ was employed to control style. We let the generator G take the source sentence $\mathbf{x}$ and the style s. If we transfer the sentence to its originated style in $\widehat{\mathbf{x}} \sim G(\mathbf{x}, s)$, the model should output the same sentence. Targeting this reconstruction, the

self loss is defined as

$$\mathcal{L}_{\text{self}}(\theta_G) = -\mathbb{E}_{s,\mathbf{x}}\left[\log p\left(G\left(\mathbf{x}, s\right) = \mathbf{x} | \mathbf{x}, s\right)\right] \tag{1}$$

which is the cross entropy between the reconstructed sentence $\widehat{\mathbf{x}}$ and source sentence $\mathbf{x}$.

While transferring the sentence to the target style in $\widehat{\mathbf{x}}' \sim G\left(\mathbf{x}, s'\right)$, the content of the sentence should be preserved. As in previous studies (Logeswaran et al., 2018; Xu et al., 2018), Style Transformer adopts the cycle loss

$$\mathcal{L}_{\text{cycle}}(\theta_G) = -\mathbb{E}_{s,\mathbf{x},\widehat{\mathbf{x}}' \sim G(\mathbf{x},s')}\left[\log p\left(G(\widehat{\mathbf{x}}', s) = \mathbf{x} | \widehat{\mathbf{x}}', s\right)\right] \tag{2}$$

which regularizes the generated sentence so that it is identical to the source sentence when re-transferred to the original style.

For style control, Style Transformer leverages an external model that discriminates the style. The discriminator D judges the consistency between the given sentence $\mathbf{x}$ and style s. The discriminator is trained separately from the generator and takes the generated sentences along with the original sentences. The training process for the discriminator optimizes

$$\mathcal{L}_{\text{disc}}(\theta_D) = -\mathbb{E}_{s,\mathbf{x}}\left[\log D\left(c \mid \mathbf{x}, s\right)\right] \tag{3}$$

where labeling $\{(\mathbf{x}, s), (\widehat{\mathbf{x}}, s)\}$ in positive as $c = 1$, $\{(\mathbf{x}, s'), (\widehat{\mathbf{x}}', s')\}$ in negative as $c = 0$. Style Transformer attempts to deceive this discriminator into classifying the generated sample $(\widehat{\mathbf{x}}', s')$ as $c = 1$:

$$\mathcal{L}_{\text{style}}(\theta_G) = -\mathbb{E}_{s,\mathbf{x},\widehat{\mathbf{x}}' \sim G(\mathbf{x},s')}\left[\log D\left(c = 1 \,\middle|\, \widehat{\mathbf{x}}', s'\right)\right] \tag{4}$$

The upper part of Figure 4 describes how each loss works in our model.

In the cycle and style loss, the gradients should be propagated into the generated sentences, but the nature of language discreteness prevents a trivial solution. To propagate the gradients directly, Style Transformer feeds the generated sentences to the discriminator in the form of a softmax distribution for each token. This soft representation of the sentences empirically reports better performance than REINFORCE (Williams, 1992) and the Gumbel softmax (Jang et al., 2017).

BART Style Transformer follows the transformer encoder-decoder structure and initializes weights by training the dataset in an autoencoding manner. In contrast, we leverage BART (Lewis et al., 2020),

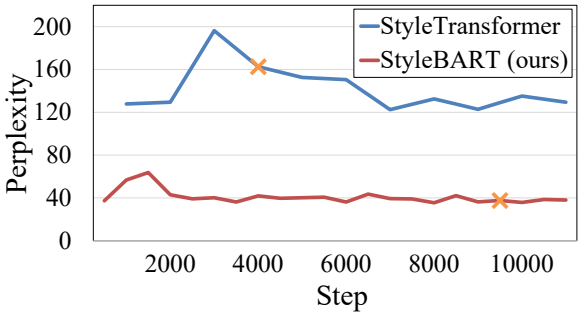

Figure 3: Perplexity change on the dev dataset during training steps in the GYAFC FR dataset. We measure perplexity by GPT-2 fine tuned on the sentences in the target style. Yellow cross-marks indicate the points that report the best style transfer performance $J(\mathrm{A}, \mathrm{S})$. Our model shows a smaller perplexity bump then Style Transformer and maintains the perplexity constant.

a denoising autoencoder for pretraining sequence-to-sequence models, to enhance Style Transformer. BART is pretrained on two tasks: text in-filling and sentence shuffling. The text in-filling task trains the model to predict the masked span from a sentence, and the sentence shuffling task reorders the shuffled sentences in the right order. Both tasks are trained with the denoising autoencoder structure which takes the corrupted sentence $\tilde{\mathbf{x}}$ and predicts the original sentence $\mathbf{x}$ in an objective:

$$\mathcal{L}(\theta) = -\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}\left[\sum_i \log\left(p(x_i \mid \mathbf{x}_{1:i-1}, \tilde{\mathbf{x}})\right)\right] \tag{5}$$

3.2 Energy-based interpretation for fluency degradation

In our preliminary study, there is a significant gap between the perplexity of the corpus in the target style and the generated sentences. Based on the energy-based interpretation (Hinton, 2002; Lecun et al., 2006), we hypothesize that fluency degradation occurs due to the style discriminator. The energy-based model estimates the dependency between the sample $\mathbf{x}$ and the label s, and outputs the scalar value implying the *energy* between them. If the energy is high, the entropy between the sample and label is high, so those are likely to be independent of each other. The energy-based classifier outputs the probability of each label by the ratio between the energy of labels. Based on this interpretation, the style discriminator can be decomposed into

$$D\left(c = 1 \,\middle|\, \widehat{\mathbf{x}}', s'\right) = \frac{\exp\left(-E(\widehat{\mathbf{x}}', s')\right)}{\exp\left(-E(\widehat{\mathbf{x}}', s')\right) + \exp\left(-E(\widehat{\mathbf{x}}', s)\right)} \tag{6}$$

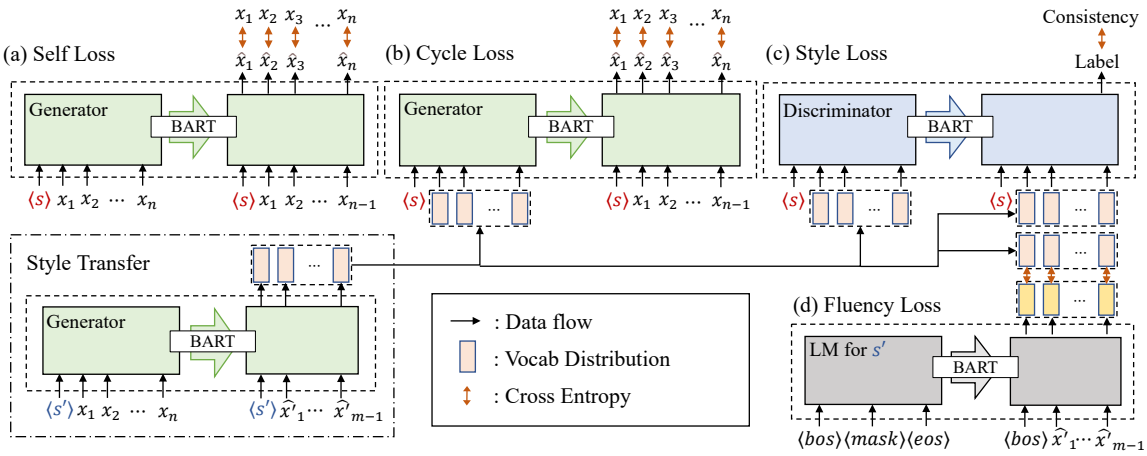

Figure 4: The overall structure of our model. We concatenate the style label in front of the input of the BART encoder and decoder except the LM depicted in (d). For the language model on (d), we feed the mask token in the form of [<bos>, <mask>, <eos>] into the encoder to tackle the problem in a way similar to the text infilling task, and we freeze the encoder while finetuning the decoder to the target corpus. Our model learns how to transfer the style and preserve the content through four mechanisms; the model (a) reconstructs the source sentence by itself, (b) cyclically reconstructs the source sentence from the transferred sentence, (c) deceives a discriminator by generating a sentence in the target style, and (d) improves the fluency of the transferred sentence by using the LM. All modules leverage the BART model and are trained in an end-to-end manner.

which is the exponential ratio of the negative energy E between the transferred sentence $\widehat{\mathbf{x}}'$ and style s or s'. This expression matches the real implementation as the discriminator takes the sentence x and style s as input and outputs of two logits. Each logit value means the negative energy of style s and s', and the discriminator calculates the softmax output between them. To deceive the style discriminator, the generator needs to minimize $E(\widehat{\mathbf{x}}', s')$ while maximizing $E(\widehat{\mathbf{x}}', s)$. Meanwhile, the energy between the sentence and style can be interpreted as the perplexity or entropy of the sentence with the original style in $E(\widehat{\mathbf{x}}', s) \approx \text{PPL}_s(\widehat{\mathbf{x}}')$. Maximizing the perplexity with the original style degrades the fluency of the generated sentences because both styles are from the corpus, sharing syntactic and semantic attributes. Figure 3 depicts this phenomena. At the beginning of training, there is a significant increase in perplexity, and a following perplexity decrease compensates for this initial increase. Thus, perplexity increases by maximizing $E(\widehat{\mathbf{x}}', s)$ in the initial steps and decreases by minimizing $E(\widehat{\mathbf{x}}', s')$ in the final. This initial increase in perplexity affects the perplexity of the final model and harms the LM performance of the pretrained model. If we generalize the discriminator $D(c = 1|x, s)$ to $D(s|x)$, this energy-based interpretation provides a mathematical reason why the adversarial model, which tries to deceive the style classifier, suffers from fluency degradation.

Inspired by the work of Yang et al. (2018), our model leverages LM to prevent the generated sentence from being out of the distribution of the corpus. As the discriminator pushes out the sentence from the distribution, we require additional power to pull it back into the corpus. Thus, we introduce a fluency loss $\mathcal{L}_{\text{fluent}}$, which pulls the generated sentence into the target distribution. For each style s, we train LM by

$$\mathcal{L}_{\text{LM}}(\theta_{\text{LM}_s}) = -\,\mathbb{E}_{\mathbf{x}}\left[\sum_i \log p_{\text{LM}_s}(x_i; \mathbf{x}_{1:i-1})\right] \tag{7}$$

in advance, and optimize the cross entropy of the generated sentence during training along with other losses as

$$\mathcal{L}_{\text{fluency}}(\theta_G) = -\sum_i p^i_G(\widehat{\mathbf{x}}'; \mathbf{x}, s') \log p^i_{\text{LM}_{s'}}(\widehat{\mathbf{x}}') \tag{8}$$

where $p^i_G(\widehat{\mathbf{x}}'; \mathbf{x}, s') = p_G(\widehat{x'}_i; \widehat{\mathbf{x}}'_{1:i-1}, \mathbf{x}, s')$ and $p^i_{\text{LM}_{s'}}(\widehat{\mathbf{x}}') = p_{\text{LM}_{s'}}(\widehat{x'}_i; \widehat{\mathbf{x}}'_{1:i-1})$. We report and analyze the fluency enhancement with this loss in Section 4.7. Finally, the total loss of our model is

$$\begin{aligned}\mathcal{L}(\theta_G) = {} & \lambda_{\text{self}}\mathcal{L}_{\text{self}}(\theta_G) + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}}(\theta_G) \\ & + \lambda_{\text{style}}\mathcal{L}_{\text{style}}(\theta_G) + \lambda_{\text{fluency}}\mathcal{L}_{\text{fluency}}(\theta_G)\end{aligned} \tag{9}$$

where each λ implies the coefficient for each loss.

3.3 Considering the structural dilemma toward adversarial training

For fluent generation, it is desirable to apply a pretrained model (Radford et al., 2019; Brown et al., 2020) to the regularizer. For fluent style control, we applied the pretrained model not only for LM, but also to the generator and discriminator. Since the Style Transformer uses the Transformer encoder-decoder structure, we can readily apply BART to the generator, but there is an architectural problem for the style discriminator and LM. When training the Style Transformer, the discriminator takes the softmax distribution of the generated sentences, and thus the discriminator needs to share the same vocabulary as the generator. This problem is not limited to just the Style Transformer but also expands to the model, requiring gradient backpropagation on the token level using the Gumbel softmax (Jang et al., 2017). As the discriminator in Style Transformer adopts the transformer encoder structure, BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) is the most feasible option, and leveraging GPT-2 (Radford et al., 2019) is most appropriate for LM. However, there is no publicly available BERT, GPT-2 model with BART vocab. This requires training BERT or GPT-2 from scratch, which takes a lot of resources.

There is, however, a rather simple solution to this problem of mismatching tokenizers: We used the same pretrained model for the generator, discriminator, and LM. In this way, we leverage the BART classifier proposed in the original BART paper. The BART classifier takes the same sequence **x** in the encoder and decoder, and predicts the class label at the <eos> token position at the decoder. For the LM, we adopt BART again to share the same vocab and tokenizer, and also take advantage of the BART decoder, which works as the language model in the text infilling task (Lewis et al., 2020). After fine tuning separate LMs for both styles, we leverage them to enhance the fluency of the generated sentences by Eq. (8). Figure 4 describes how we adopt BART for the discriminator and LM. Since the generator, discriminator, and LM share the same BART vocab, the softmax distribution on the vocab could be transferred in an end-to-end manner. In addition, using such a single pretrained model has advantages over using RoBERTa, which has the same vocabulary as BART, because further pretraining for the domain adaptation (Gururangan et al., 2020) or additional datasets (Lai et al., 2021a) can be done by training one model, and this saves a lot of training resources.

Algorithm 1: Training Procedure

Data: Non-parallel corpus $\mathbf{X}, \mathbf{X}'$ with style s, s'
Initialize the discriminator θ_D, generator θ_G, and language models $\theta_{\text{LM}_s}, \theta_{\text{LM}_{s'}}$ from the BART weights;
Retrain θ_G by Eq. (1);
Retrain θ_D by Eq. (3);
Retrain θ_{LM_s} and $\theta_{\text{LM}_{s'}}$ by Eq. (7);
repeat
  for n_D *step* **do**
    Sample minibatches $\mathbf{x}_i \sim \mathbf{X}, \mathbf{x}'_j \sim \mathbf{X}'$;
    $\widehat{\mathbf{x}}_i = G(\mathbf{x}_i, s), \widehat{\mathbf{x}}'_i = G(\mathbf{x}_i, s')$;
    $\widehat{\mathbf{x}}'_j = G(\mathbf{x}'_j, s'), \widehat{\mathbf{x}}_j = G(\mathbf{x}'_j, s)$;
    Label $\{(\mathbf{x}_i, s), (\widehat{\mathbf{x}}_i, s), (\mathbf{x}'_j, s'), (\widehat{\mathbf{x}}'_j, s')\}$ as 1, $\{(\mathbf{x}_i, s'), (\widehat{\mathbf{x}}'_i, s'), (\mathbf{x}'_j, s), (\widehat{\mathbf{x}}_j, s)\}$ as 0;
    Optimize θ_D by Eq. (3)
  end
  for n_G *step* **do**
    Sample minibatches $\mathbf{x}_i \sim \mathbf{X}, \mathbf{x}'_j \sim \mathbf{X}'$;
    $\widehat{\mathbf{x}}_i = G(\mathbf{x}_i, s), \widehat{\mathbf{x}}'_j = G(\mathbf{x}'_j, s')$;
    Compute $\mathcal{L}_{\text{self}}(\theta_G)$ by Eq. (1);
    $\widehat{\mathbf{x}}'_i = G(\mathbf{x}_i, s'), \widehat{\mathbf{x}}_j = G(\mathbf{x}'_j, s)$;
    Compute $\mathcal{L}_{\text{cycle}}(\theta_G)$ by Eq. (2);
    Compute $\mathcal{L}_{\text{style}}(\theta_G)$ by Eq. (4);
    Compute $\mathcal{L}_{\text{fluency}}(\theta_G)$ by Eq. (8);
    Optimize θ_G by Eq. (9)
  end
until;

Algorithm 1 describes the entire training procedure of our method. From the BART weights, we retrain the generator, discriminator, and language models with the source sentences and labels using the training corpus. During the main training procedure, the language models are frozen, and the generator and the discriminator are fine tuned again in an end-to-end manner. The training procedure is similar to the training strategy of GAN (Goodfellow et al., 2014), in that we train the discriminator several times while the generator takes one step. For all datasets, we use $\lambda_{\text{self}} = 0.1, \lambda_{\text{cycle}} = 0.25, \lambda_{\text{style}} = 1.0, \lambda_{\text{fluency}} = 0.05$, and $n_D = 2, n_G = 1$. The other details on the architecture and training procedure are available in the Appendix A.

4 Experiments

4.1 Datasets

For the experiments, we used three widely-used English datasets (Shen et al., 2017; Wang et al., 2019; Lai et al., 2021b): Grammarly's Yahoo Answers Formality Corpus (GYAFC), Amazon and

Approach	Entertainment & Music					Family & Relationships				
	Acc.	*ref*-Sim.	$J(A,S)$	CoLA	PPL ↓	Acc.	*ref*-Sim.	$J(A,S)$	CoLA	PPL ↓
Source Copy	4.2	82.4	3.8	88.0	79	5.4	81.0	4.8	90.5	52
Human Ref.	87.9	-	-	89.8	54	90.2	-	-	92.4	36
CrossAlign (Shen et al., 2017)	34.7	48.6	15.8	42.1	386	33.1	58.0	17.8	50.1	253
ST (Dai et al., 2019)	63.4	74.5	46.2	45.6	233	60.6	71.7	42.3	45.4	167
Masker (Malmi et al., 2020)	28.0	79.5	22.8	76.6	100	26.0	**77.0**	20.6	78.9	63
StyIns (Yi et al., 2020)	69.5	75.0	51.0	53.9	136	72.9	74.7	53.7	57.8	78
Ours	**93.0**	**83.1**	**77.1**	**76.7**	**63**	**95.4**	76.3	**72.9**	**79.8**	**34**

Table 1: Experimental results on Entertainment & Music (EM) and Family & Relationships (FR) set of *GYAFC* dataset. ↓ indicates the smaller the better. Among the methods except for Source Copy and Human Ref., the best result is shown in **bold**.

Approach	Amazon					Yelp				
	Acc.	*ref*-Sim.	$J(A,S)$	CoLA	PPL ↓	Acc.	*self*-Sim.	$J(A,S)$	CoLA	PPL ↓
Source Copy	11.9	76.0	9.0	86.1	40	0.9	-	-	89.6	55
Human Ref.	54.6	-	-	85.3	47	-	-	-	-	-
Target Copy	88.4	-	-	86.1	26	99.1	-	-	89.6	29
CrossAlign (Shen et al., 2017)	16.2	65.1	9.9	69.8	265	72.0	25.3	17.5	24.7	232
ST (Dai et al., 2019)	67.5	47.5	31.0	21.8	528	66.2	71.5	46.1	37.1	164
Masker (Malmi et al., 2020)	27.7	**68.8**	18.6	**72.0**	49	28.2	**90.4**	24.7	**69.6**	76
StyIns (Yi et al., 2020)	44.0	66.6	27.2	63.4	70.4	**86.5**	39.3	33.0	3.2	505
Ours	**71.1**	58.8	**40.1**	66.5	**47**	86.1	77.5	**65.4**	59.7	**56**

Table 2: Experimental results on the *Amazon* and *Yelp* dataset. ↓ indicates the smaller the better. Acc., PPL indicate accuracy, perplexity, respectively. Among the methods except for Source and Target Copy, the best result is shown in **bold**, and the second-highest result is underlined.

Yelp reviews. We used each dataset in raw text to fairly evaluate performance in the wild, and the details of the dataset preprocessing and statistics are explained in Appendix B.

The **GYAFC** dataset (Rao and Tetreault, 2018) was originally a question-and-answer dataset on an online forum, consisting of informal and formal sentences from the two categories: Entertainment & Music (EM) and Family & Relationships (FR). The **Amazon** dataset is a product review dataset, labeled as either a positive or negative sentiment. The **Yelp** dataset[1] is a restaurant and business review dataset with positive and negative sentiments.

4.2 Baselines

We chose the four unsupervised baselines, **CrossAlign** (Shen et al., 2017), Style Transformer (**ST**) (Dai et al., 2019), **Masker** (Malmi et al., 2020), and **StyIns** (Yi et al., 2020), since they are similar to our proposed method. CrossAlign is based on adversarial learning, and Style Transformer is the basis of our model architecture. Masker utilizes the pretraining process of BERT. StyIns is also one of the adversarial learning methods, which learns discriminative and expressive latent style space. Implementation and experiment details for each baseline are described in Appendix B. We report **Source Copy**, **Target Copy** and **Human Ref.** (if available), which evaluate the source, target corpus and hand-crafted reference using the same evaluation metrics, respectively.

4.3 Evaluation metrics

An ideal output is a sentence whose style is transferred to the target style while preserving the original content without losing fluency. Therefore, performance is measured using three criteria: 1) style transfer accuracy, 2) content preservation, and 3) fluency.

Style transfer accuracy This metric indicates how many generated sentences are correctly transferred into the target style. Following Krishna et al. (2020), we leverage RoBERTa (Liu et al., 2019) as a style classifier fine tuned on each corpus. We denote this metric as Acc.

Content preservation We measure the similarity between sentences with a subword embedding model (Wieting et al., 2019) which captures semantic textual similarity (Agirre et al., 2016). When the similarity is measured between generated sentences and original sentences, we denote this metric as *self*-Sim. When hand-crafted reference sentences

[1] https://www.yelp.com/dataset

are available, the similarity between generated sentences and reference sentences is measured, and we denote this metric as *ref*-Sim.

Fluency Following Krishna et al. (2020), we leverage the accuracy from the RoBERTa classifier on the CoLA dataset (Warstadt et al., 2019) as a metric to capture the grammatical correctness. We denote this metric as CoLA. In addition, to evaluate fluency not related to the grammar, we measure the average perplexity (PPL) of the generated sentences using a GPT-2 model (Radford et al., 2019) fine tuned on the target style sentences. We denote this metric as PPL.

Overall metric Following the work of Krishna et al. (2020) which evaluates the overall performance of style transfer, we evaluate the joint measure as

$$J(\mathrm{A}, \mathrm{S}) = \frac{1}{|\mathbf{X}|} \sum_{x \in \mathbf{X}} \mathrm{Acc}(x) \cdot \mathrm{Sim}(x) \qquad (10)$$

for our model and baseline, and we select the model with the highest score in the dev dataset and plot the selected model along with the fluency measure.

4.4 Quantitative Results

Table 1 and 2 show the experimental results for the GYAFC, Amazon, and Yelp datasets.

The perplexity of the source copy is not extremely high when compared to human references, as shown in Table 1. This is because the source and target sentences come from the same corpus and thus share a common topic, such as entertainment or human relationships. Therefore, this numerically proves that the energy for each style is similar, so the text style transfer models should maintain low perplexity while transferring the sentences.

Style Transformer typically shows convincing accuracy, but it reported high perplexity because the generated sentences deviated from the corpus distribution. On the other hand, especially for GYAFC EM & FR, our method reported state-of-the-art performance on the overall metric of style transfer and content preservation, while reporting higher or similar fluency scores than the others.

On the other hand, on the Amazon dataset, our model showed lower content preservation than the baseline. However, the human reference in the Amazon dataset only reports 54.6% style accuracy, and this implies the reference may not represent the target distribution, so the similarity score to the Amazon human reference is questionable. For the Yelp dataset, the similarity score was significantly lower than that of Masker, but it should be noted that this self-similarity is calculated by comparing the original sentence, so the high similarity implies the generator did not make any changes to the sentences. Nevertheless, our model exhibited the best $J(\mathrm{A}, \mathrm{S})$ score for the Amazon and Yelp datasets, which indicates the overall performance of style transfer and content preservation. Therefore, our model performed better on style transfer than other baselines, while maintaining fluency.

Dataset	Model	Fluency	Style	Content
GYAFC EM	ST	4.62	3.85	5.00
	Masker	22.31	13.85	**44.62**
	StyIns	10.77	10.58	10.19
	Ours	**62.12**	**71.54**	<u>40.00</u>
Amazon	ST	4.35	6.09	3.48
	Masker	36.30	**35.43**	**40.87**
	StyIns	21.30	23.26	<u>34.57</u>
	Ours	**37.61**	<u>34.78</u>	20.65

Table 3: Human evaluation on GYAFC EM and Amazon datasets. Each number represents the proportion (%) of being preferred. The best result is shown in **bold** and the second best result is <u>underlined.</u>

4.5 Human Evaluation

To analyze the experimental results qualitatively, we also conduct human evaluation on GYAFC EM and Amazon datasets. As with automatic evaluation, we evaluate three criteria: *fluency*, *style control*, and *content preservation*. We evaluate the GYAFC EM dataset from informal to formal style, and the Amazon dataset from negative to positive style. For each dataset, a total of 20 source sentences are randomly selected. For each source sentence, four sentences are presented as answer options, one from our model and three from each baseline except CrossAlign. The results of the GYAFC EM and Amazon datasets are shown in Table 3.

4.6 Qualitative Results

Table 4 shows several style transfer results for each model on the GYAFC FR and Amazon datasets.[2] Style Transformer generates non-fluent sentences that are grammatically misaligned. Masker outputs source sentences not much different than the original. StyIns fails to change the text style and generates sentences with grammatical errors. Ours performs well on formality text style transfer, in-

[2]More examples are available in Appendix C.

GYAFC Family & Relationships (informal → formal)	
Original	how do i get he to stop nagging me and leave me alone.
ST	~~Yes~~ do ~~am~~ get he to stop nagging me and leave me alone.
Masker	how do I get he to stop nagging me and leave me alone.
StyIns	~~You~~ do ~~not~~ get he to stop ~~asagging~~ me and leave me alone.
Ours	How do I get him to stop nagging me and leave me alone?
Amazon (negative → positive)	
Original	This item does not provide full coverage as the picture suggests.
ST	This item ~~does am scratch~~ full ~~retains~~ as the ~~manual magnets~~.
Masker	This item is not full coverage as the picture suggests.
StyIns	This item does not provide full coverage as the picture suggests.
Ours	This item provides full coverage as the picture suggests.

Table 4: Case Study on *GYAFC Family & Relationships* and *Amazon* dataset. The red and blue words indicate bad and good transfer, respectively. Texts with strikethrough(-) are non-fluent parts of the generated sentences, including grammatical errors.

Approach	Acc.	*ref*-Sim.	J(A,S)	CoLA	PPL ↓
Ours	95.4	76.4	72.9	79.8	34
- $\mathcal{L}_{\text{fluency}}$	90.8	82.4	74.5	75.4	41
- D_{PT}	90.8	74.7	67.5	65.1	44
- $\mathcal{L}_{\text{fluency}}$ - D_{PT}	92.0	74.6	68.7	51.7	54
ST	63.4	74.5	46.2	45.6	167

Table 5: Ablation study on the **GYAFC FR** dataset. -$\mathcal{L}_{\text{fluency}}$ indicates the model without the fluency loss. -D_{PT} indicates the pretrained models were not applied to the discriminator.

cluding capitalization and punctuation transformation, and also on sentiment text style transfer, from negative to positive style. Also, our outputs are grammatically correct and fluent.

4.7 Ablation study

We conducted an ablation study to understand the contribution of each component in our proposed method. The results of the ablation study on the GYAFC FR dataset are shown in Table 5. Without a fluency loss (-$\mathcal{L}_{fluency}$), the fluency in both metrics is degraded even though there is an improvement in the J(A,S) score. This implies that even though the overall performance of style transfer and content preservation may be improved, the model generates unnatural sentences. If we do not leverage the pretrained discriminator and train it from scratch (-D_{PT}), there is degradation in all metrics, and this shows that joint training with the pretrained generator and discriminator helps in all aspects. Lastly, the performance degradation occurs more severely when both components are disabled (-$\mathcal{L}_{fluency}$, -D_{PT}), which means both components are important for fluent sentence style transfer. In addition, the comparison between this model and Style Transformer again confirms the importance of using the pretrained generator.

5 Limitations

As our model uses adversarial learning, training is somewhat unstable, like GAN (Goodfellow et al., 2014; Arjovsky et al., 2017). For this reason, continuing training does not guarantee good results, and the text style transfer performance fluctuates. Although we have paid attention to model selection to compensate for the unstable training, the instability of adversarial learning remains an issue.

In addition, we have only conducted experiments on widely used datasets, to compare our work with previous studies. These datasets are composed of binary style classes, such as positive and negative sentiments. Therefore, conducting experiments using multi-class datasets (Lample et al., 2019) should be considered.

6 Ethical consideration

Our model may generate negative and rude expressions about a specific person or a commercial site because of the data distribution of the Yelp, Amazon, and GYAFC datasets. However, we propose our work in anticipation of positive applicability as shown in previous studies.

7 Conclusion

Using energy-based interpretation, we found that fluency is inevitably degraded when deceiving the discriminator in Style Transformer (Dai et al., 2019). The problem is solved by adding an LM-based regularizer and training the pretrained generator, discriminator and LM together. Our model shows comparable performance in text style transfer and content preservation while preserving fluency, and we demonstrated the robustness of our model by conducting extensive experiments on various styles in raw text.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)), the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2022R1A2B5B02001913 and NRF-2020H1D3A2A03100945), and the NCSOFT Corporation.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Abdullah Al Nahas, Murat Salih Tunali, and Yusuf Sinan Akgul. 2019. Supervised text style transfer using neural machine translation: Converting between old and modern turkish as an example. In *2019 27th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Anton Bakhtin, Yuntian Deng, Sam Gross, Myle Ott, Marc'Aurelio Ranzato, and Arthur Szlam. 2021. Residual energy-based models for text. *Journal of Machine Learning Research*, 22(40):1–41.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tong Che, Ruixiang ZHANG, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yuan Cao, and Yoshua Bengio. 2020. Your gan is secretly an energy-based model and you should use discriminator driven latent sampling. In *Advances in Neural Information Processing Systems*, volume 33, pages 12275–12287. Curran Associates, Inc.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2020. Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2703–2711, Online. Association for Computational Linguistics.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. 2020. Residual energy-based models for text generation. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3168–3180, Minneapolis, Minnesota. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. 2017. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1352–1361. PMLR.

Geoffrey E. Hinton. 2002. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800.

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6521–6528.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 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. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2213–2223, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

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

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021a. Generic resources are what you need: Style transfer tasks without task-specific parallel training data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4241–4254, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021b. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online. Association for Computational Linguistics.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1442–1461, Online. Association for Computational Linguistics.

Yann Lecun, Sumit Chopra, Raia Hadsell, Marc Aurelio Ranzato, and Fu Jie Huang. 2006. *A tutorial on energy-based learning*. MIT Press.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

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. Cite arxiv:1907.11692.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with

attribute controls. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5116–5122. International Joint Conferences on Artificial Intelligence Organization.

Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. FELIX: Flexible text editing through tagging and insertion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1244–1255, Online. Association for Computational Linguistics.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8671–8680, Online. Association for Computational Linguistics.

Jiquan Ngiam, Zhenghao Chen, Pang Wei Koh, and Andrew Ng. 2011. Learning deep energy models. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 1105–1112, New York, NY, USA. ACM.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality 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 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9008–9015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256.

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

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, Melbourne, Australia. Association for Computational Linguistics.

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

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2020. Text style transfer via learning style instance supported latent space. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3801–3807. International Joint Conferences on Artificial Intelligence Organization. Main track.

Junbo Zhao, Michael Mathieu, and Yann LeCun. 2017. Energy-based generative adversarial networks. In *International Conference on Learning Representations*.

A Implementation details

Architecture details Our implementation is based on `bart-base`[3] of the Huggingface's Transformers library (Wolf et al., 2020), which has 140 million parameters in total. On inference time, the next token is decoded in a greedy fashion, and we constrain an n-gram whose n is bigger than three not to be generated again.

[3]Details of model are available in https://huggingface.co/facebook/bart-base.

Training details For model selection, we record the model checkpoint per 500 steps, and the model with the highest $J(\mathrm{A}, \mathrm{S})$ in a single run is selected as our final model. Our model takes about 14 hours on a single NVIDIA RTX A6000 machine to train the *GYAFC* dataset. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a batch size of 64. The initial learning rate of the un-pretrained layers, such as the last linear layer of the discriminator, are set to $2e^{-4}$, and all the others are set to $3e^{-5}$. In addition, a cross entropy of the last linear layer in the discriminator is label-smoothed (Szegedy et al., 2016) with $\alpha = 0.1$.

B Experimental Setup

Previous studies based on the Amazon and Yelp datasets (Shen et al., 2017; Dai et al., 2019) mainly use tokenized datasets in lowercase, which are provided by Shen et al. (2017) and Wang et al. (2019). They are not appropriate for evaluating fluency, because the generated sentences are far away from the raw texts. In addition, since GYAFC (Rao and Tetreault, 2018) datasets are raw texts, it is difficult to compare with existing methods that utilize tokenized datasets. Furthermore, since Masker (Malmi et al., 2020) and our model utilize the pretrained models, they cannot change the tokenizer and do not work on the tokenized datasets. Therefore, we converted all datasets into raw text form and trained our baselines accordingly.

Dataset For the Amazon dataset, we borrowed the preprocessed dataset provided by Wang et al. (2019) and detokenized the dataset to make raw-text sentences. For the Yelp dataset, we created a raw-text dataset following (Shen et al., 2017). Using the raw-text Yelp reviews, only reviews between 10 and 180 in character length were included, and reviews with a rating of 5 were labeled positive, and reviews with ratings of 1 and 2 were labeled negative. For the GYAFC dataset, we use the raw datasets as is.

Baseline For a fair comparison between baselines in raw text, the tokenizer of each model was changed to a subword tokenizer (Gage, 1994; Sennrich et al., 2016). We changed the tokenizers of Style Transformer (Dai et al., 2019) and Styins (Yi et al., 2020) to the BART tokenizer, which is the same as our model. Since our model is based on Style Transformer, we trained Style Transformer using our code while setting the hyperparameters

same to the original paper. For StyIns, we trained the model using the code provided by authors while feeding the text tokenized by the BART tokenizer. For Masker, we reproduced the model as written in the paper because the authors did not release the code.

B.1 Dataset statistics

Dataset	Style	Train	Dev	Test	Avg.Len
Amazon	Positive	277K	1K	0.5K	64.1
	Negative	278K	1K	0.5K	
Yelp	Positive	192K	5K	1K	131.1
	Negative	192K	5K	1K	
GYAFC EM	Formal	53K	2K	1K	53.2
	Informal	53K	3K	1K	
GYAFC FR	Formal	52K	2K	1K	53.8
	Informal	52K	3K	1K	

Table 6: The datasets statistic. The amount of each dataset (train, dev, test) is rounded up from a position of a thousand. Avg.Len indicates the average character length of the dataset.

C More examples of qualitative evaluation

GYAFC Family & Relationships (informal → formal)	
Original	so so sexy YES it look very nice on most any girl
ST	~~Have~~ so sexy ~~however~~ it look very nice on most any girl.
Masker	That look so so sexy YES it look very nice on most any girl.
StyIns	~~I~~ so attractive, it look very nice on most any ~~information~~.
Ours	So attractive, it look very nice on most women.
Original	keep away ur fear and tell him.
ST	~~Do~~ away ~~likely~~ fear and tell him.
Masker	keep away ~~fromr~~ fear and tell him.
StyIns	~~In~~ away your fear and tell him.
Ours	Keep away your fear and tell him.
Original	tell him, and it wouldn't seem psycho cuz u have kno each other for a long time
ST	~~Let~~ him, and it ~~not~~ not seem psycho ~~c female will have specifyo~~ each other for a long time.
Masker	tell him, and it wouldn't seem psycho cuz u have kno each other for a long time.
StyIns	Tell him, and it wouldn ~~to~~ seem psycho, as you have ~~areo~~ each other for a long time.
Ours	Tell him, and it would not seem awful because you have known each other for a long time.
Original	i can onli say...women r complicated...
ST	~~He~~ can onli say.
Masker	i can onli say... women r complicated.
StyIns	I can ~~on unusual,. They~~ are complicated.
Ours	I can say that women are complicated.

GYAFC Entertainment & Music (informal → formal)	
Original	i will call u to help me.
ST	I will call ~~are~~ to help me.
Masker	I will call u to help me.
StyIns	I will call ~~information~~ to help me.
Ours	I will call you to help me.
Original	I am a big superman fan, i bet u can tell by now.
ST	I am a big superman fan, ~~is~~ bet ~~are~~ can tell by ~~five~~.
Masker	I am a big superman fan, I can tell by now.
StyIns	I am a big superman fan, I bet you can tell by now.
Ours	I am a big Superman fan, I bet that you can tell by now.
Original	Hmmmmmmmm...wow that's a tough one !
ST	~~H Eminem Eminem~~. ~~I~~ that's a tough one available.
Masker	Hmmmmmmmm.. wow that's a tough one!
StyIns	Hmmmmmmmm.wow that is a tough one.
Ours	Yes, that is a tough one.
Original	im famous for inventing a hanger
ST	im famous for inventing a ~~thatanger~~.
Masker	im famous for inventing a hanger.
StyIns	I famous for inventing a ~~veryanger~~.
Ours	I am famous for inventing a hanger.

Table 7: Case Study on *GYFAC Entertainment & Music* and *GYAFC Family & Relationships* dataset. The red and blue words indicate bad and good transfer, respectively. Texts with strikethrough(-) are non-fluent parts of the generated sentences, including grammatical errors.

	Yelp (negative → positive)
Original	Horrible customer service. Employees not knowledgeable of furniture of delivery . Terrible
ST	Great! customer service. Employees excellent knowledgeable of furniture of delivery. Good! Excellent superb!
Masker	Great ~~food~~ service. Employees ~~but~~ not knowledgeable of furniture of delivery. Terrible service ~~Te~~ Terrible
StyIns	Great ~~customer~~ customer service. ~~Great great helpful of beer of.. Great atmosphere~~
Ours	Great customer service. Great knowledgeable of furniture of delivery. Great
Original	Same sad review as others. Called nice person and then got no response. Don't waste your time.
ST	Same delicious review as others. Called nice person and the got great response.
Masker	Same sad ~~face~~ as others. Called nice person and got response. Stop place. waste your time.
StyIns	~~Ec fried food as amazing. Great nice atmosphere~~ and always no response. ~~Can't~~ your time.
Ours	Same good review as others. Called nice person and then got great! Can't wait to go back!
Original	the people here are really rude and slow... and there are bugs crawling all over the bar.
ST	the people here are really friendly and ~~great~~... and there are~~! I all super~~ the bar. Excellent superb!
Masker	the people here are really nice... and there are bugs crawling all over the bar.
StyIns	the people here are really friendly and ~~friendly!~~ and there are always all the best..
Ours	the people here are really friendly and fast... and there are great all over the bar.
Original	Service is good. But, the food was salty and expensive. I'll wait a while before I go back there again.
ST	~~Great~~ is good. ~~The,~~ the food was ~~! and week.~~ I'll wait a while before I go back there again. Excellent superb!
Masker	So good. But, the food was salty and not expensive. I'll wait a while before I go back there!
StyIns	Service is good. Cheap, the food was ~~friendly~~ and expensive. I'll ~~a~~ go again
Ours	Service is good. Delicious, the food was great and reasonable. I'll wait a while before I go back but I'll definitely be back for sure!

	Amazon (negative → positive)
Original	However, definitely does not work like a new battery.
ST	However, definitely does ~~abandon~~ work like a new battery.
Masker	However, definitely does work like a new battery.
StyIns	However, definitely does ~~the job~~ like a new battery.
Ours	However, definitely works like a new battery.
Original	Overall it is inconvenient to keep the counter and the unit clean.
ST	Overall it is ~~announce~~ to keep the counter and the unit clean.
Masker	Overall ~~in~~ inconvenient to keep the counter and the unit clean.
StyIns	Overall it is inconvenient to keep the counter and the unit clean.
Ours	Overall it is easy to keep the counter and the unit clean.
Original	It almost fit, but not enough to snap completely together and fit.
ST	It almost fit, ~~Uses~~ and enough to snap connect together and fit.
Masker	It almost fit, but not enough to snap ~~it~~ completely together and fit.
StyIns	It almost fit, but not enough to snap completely together and fit.
Ours	It almost fits, but enough to snap completely together and fit.
Original	The quality of materials and workmanship is noticeably less.
ST	The quality mixer cord and workmanship is noticeably less.
Masker	The quality of materials and workmanship is noticeably superior.
StyIns	The quality of materials and workmanship is noticeably less.
Ours	The quality of materials and workmanship is noticeably better.

Table 8: Case Study on *Yelp* and *Amazon* dataset. The red and blue words indicate bad and good transfer, respectively. Texts with strikethrough(-) are non-fluent parts of the generated sentences, including grammatical errors.