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

Anonymous ACL submission

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 with the energy-based interpretation and leverage a pretrained language model to improve fluency. Specifically, we propose a novel approach of applying 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 evaluate our model on four public benchmarks Amazon, Yelp, GYAFC, and Civil Comments and achieve 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 has transferred from the original sentence to the generated one, 2) *content preservation*, how well the generated sentence has retained the semantics from the original, and 3) *fluency*, how natural the generated sentence is.

Text style transfer is challenging in that converting a style of the sentence fluently often conflicts with preserving the content (John et al., 2019; Prabhumoye et al., 2018; Gong et al., 2019). To address this challenge, several supervised text style transfer methods have been attempted (Al Nahas et al., 2019; Lai et al., 2021), but style-labeled sentence pairs are not often available, making it less practical in a real-world setting. Therefore, unsupervised text style transfer approaches have been popular using autoencoder (Hu et al., 2017), back-translation (Prabhumoye et al., 2018; Lample et al., 2018), and reinforcement learning (Xu et al., 2018; Luo et al., 2019). Among the previous studies, Style Transformer (Dai et al., 2019) achieved fine-grained style control by deceiving the style discriminator through adversarial training. Aside from its strength, however, the adversarial models including Style Transformer suffer from fluency degradation for the generated sentences.

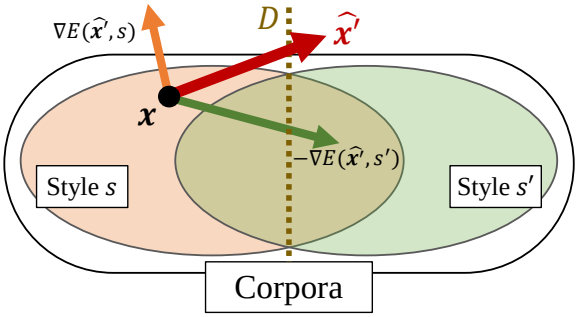

Figure 1: Energy-based interpretation for fluency degradation. Deceiving energy-based discriminator D requires 1) minimizing the energy E between the transferred sentence $\widehat{\mathbf{x}}'$ and the target style s', and 2) maximizing the energy between the sentence $\widehat{\mathbf{x}}'$ and the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

In this paper, we analyze the reason behind the fluency degradation in adversarial models by reviewing Style Transformer. To interpret what exactly fluency is, we introduce the notion of *energy* (Hinton, 2002; Lecun et al., 2006), which is the entropy in 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, a formal/informal sentence would likely have low energy in formality corpora, while sentences that has nothing to do with formality (e.g., political expressions) would have high energy in the corpora.

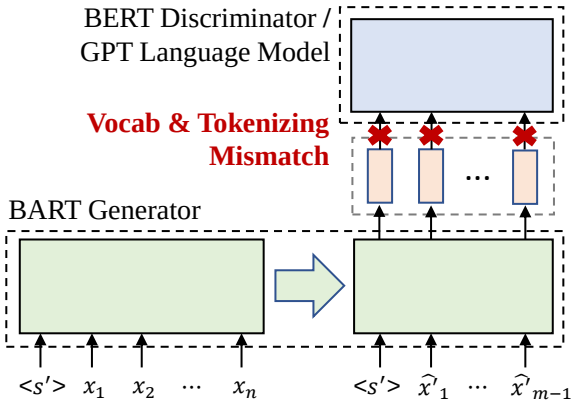

Figure 2: A structural dilemma to apply pretrained models to adversarial learning. To propagate gradients to generated tokens, the generator, discriminator, and LM should have the same vocabulary and tokenizer. However, publicly available pretrained models (e.g., BERT and GPT) use their own tokenizers,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

Hence, we define *fluency* as having low energy in particular corpora, in which the fluent sentences express one of the styles in the corpora. As illustrated in Figure 1, fluency degrades while deceiving the discriminator, since the adversarial learning maximizes the energy to the source style and drives the generated sentence far away from the distribution of the corpora. 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 corpora by pulling the sentence to the target corpus.

To apply the LM-based regularizer, it is desirable to bring the pretrained model such as GPT-2 (Radford et al., 2019) for fluent generation. Moreover, fluency is expected to improve further 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 arises in that the two of three modules may need to be re-trained from scratch because the existing pretrained models are based on different tokenizers. We restructure the discriminator and LM such that the pretrained model is applied to all three modules: the generator, discriminator, and LM.

Our contributions can be summarized as follows:

- We analyze the fluency degradation in adversarial training with the 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 pretrained language model can be employed in the text style transfer framework.
- We achieve new state-of-the-art results on Amazon, Yelp, GYAFC, Civil Comments 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 aimed to learn disentangled representations of text by separating the meaning of sentences into content and style in the 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 a language model as a discriminator to enhance the informativeness and stability of adversarial training. These works with disentangled representations showed a reasonable performance with high interpretability, but disentangled content representations could still contain style-relevant information as pointed out by Lample et al. (2018). In addition, there is a limitation in that the meaning of the input sentence must be expressed in a fixed-size vector with limited capacity (Dai et al., 2019).

In contrast, there have been methods without disentangled representations that did not explicitly disentangle the content and style of text. There also have been several approaches using reinforcement learning (Xu et al., 2018; Luo et al., 2019) and back-translation (Lample et al., 2018; Prabhumoye et al., 2018). Dai et al. (2019) proposed a new style transfer model based on the transformer architecture without disentangled representations. Wang et al. (2019) proposed an unsupervised framework by editing entangled latent representations. Our work proposes a new way of effectively leveraging a pretrained language model into an unsupervised text style transfer task.

2.2 Style transfer with pretrained models

Recently, pretrained language models have had a huge success on various NLP tasks such as machine translation (Chronopoulou et al., 2020) and text summarization (Liu and Lapata, 2019). Furthermore, the models are also being used in text style transfer task. Sudhakar et al. (2019) leveraged GPT (Radford et al., 2018) to capture a representation of content words in a source sentence with that of attribute words which are retrieved from a target style corpus. Malmi et al. (2020) used a padded masked language model (Mallinson et al., 2020) variant, which is pretrained on the same corpora that BERT (Devlin et al., 2019) used. It removed the necessity to predetermine the number of tokens to be infilled on a source sentence. Although they exploited the power of pretrained models, our approach differs in the fact that we train our model adversarially in an end-to-end manner.

For the purpose of transferring writing styles between authors, Syed et al. (2020) pretrained a language model from scratch on the author corpus with masked language modeling. Laugier et al. (2021) detoxified toxic texts by finetuning a pretrained T5 with two additional objectives: a denoising objective and a cycle-consistency objective. However, these studies only focused on a specific domain. Meanwhile, Lai et al. (2021) finetuned BART (Lewis et al., 2020) using parallel data with policy gradient (Sutton et al., 1999) which maximized two rewards: a style classifier reward and a BLEU score reward. However, as the model was trained with supervision, it is infeasible to directly compare it with our work. Our work incorporates pretrained models with adversarial training, and it shows great performance on various domains while trained on the nonparallel 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) With the energy-based model, we can classify x by comparing the energy of each label, or generate x by optimizing $\arg\min_x E(x)$.

There are several works leveraging the energy-based model in image generation (Ngiam et al., 2011; Zhao et al., 2017), text generation (Deng et al., 2019; 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 the energy function. 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 the unsupervised setting, we assume the nonparallel corpus $\mathbf{X} = \{\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}\}$ and $\mathbf{X}' = \{\mathbf{x}'^{(0)}, \mathbf{x}'^{(1)}, \cdots, \mathbf{x}'^{(m)}\}$, 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 the similar content but the style of the other corpus.

3.1 Preliminaries

Style Transformer Dai et al. (2019) proposed the unsupervised style transfer model based on the transformer architecture (Vaswani et al., 2017). On their work, the self loss $\mathcal{L}_{\text{self}}$ and cycle loss $\mathcal{L}_{\text{cycle}}$ are used to preserve content, while the style loss $\mathcal{L}_{\text{style}}$ aims to control style. 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

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

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 $\widehat{\mathbf{x}}' \sim G(\mathbf{x}, s')$, the content of the sentence should be preserved. Along with the previous studies (Logeswaran et al., 2018; Xu et al., 2018), Style Transformer adopts the cycle loss

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

which regularizes the generated sentence to be identical with 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 attribute s. The discriminator is trained separately from the generator and

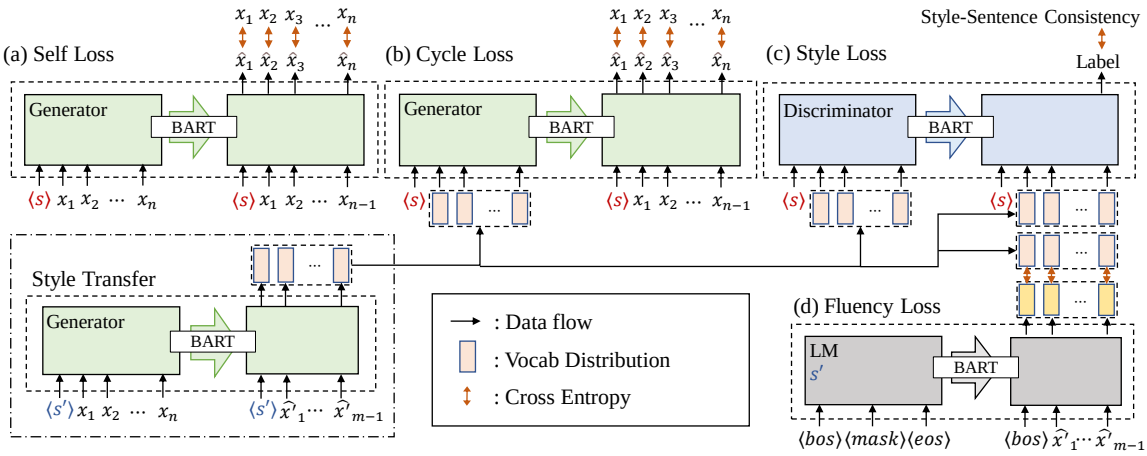

Figure 3: Overall structure of our model. Our model learns how to transfer the style and preserve the content through four mechanisms; a model (a) reconstructs a source sentence, (b) cyclically reconstructs a sentence in a style of source corpus while preserving its content, (c) deceives a discriminator by generating a target corpus style-like sentence, and (d) improves fluency of generated sentence by using langauge model. All modules leverages BART and trained in the end-to-end manner.

takes the generated sentences also as negative samples along with the original sentences. The training process for the discriminator optimizes

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

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 targets to deceive the discriminator by generating sentences with the target style:

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

The upper part of Figure 3 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 discreteness of language prevents the trivial solution. To propagate the gradients directly, Style Transformer feeds the generated sentences to the discriminator in the form of the 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), 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 auto-regressive manner:

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

3.2 Energy-based interpretation for fluency degradation

In our preliminary study, we found that there is a significant gap between the perplexity of the target corpus 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 *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. With this interpretation, the style discriminator could be decomposed into

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

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 the two logits for each label. 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 could 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 corpora sharing syntactic and semantic attributes. 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 a language model to prevent the generated sentence from being out of the distribution of the corpora. As the discriminator pushes out the sentence from the distribution, we require an additional power to pull it back into the corpora. Thus, we introduce a fluency loss $\mathcal{L}_{\text{fluent}}$ which pulls the generated sentence into the target corpus distribution. For each style s, we train a language model with $\mathcal{L}_{\text{LM}}(\theta_{\text{LM}_s}) = -\mathbb{E}\left[\sum_i \log p_{\text{LM}_s}(x_i; x_{1:i-1})\right]$ in advance, and optimize the cross entropy of the generated sentence during training along with other losses as

$$\mathcal{L}_{\text{fluency}}(\theta_G) = -\sum_i p^i_G(\widehat{\mathbf{x}'}; \mathbf{x}, s') \log p^i_{\text{LM}_{s'}}(\widehat{\mathbf{x}'}) \tag{7}$$

where $p^i_G(\widehat{\mathbf{x}'}; \mathbf{x}, s') = p_G(\widehat{x'}_i; \widehat{x'}_{1:i-1}, \mathbf{x}, s')$ and $p^i_{\text{LM}_{s'}}(\widehat{\mathbf{x}'}) = p_{\text{LM}_{s'}}(\widehat{x'}_i; \widehat{x'}_{1:i-1})$. We report and analyze the fluency enhancement with this loss in Section 4.6.

3.3 Consideration for structural dilemma toward adversarial training

For fluent generation, it is desirable to apply the pretrained model (Radford et al., 2019; Brown et al., 2020) to the regularizer. Not only for the LM, we apply the pretrained model also to the generator and discriminator for fluent style control. As 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 the language model. When training Style Transformer, the style discriminator takes the softmax distribution of the generated sentences, and thus the discriminator should share the same vocabulary as the generator. This problem is not only limited to Style Transformer but also expands to the model requiring gradient back-propagation on token level using gumbel softmax (Jang et al., 2017). As the discriminator in Style Transformer adopts the transformer encoder structure, BERT (Devlin et al., 2019) is the most feasible option, but there is no publicly available BERT model with the BART vocab. This requires training BERT from scratch which needs a lot of resources. If the discriminator is not based on the pretrained model, the fluency degrades significantly as shown in section 4.6.

There is, however, a rather simple solution to this problem of mismatching tokenizers: We use the same pretrained model for the generator and discriminator. Thus, we leverage the BART classifier proposed by 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. Figure 3 describes the style discriminator with BART. As the generator and discriminator share the same BART vocab, the softmax distribution on the vocab could be transferred in an end-to-end manner.

Just like the generator and the discriminator, we use the pretrained model also for the language model. 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). Figure 3 shows how we adopt BART to the language model. To tackle the problem similarly with the text infilling task, we feed the mask token in form of [`<bos>`, `<mask>`, `<eos>`] into the encoder and freeze it while finetuning the decoder to the target corpus. After finetuning separate language models for both styles, we leverage them to enhance the fluency of the generated sentences by Eq. (7). We concatenate the target style label in front of the input of the BART encoder and decoder as depicted on Figure 3. The other details on architecture and training are available in Appendix A.

4 Experiments

4.1 Datasets

For the experiments, we use four widely-used English datasets: Amazon, Yelp reviews, Grammarly's Yahoo Answers Formality Corpus (GYAFC), and Civil Comments. Dataset statistics are available in Appendix B.

Amazon The Amazon dataset is a product review dataset, labeled as either a positive or negative sentiment style. We use the preprocessed dataset provided by Wang et al. (2019) but use a raw sentence due to the pretrained model having its own tokenizer.

Yelp Following the work of Shen et al. (2017), we conduct experiments on the Yelp dataset[1]. The Yelp dataset is a restaurant and business review dataset with positive and negative sentiments. Only reviews between 10 and 180 in character lengths are included, and reviews with a rating of 5 are labeled positive, and reviews with ratings of 1 and 2 are labeled negative.

GYAFC The GYAFC dataset (Rao and Tetreault, 2018) is a question and answer dataset on the online forum, consisting of informal and formal sentences from the two categories: Entertainment & Music and Family & Relationships.

Civil Comments The Civil Comments dataset[2] consists of comments to worldwide news and their toxicity measured by crowd raters (Borkan et al., 2019). Referring to the work of Laugier et al. (2021), we label the comment over a score of 0.5 as toxic, and label as non-toxic if the comment has a zero toxicity.

4.2 Baselines

We choose four unsupervised baselines, **CrossAlign** (Shen et al., 2017), Style Transformer (**ST**) (Dai et al., 2019), **Masker** (Malmi et al., 2020) , and **Thk.BART** (Lai et al., 2021), since they are similar to our proposed method. [3] CrossAlign is based on adversarial learning and Style Transformer is the basis of our model architecture. Masker utilized the pretraining process of BERT and used pretrained models, GPT-2 and BART. We report **Source Copy** and **Target Copy** which evaluates the source and target corpus with the same metrics.

4.3 Evaluation metrics

An expected output sentence is a sentence transferred to the target style while preserving the content of the input sentence and maintaining its fluency. Therefore, the performance is measured by three criteria: 1) style transfer accuracy 2) content preservation 3) fluency.

Style transfer accuracy This metric indicates how many generated sentences are accurately transferred to the target style. This is measured by the prediction accuracy of the style classifier implemented by a finetuned BERT classifier.

Content preservation This metric is computed by BLEU score (Papineni et al., 2002) between the generated sentences and inputs themselves [4]. We denote this metric as *self*-BLEU.

For GYAFC task, as human-written reference sentences are available, we additionally measure the BLEU score between generated sentences and human-written references. We denote this metric as *ref*-BLEU.

Fluency This metric is measured by the average perplexity (PPL) of the generated sentences using a finetuned GPT-2 model.

Overall metric Since the style transfer accuracy and content preservation are trade-off, we report the harmonic mean of the classifier's accuracy and BLEU (*self*-BLEU, *ref*-BLEU) as the overall performance (Luo et al., 2019; Lai et al., 2021), except for the perplexity.

4.4 Quantitative Results

Table 1 and 2 show the experimental results on the GYAFC dataset and other datasets, Amazon, Yelp, Civil Comments, respectively.

The perplexity of the source copy is not extremely high when compared to human references in Table 1. This is because the source and target corpus are monolingual and share a common topic such as entertainment or human relationships. Therefore, it numerically proves that the energy functions for each style are similar, so text style transfer models should maintain low perplexity while transferring the sentence. Even without

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

[2] https://www.tensorflow.org/datasets/catalog/civil_comments

[3] **Thk.BART** is originally a supervised model, so we only employ the refinement part with reinforcement learning while inputting a source sentence instead of the transferred one.

[4] We measure BLEU score by NLTK word tokenizer (Loper and Bird, 2002) and `multi-bleu.perl`.

Approach	Entertainment & Music					Family & Relationships				
	ref-B.	*self*-B.	Acc.	H.mean	PPL ↓	*ref*-B.	*self*-B.	Acc.	H.mean	PPL ↓
Source Copy	40.3	100	11.1	24.0	79.0	41.0	100	11.1	24.2	52.1
Human Ref.	100	21.4	86.3	43.9	47.4	100	22.8	86.8	45.9	31.0
CrossAlign (Shen et al., 2017)	3.81	3.41	**73.8**	5.3	119	2.93	2.86	63.4	4.24	72.2
ST (Dai et al., 2019)	32.3	55.1	59.3	45.5	428	35.1	55.7	49.7	45.0	193
Masker (Malmi et al., 2020)	38.6	75.7	25.2	38.0	98.9	39.1	77.1	22.9	36.5	61.6
Thk.BART (Lai et al., 2021)	40.1	**99.3**	11.3	24.3	77.8	40.8	**99.0**	11.5	24.7	53.5
Ours	**48.4**	60.0	62.3	**56.2**	**50.6**	**50.9**	57.6	**70.0**	**58.5**	**38.1**

Table 1: Experimental results on Entertainment & Music and Family & Relationships set of *GYAFC* dataset. ↓ indicates the smaller the better. Acc., *self*-B., *ref*-B., and H.mean indicate accuracy, *self*-BLEU, *ref*-BLEU, and harmonic mean, respectively. Among the methods except for Source Copy and Human Reference, the best result is shown in **bold**, and the second-highest result is underlined.

Approach	Amazon				Yelp				Civil Comments			
	self-B.	Acc.	H.mean	PPL ↓	*self*-B.	Acc.	H.mean	PPL ↓	*self*-B.	Acc.	H.mean	PPL ↓
Source Copy	100	13.1	23.2	40.3	100	3.4	6.5	55.1	100	14.9	26.0	70.7
Target Copy	-	86.9	-	26.2	-	96.6	-	29.4	-	85.1	-	57.5
CrossAlign	-	-	-	-	11.3	36.7	17.2	396	-	-	-	-
ST	91.1	21.8	35.3	58.9	21.0	51.1	29.7	192	23.2	33.9	27.5	187
Masker	73.1	28.3	40.8	67.3	71.9	26.8	39.0	69.0	73.0	28.6	41.1	102
Thk.BART	**99.5**	13.1	23.2	**40.1**	**86.8**	3.25	6.27	**57.8**	**75.7**	35.2	48.1	**63.3**
Ours	60.1	**74.6**	**66.6**	51.0	66.0	**72.4**	**69.1**	58.9	62.1	**74.5**	**67.7**	88.1

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

human references, we can observe the same phenomenon by comparing the perplexity of the target copy as in Table 2.

Cross Align, which is based on autoencoder, has very low *self* BLEU due to the poor reconstruction on the input sentence. Style Transformer shows high accuracy in general, but unfortunately it reports high perplexity. As we mentioned before, the perplexity of Style Transformer is high because the generated sentence deviates from the source style corpus. Thk.BART has a considerably high *self*-BLEU at the expense of the accuracy. In other words, this means that the input sentence is being outputted without text style transfer through the model. Therefore, it shows a relatively low perplexity, which is similar to the perplexity of the source copy.

Considering the harmonic mean as overall performance, our model outperforms the baselines on all datasets. Futhermore, our model achieves the lowest perplexity (53.8, 41.8) on each GYAFC EM, FR dataset, which is a little difference from the perplexity of human reference (47.4, 31.0). The model yields a sufficiently desirable perplexity for the rest of the datasets in Table 2.

4.5 Qualitative Results

Table 3 shows some of style transfer results by each model on GYAFC Entertainment & Music and Yelp dataset. [5]

An original example of GYAFC dataset is an informal sentence that 'u', the beginning of the sentence, begins in a lowercase letter. CrossAlign and ST do not properly transfer the text style. CrossAlign completely fails to transfer the style, while ST converts an unrelated verb into 'Look' and the fluency was also degraded with the appearance of two verbs ('Look Mean') in a row. In the case of our model, a desirable sentence is generated by converting 'u' into a formal form of 'You' and generating 'Y' of 'You' that starts with a capital letter.

An original example of Yelp dataset is a negative sentence that the author does 'not recommended' the bubble tea because it is 'not good' due to 'careless' boss or server. Our model transferred into a positive sentence well by excluding 'not' and replacing 'careless' with 'caring' and 'not' with 'highly'. CrossAlign correctly excludes 'not', but the fluency is declined by generating an additional word 'great'. ST does not preserve the content and

[5] More examples are available in Appendix D.

GYAFC Entertainment & Music (informal → formal)	
Original	u mean all actors who've ever played superman?
CrossAlign	Can you find out about this year that knows?
ST	Look mean all actors who've ever played superman?
Masker	Do u mean all actors who've ever played superman?
Thk. BART	u mean all actors who've ever played superman?
Ours	You refer to all actors who have ever played superman?
Yelp (negative → positive)	
Original	The bubble tea is really not good. The boss or sever has some kind of careless. Not recommend
CrossAlign	The bubble tea is really good great. The thorough or <unk>
ST	The bubble tea is really good. ~~The boss or sever has some kind of careless. Not recommend~~
Masker	The bubble tea is really good. The sever has some kind of careless. Highly recommend
Thk. BART	The bubble tea is really not good. The boss or sever has some kind of careless. Not recommend
Ours	The bubble tea is really good. The boss or sever has some kind of caring. Highly recommend

Table 3: Case Study on *GYAFC Entertainment & Music* and *Yelp* dataset. The red and blue words indicate bad and good transfer, respectively. Texts with strikethrough(-) are a part of the original sentence whose content is not preserved in a generated sentence.

even Masker, which generated a qualified sentence, could not manage to convert the 'careless' word.

	Yelp			
Approach	*self*-B.	Acc.	H.mean	PPL ↓
ST (Dai et al., 2019)	21.0	51.1	29.7	192
+ G_{PT}	65.5	65.0	65.2	101
+ G_{PT} + $\mathcal{L}_{\text{fluency}}$	**66.0**	64.5	65.2	78.3
+ G_{PT} + D_{PT}	63.8	**76.6**	**69.6**	63.5
Ours	**66.0**	72.4	69.1	**58.9**

Table 4: Ablation study on **Yelp** dataset. G_{PT} and D_{PT} indicates applying the pretrained models to the generator and discriminator respect, and **-**$\mathcal{L}_{LM}$ indicate the fluency loss without leveraging pretrained parameters

4.6 Ablation study

We conduct an ablation study to understand the contribution of each component in our proposed method. The results of ablation study on the Yelp dataset are shown in Table 4. We choose Yelp for the study because it has the longest average sentence lengths and the fluency degradation significantly occurs in Style Transformer which is our base model. As we can see, when a pretrained model is applied to the generator, all the metrics are improved. This implies that leveraging a pretrained model is helpful when generating plausible sentences on style transfer task. We evaluate the impact of the fluency loss $\mathcal{L}_{\text{fluency}}$ by employing the vanilla transformer decoder trained on the target corpus. Although there is no change on the harmonic mean, fluency has improved considerably, which means there is a potential on Style Transformer to improve fluency degraded by adversarial training. We check the effect of initializing the multiple components with the pretrained weight by applying BART also to the discriminator. Along with the pretrained discriminator, the harmonic mean and fluency has significantly enhanced, so applying this means pretrained discriminator is helpful on improving overall performance of style transfer. At last, when pretrained models are applied to all component including the genrator, discriminator, and LM, the output has enhanced at fluency showing the best performance among all the other cases at the expense of only a small performance drop on harmonic mean.

5 Discussion

We have only conducted experiments on the 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. In addition, an objective human evaluation of a third party can also be introduced.

6 Conclusion

Through the energy-based interpretation, we find that the fluency is inevitably degraded when deceiving the discriminator of Style Transformer (Dai et al., 2019). The problem was solved by adding an auxilliary LM-based regularizer. To apply the regularizer to Style Transformer, we leverage the pretrained model BART. Our model shows state-of-the-art performance on text style transfer, content preservation and fluency. Furthermore, we prove the robustness of our model by conducting extensive experiments on various styles.

References

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.

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.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. pages 491–500.

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.

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.

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. 2019. 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.

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 J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA. MIT Press.

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*, ICML'17, page 1352–1361. JMLR.org.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800.

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*, pages 1587–1596.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparametrization with gumbel-softmax. In *Proceedings International Conference on Learning Representations (ICLR)*.

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.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. 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. 2018. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

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, and Raia Hadsell. 2006. *A tutorial on energy-based learning*.

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.

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.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *CoRR*, cs.CL/0205028.

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, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019*.

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 masked language models. In *EMNLP*.

Jiquan Ngiam, Zhenghao Chen, Pang Wei Koh, and Andrew Y. Ng. 2011. Learning deep energy models. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1105–1112, Madison, WI, USA. Omnipress.

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.

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.

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.

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 *EMNLP*.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 1057–1063, Cambridge, MA, USA. MIT Press.

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 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*. Curran Associates Inc., Red Hook, NY, USA.

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, Rémi 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 M. 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.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7298–7309.

Junbo Zhao, Michael Mathieu, and Yann LeCun. 2017. Energy-based generative adversarial networks. Publisher Copyright: © ICLR 2019 - Conference Track Proceedings. All rights reserved.; 5th International Conference on Learning Representations, ICLR 2017 ; Conference date: 24-04-2017 Through 26-04-2017.

A Experimental setup

Implementation details We first finetune a discriminator and language models with the source sentences and labels, and train the generator with the autoencoding objective using the training corpus. The language models are frozen, and the generator and the discriminator are finetuned again in an end-to-end manner. The main training procedure is similar to the training strategy of GAN (Goodfellow et al., 2014) in that we train the discriminator for several times while the generator takes one step.

Architecture details Our implementation is based on `bart-base` [6] of the Huggingface's Transformers library (Wolf et al., 2020), which has 406 million parameters in total. On inference time, next token is decoded in a greedy fashion, and we constrain n-gram whose n is bigger than three not to be generated again.

Training details We perform a hyperparameter tuning for every dataset. We record the model checkpoint per 500 steps, and the model with the highest harmonic mean in a single run is selected as our final model. Our model is trained on a single NVIDIA RTX A6000 machine, and it takes about 30 hours to train *Yelp* dataset. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a batch size of 64. 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 last linear layer in the discriminator is label-smoothed (Szegedy et al., 2016) with $\alpha = 0.1$.

B 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	
Civil Comments	Toxic	73K	4K	4K	99.3
	Non-Toxic	73K	4K	4K	

Table 5: The datasets statistic. The amount of each dataset (train, dev, test) is rounded up from a position of thousand. Avg.Len indicates the average character length of the dataset.

C Ethical consideration

We paid attention to avoid ethical issues because some of the civil comments dataset includes insulting remarks regarding a particular party or politician. However, just like changing toxicity to non-toxicity, we proposed our work in anticipation of positive applicability as in previous studies.

D More example for qualitative evaluation

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

GYAFC Entertainment & Music (informal → formal)	
Original	i cant believe u asked that question, but i'll answer it.
CrossAlign	I ~~cant~~ believe that people know, I do not know about.
ST	I prefer believe that ~~u~~ asked that question, but I'll answer it.
Masker	I cant believe u asked that question, but i'll answer it.
Thk. BART	i cant believe u asked that question, but i'll answer it.
Ours	I cannot believe you asked that question, but I'll answer it.
Original	This is sooo funny, it has perverted humor to it.
CrossAlign	This is not good, but it is going to go.
ST	This is sooo funny, it has perverted humor to it.
Masker	This is sooo funny, it has perverted humor. ~~to it~~.
Thk. BART	This is sooo funny, it has perverted humor to it.
Ours	This is very funny, it has perverted humor to it.
Original	they make u think of who your with, and love them more than u did already
CrossAlign	They need out of them, and they are in <unk>.
ST	~~they~~, make u think of who your with, and love them more than u did already.
Masker	they make u think of who your with, and love them more than u did already.
Thk. BART	they make u think of who your with, and love them more than u did already
Ours	They make you think of who you are with, and love them more than you did previously.
Original	It goes somethin like that for the chorus.
CrossAlign	It's out for the other that time.
ST	It goes involvinghin like that for the chorus.
Masker	It goes some like that for the chorus.
Thk. BART	It goes somethin like that for the chorus.
Ours	It goes something similar to that for the chorus.

GYAFC Family & Relationships (informal → formal)	
Original	I love curly hair I like to run my fingers thru it
CrossAlign	I wanted <unk> to me and I wanted it is the guy.
ST	I love ~~curly~~ hair I like to run my fingers horror it horror tub me
Masker	I would like to run my fingers thru it
Thk. BART	I love curly hair I like to run my fingers thru it
Ours	I love curly hair I like to run my fingers through it.
Original	give a rose and express ur love
CrossAlign	Ask him a little and kiss her.
ST	Ask a rose and express you love.
Masker	give a rose and express yourr love
Thk. BART	give a rose and express ur love
Ours	Give a rose and express your love.
Original	u'll see that no matter how hurt u r it's still the simplest reason y u smile.
CrossAlign	They do not worry about it.
ST	A cannot see that no matter how hurt, using it is still the reason last someone you in fre simplest freind.
Masker	You'll see that no matter how hurt u r it's still the simplest reason y u smile.
Thk. BART	u'll see that no matter how hurt u r it's still the simplest reason y u smile.
Ours	You will see that no matter how hurt you are, it is still the simplest reason you smile.
Original	no, you are punish, and no tv for a month
CrossAlign	~~no~~, you are, and, and you are a good thing.
ST	Perhaps provide you are punish provide and noaith for a month a month a truth a sign a month provide a email a good a her a her a her down
Masker	no, you are punishing by no tv for a month
Thk. BART	no , you are punish , and no tv for a month
Ours	No, you are punished, and no television for a month.

Table 6: 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 a part of the original sentence whose content is not preserved in a generated sentence.

Yelp (negative → positive)	
Original	Very rude waitress. We felt unwelcome and very uncomfortable. We will not return to this location...ever.
CrossAlign	Very rude waitress. We loved it, very <unk> ~~We will not return to this location...ever.~~
ST	Very friendly waitress. ~~We felt unwelcome and very uncomfortable. We will not return to this location...ever.~~
Masker	Very nice and waitress. We felt verylcome and very uncomfortable. We will definitely return to this location... ever.
Thk. BART	Very rude waitress. We felt unwelcome and very uncomfortable. We will not return to this location...ever.
Ours	Very friendly waitress. We felt welcome and very comfortable. We will definitely return to this location...ever.
Original	Absolutely horrible! Called twice, never showed up... Avoid by all costs
CrossAlign	Absolutely horrible! Fresh here, never showed up~~... Avoid by all costs~~
ST	Absolutely fantastic! ~~Called twice, never showed up... Avoid by all costs~~
Masker	Absolutely amazing! Called twice, never showed up... Avoid all costs
Thk. BART	Absolutely horrible! Called twice, never showed up ... Avoid by all costs
Ours	Absolutely awesome! Called twice, always showed up... Thank you by all costs
Original	Poor customer service Very expensive Macarons and bread are good!
CrossAlign	Excellent customer service, affordable people and ~~bread are good!~~
ST	Excellent customer service Very knowledgeable Macarons and bread are good!
Masker	Great customer service Very expensive Macarons and bread are good!
Thk. BART	Poor customer service Very expensive Macarons and bread are good!
Ours	Excellent customer service Very affordable Macarons and bread are good!
Original	Shitty food for an exorbitant price. $4.50 for powdered hot chocolate, just as a sample.
CrossAlign	Shitty food for an hour price. $~~4.50 for powdered hot chocolate, just as a sample.~~
ST	Shped food for an ex hilariousant price. ~~$4.50 for powdered hot chocolate, just as a sample.~~
Masker	Shitty food for a exorbitant price. $ 4. 99 for powdered hot chocolate, just as a sample.
Thk. BART	Shitty food for an exorbitant price. $4.50 for powdered hot chocolate, just as a sample.
Ours	Great food for an great price. $4.50 for powdered hot chocolate, just as a sample.

Amazon (negative → positive)	
Original	I purchased them and when they arrived they were torn.
ST	I purchased them and when they arrived they were torn.
Masker	I purchased them and when they arrived they were torn apart.
Thk. BART	I purchased them and when they arrived they were torn.
Ours	I purchased them and when they arrived they were perfect.
Original	I bought them and find them pretty annoying.
ST	I bought them and find them pretty annoying.
Masker	I bought two of them and find them pretty annoying.
Thk. BART	I bought them and find them pretty annoying.
Ours	I bought them and find them pretty nice.
Original	Yes, they are light, and are not durable at all.
ST	Yes, they are light, and are not durable at all.
Masker	Yes, they are light weight, and are not durable at all.
Thk. BART	Yes, they are light, and are not durable at all
Ours	Yes, they are heavy, and are durable perfectly.
Original	After a few days I noticed I was unusually fatigued.
ST	After a few days I noticed I was unusually fatigued.
Masker	After a few days of use I noticed I was unusually fatigued.
Thk. BART	After a few days I noticed I was unusually fatigued.
Ours	After a few days I noticed I was unusually happy.

Table 7: Case Study on *Yelp* and *Amazon* dataset. The red and blue words indicate bad and good transfer, respectively. Texts with strikethrough(-) are a part of the original sentence whose content is not preserved in a generated sentence.

Civil Comments (toxicity → non-toxicity)	
Original	Not that I don't disagree with you but this isn't the time or place. Feel sorry for those 2 little babies that they had an idiot mother
ST	Not that I don't disagree with you but this isn't the time or place. ~~Feel sorry for those 2 little babies that they had an idiot mother~~
Masker	I don't disagree with you but this isn't the time or place. Feel sorry for those ~~2 little~~ babies that they had an idiot mother
Thk. BART	Not that I don't disagree with you but this isn't the time or place. Feel sorry for those 2 little babies that they had an idiot
Ours	Not that I don't disagree with you but this isn't the time or place. Feel sorry for those 2 little babies
Original	Sorry, this may be banned, but your comment is Stupid!
ST	Sorry, this may be banned, but your comment is Thanks!
Masker	Sorry, this may be banned, but your comment is Stupid.
Thk. BART	Sorry, this may be banned, but your comment is Stupid!
Ours	Sorry, this may be banned!
Original	XXXXX, I see you like your own comments. Troll on, buddy!
ST	XXXXX, I see you like your own comments.
Masker	XXXXX, I see you like your comments. Troll on, buddy.
Thk. BART	XXXXX, I see you like your comments. Troll on, buddy!
Ours	XXXXX, I see you like your own comments!
Original	Asking a portfolio manager for housing advice is like asking a dentist for automotive advice. DUMB.
ST	Asking a portfolio manager for housing advice is like asking a dentist for automotive advice.
Masker	Asking a portfolio manager for housing advice is like asking a portfolio for for automotive advice. DB.
Thk. BART	Asking a portfolio manager for housing advice is like asking a dentist for automotive advice. DUMB.
Ours	Asking a portfolio manager for housing advice is like asking a dentist for automotive advice.

Table 8: Case Study on *Civil Comments* dataset. The red and blue words indicate bad and good transfer, respectively. Texts with strikethrough(-) are a part of the original sentence whose content is not preserved in a generated sentence.