# Revision in Continuous Space: Unsupervised Text Style Transfer without Adversarial Learning

**Dayiheng Liu,[†] Jie Fu,[‡] Yidan Zhang,[†] Chris Pal,[‡] Jiancheng Lv[†*]**
[†]College of Computer Science, Sichuan University
[‡]Québec Artificial Intelligence Institute (Mila), Polytechnique Montréal
losinuris@gmail.com, lvjiancheng@scu.edu.cn

**Abstract**

Typical methods for unsupervised text style transfer often rely on two key ingredients: 1) seeking the explicit disentanglement of the content and the attributes, and 2) troublesome adversarial learning. In this paper, we show that neither of these components is indispensable. We propose a new framework that utilizes the gradients to revise the sentence in a *continuous* space during inference to achieve text style transfer. Our method consists of three key components: a variational auto-encoder (VAE), some attribute predictors (one for each attribute), and a content predictor. The VAE and the two types of predictors enable us to perform gradient-based optimization in the *continuous* space, which is mapped from sentences in a *discrete* space, to find the representation of a target sentence with the desired attributes and preserved content. Moreover, the proposed method naturally has the ability to simultaneously manipulate multiple *fine-grained* attributes, such as sentence length and the presence of specific words, when performing text style transfer tasks. Compared with previous adversarial learning based methods, the proposed method is more *interpretable*, *controllable* and *easier* to train. Extensive experimental studies on three popular text style transfer tasks show that the proposed method significantly outperforms five state-of-the-art methods.

## 1 Introduction

Text style transfer, which is an under-explored challenging task in the field of text generation, aims to convert some attributes of a sentence (e.g., negative sentiment) to other attributes (e.g., positive sentiment) while preserving attribute-independent content. In other words, text style transfer can generate sentences with desired attributes in a controlled manner. Due to the difficulty in obtaining training sentence pairs with the same content and differing styles, this task usually works in an unsupervised manner where the model can only access non-parallel, but style labeled sentences.

Most existing methods (Hu et al. 2017; Shen et al. 2017; Fu et al. 2018; Li et al. 2018) for text style transfer usually first explicitly disentangle the content and the attribute through an adversarial learning paradigm (Goodfellow et al. 2014). The attribute-independent content and the desired attribute vector are then fed into the decoder to generate the target sentence. However, some recent evidence suggests that using adversarial learning may not be able to learn representations that are disentangled (Li et al. 2018; Guillaume Lample 2019). Moreover, vanilla adversarial learning is designed for generating real-valued and continuous data but has difficulties in generating sequences of discrete tokens directly . As a result, algorithms such as REINFORCE (Sutton et al. 2000) or those that approximate the discrete tokens with temperature-softmax probability vectors (Kusner and Hernández-Lobato 2016; Zhang et al. 2017) are used. Unfortunately, these methods tend to be unstable, slow, and hard-to-tune in practice (Guillaume Lample 2019).

Is it really a necessity to explicitly disentangle the content and the attributes? Also, do we have to use adversarial learning to achieve text style transfer? Recently, the idea of mapping the discrete input into a continuous space and then performing gradient-based optimization with a predictor to find the representation of a new discrete output with desired property has been applied for sentence revision (Mueller, Gifford, and Jaakkola 2017) and neural architecture search (Luo et al. 2018). Motivated by the success of these works, we propose a new solution to the task of content-preserving text style transfer.

The proposed approach contains three key components: (a) A **variational auto-encoder (VAE)** (Kingma and Welling 2013), whose encoder maps sentences into a smooth continuous space and its decoder can map a continuous representation back to the sentence. (b) Some **attribute predictors** that take the continuous representation of a sentence as input and predict the attributes of its decoder output sentence, respectively. These attribute predictors enable us to find the target sentence with the desired attributes in the continuous space. (c) A **content predictor** that takes the continuous representation of a sentence as input and predicts the Bag-of-Word (BoW) feature of its decoder output sentence. The purpose of component (c) is threefold: First, it could enhance the content preservation during style transfer; Second, it enables the target sentence to contain some specific words; Third, it can tackle the vanishing latent variable problem of VAE (Zhao,

*Correspondence to Jiancheng Lv.
Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Zhao, and Eskenazi 2017). With the gradients obtained from these predictors, we can revise the continuous representation of the original sentence by gradient-based optimization to find a target sentence with the desired fine-grained attributes, and achieve the content-preserving text style transfer.

The method we propose has three major advantages compared to previous methods:

- The method can be *easily* trained on the non-parallel dataset, avoiding the problem of training difficulties caused by adversarial learning and achieving higher performance.
- Unlike previous methods directly generate the target-style sentence through once feed-forward in the inference stage, our method revises the original sentence with gradient information for several steps during inference, which explicitly presents the process of the style transfer and can easily provide us multiple results with tuning the gradients. Therefore, the proposed method has higher *interpretability* and is more *controllable*.
- Most previous text style transfer methods that only control a single binary attribute (e.g., positive and negative sentiments). In contrast, our approach is more *generic* in the sense that it naturally has the ability to control multiple fine-grained attributes, such as sentence length and the existence of specific words.

Extensive experimental comparisons on three popular text style transfer tasks show that the proposed method significantly outperforms five state-of-the-art methods. The source code is available at https://github.com/dayihengliu/Fine-Grained-Style-Transfer.

## 2 Methodology

Let $\mathcal{D} = \{(x^1, s^1), ..., (x^n, s^n)\}$ denotes a dataset which contains $n$ sentences $x^i$ paired with a set of attributes $s^i$. Each $s$ has $k$ attributes of interest $s = \{s_1, ..., s_k\}$. Unlike most previous methods (Shen et al. 2017; Fu et al. 2018; Prabhumoye et al. 2018; Li et al. 2018; Yang et al. 2018) that only consider a single binary attribute (e.g., positive or negative sentiments), our approach naturally has the ability to control multiple fine-grained attributes during style transfer. Here we take two fine-grained attributes, sentence length and the presence of specific words (e.g., a pre-defined subject noun), as the case study. For example, given a original sentence $x$ ="*the salads are fresh and delicious.*", its attribute set can be $s$={sentiment=*positive*, length=*7*, subject_noun=*salads*}. Our task is to learn a generative model $G$ that can generate a new sentence $x^*$ with the required attributes $s^*$, and retain the attribute-independent content of $x$ as much as possible.

### 2.1 Model Structure

The proposed model consists of three components: a variational auto-encoder (VAE), attribute predictors, and a content predictor.

**Variational auto-encoder** $G$. The VAE integrates stochastic latent representation $z$ into the auto-encoder architecture. Its RNN encoder maps a sentence $x$ into a continuous latent representation $z$:

$$z \sim G_{\text{enc}}(\theta_{\text{enc}}; x) = q_E(z|x), \tag{1}$$

and its RNN decoder maps the representation back to reconstruct the sentence $x$:

$$x \sim G_{\text{dec}}(\theta_{\text{dec}}; z) = p_G(x|z), \tag{2}$$

where $\theta_{\text{enc}}$ and $\theta_{\text{dec}}$ denote the parameters of the encoder and decoder. The VAE is then optimized to minimize the reconstruction error $\mathcal{L}_{\text{rec}}$ of input sentences, and meanwhile minimize the KL term $\mathcal{L}_{\text{KL}}$ to encourages the $q_E(z|x)$ to match the prior $p(z)$:

$$\begin{aligned}&\mathcal{L}_{\text{VAE}}(\theta_{\text{enc}}, \theta_{\text{dec}}) = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} \\ &= -\mathbb{E}_{q_E(z|x)}\left[\log p_G(x|z)\right] + \mathcal{D}_{\text{KL}}(q_E(z|x)\|p(z)),\end{aligned} \tag{3}$$

where $\mathcal{D}_{\text{KL}}(\cdot\|\cdot)$ is the KL-divergence. Compared with traditional deterministic auto-encoder, the VAE offers two main advantages in our approach:

(1) Deterministic auto-encoders often have "holes" in their latent space, where the latent representations may not able to generate anything realistic (Roberts et al. 2018). In contrast, by imposing a prior standardized normal distribution $\mathcal{N}(z; 0, I)$ on the latent representations, the VAE learns latent representations not as single isolated points, but as soft dense regions in continuous latent space which makes it be able to generate plausible examples from every point in the latent space (Bowman et al. 2016). This characteristic avoids the problem that the representation $z^*$ revised (optimized) by the gradient not being able to generate a plausible sentence.

(2) This continuous and smooth latent space learned by the VAE enables the sentences generated by adjacent latent representation to be similar in content and semantics (Bowman et al. 2016; Semeniuta, Severyn, and Barth 2017; Goyal et al. 2017; Yang et al. 2017; Shen et al. 2018). Therefore, if we revise the representation $z$ within a reasonable range (i.e., small enough), the resulting new sentence would not differ much in content from the original sentence.

**Attribute predictors** $f_1, ..., f_k$. Each of them takes the representation $z$ as input and predict one attribute $s_j$ of the decoder output sentence $\hat{x}$ generated by $z$. For example, the attribute predictor can be a binary classifier for positive-negative sentiment prediction or a regression model for sentence length prediction. With the gradients provided by the predictors, we can revise the continuous representation $z$ of the original sentence $x$ by gradient-based optimization to find a target sentence $x^*$ with the desired attributes $s^*$.

The attribute predictors $f_1, ..., f_k$ are jointly trained with VAE. For M-classification predictors, we have

$$\mathcal{L}_{\text{Attr},s_j}(\theta_{s_j}, \theta_{\text{enc}}) = -\mathbb{E}_{q_E(z|x)} \log\left[f_j(z)\right], \tag{4}$$

where $f_j(z) = \text{MLP}_j(z) = p(s_j|z) \in \mathbb{R}^{\text{M}}$. And for the regression predictors, we have

$$\mathcal{L}_{\text{Attr},s_j}(\theta_{s_j}, \theta_{\text{enc}}) = \mathbb{E}_{q_E(z|x)}\left[(s_j - f_j(z))^2\right], \tag{5}$$

where $f_j(z) = \text{MLP}_j(z) \in \mathbb{R}^1$. In this joint training, we take the attributes of the input sentence $x$ as the label of predictors. Since the predictor are designed to predict the

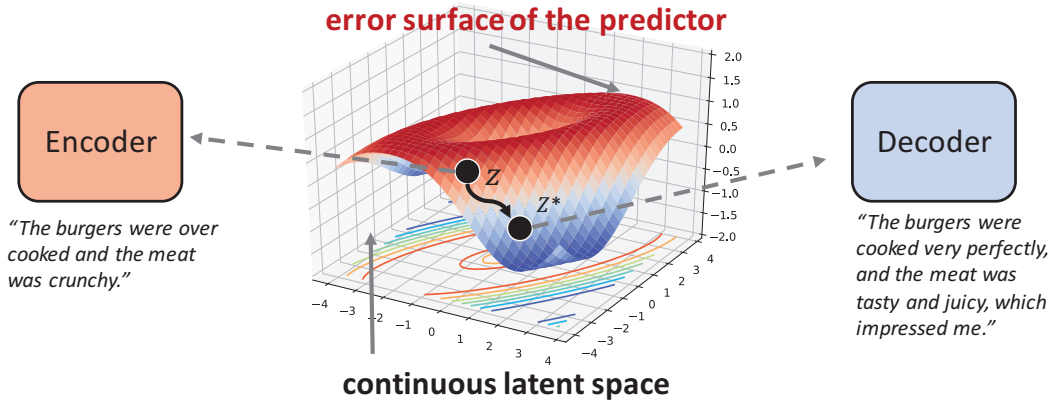


Figure 1: There is an example of content-preserving text sentiment transfer, and we hope to further increase the length of the target sentence compared with the original sentence. The original sentence $x$ with negative sentiment is mapped to continuous representation $z$ via encoder. Then $z$ is revised into $z^*$ by minimizing the error $\mathcal{L}_{\text{Attr},s_1}(\theta_{s_1}; s_1 = \{\text{sentiment} = \textit{positive}\}) + \mathcal{L}_{\text{Attr},s_2}(\theta_{s_2}; s_2 = \{\text{length} = \textit{20}\}) + \lambda_{\text{bow}}\mathcal{L}_{\text{BOW}}(\theta_{\text{bow}}; x_{\text{bow}} = [\textit{burgers}, \textit{meat}])$ with the sentiment predictor $f_1$, length predictor $f_2$, and the content predictor $f_{\text{bow}}$. Afterwards the target sentence $x^*$ is generated by decoding $z^*$ with beam search via decoder [best viewed in color].

attribute of the sentence $\hat{x}$ generated by $z$, we further train each predictor individually after joint training. We sample $z$ from $\mathcal{N}(z; 0, I)$ and feed it into the decoder to generate a new sentence $\hat{x}$. Afterwards we feed $\hat{x}$ into the CNN text classifiers (Kim 2014) which are trained on the training set to predict its attributes[1] as the label of the predictors:

$$\begin{aligned}\mathcal{L}'_{\text{Attr},s_j}(\theta_{s_j}) &= -\mathbb{E}_{p(z)p_G(\hat{x}|z)} \log\left[p(\text{CNN}(\hat{x})|z)\right], \\ \mathcal{L}'_{\text{Attr},s_j}(\theta_{s_j}) &= \mathbb{E}_{p(z)p_G(\hat{x}|z)}\left[(\hat{s}_j - f_j(z))^2\right].\end{aligned} \tag{6}$$

**Content predictor** $f_{\text{bow}}$. It is a multi-label classifier that takes $z$ as input and predicts the Bag-of-Word feature $x_{\text{bow}}$ of its decoder output sentence:

$$f_{\text{bow}}(z) = \text{MLP}_{\text{bow}}(z) = p(x_{\text{bow}}|z). \tag{7}$$

We assume $p(x_{\text{bow}}|z)$ as $|x|$-trial multimodal distribution:

$$\log p(x_{\text{bow}}|z) = \log \prod_{t=1}^{|x|} \frac{e^{f_{\text{bow}}^{(x_t)}}}{\sum_j^{\mathcal{V}} e^{f_{\text{bow}}^{(x_j)}}}, \tag{8}$$

where $\mathcal{V}$ is the size of vocabulary, $|x|$ is the length of $x$, and $f_{\text{bow}}^{(x_j)}$ is the output value of $j$-th word in $f_{\text{bow}} \in \mathbb{R}^{\mathcal{V}}$.

The training of content predictor $f_{\text{bow}}$ is similar to attribute predictors. It is jointly trained with VAE:

$$\mathcal{L}_{\text{BOW}}(\theta_{\text{bow}}, \theta_{\text{enc}}) = -\mathbb{E}_{q_E(z|x)} \log\left[p(x_{\text{bow}}|z)\right]. \tag{9}$$

After joint training, it is trained separately through:

$$\mathcal{L}'_{\text{BOW}}(\theta_{\text{bow}}) = -\mathbb{E}_{p(z)p_G(\hat{x}|z)} \log\left[p(\hat{x}_{\text{bow}}|z)\right]. \tag{10}$$

During text style transfer, we can similarly revise the representation $z$ with the gradient provided by the content predictor $f_{\text{bow}}$ to enhance content preservation. Here we consider two ways to enhance content preservation during style transfer. We can set $x_{\text{bow}}$ to contain all the words in the original sentence $x$, which means that we try to find a sentence $x^*$ with the desired attributes $s^*$ and keep all the words of the original sentence as much as possible to achieve content preservation. However, retaining all the words is often not what we want. For example, $x^*$ should not contain the original emotional words in the task of text sentiment transfer. Instead, the noun in the original sentence should be retained in such a task (Melnyk et al. 2017; Li et al. 2018; John et al. 2019). Therefore, we can set $x_{\text{bow}}$ to contain only all nouns in $x$. Furthermore, we can set $x_{\text{bow}}$ to contain some desired specific words to achieve finer-grained control of target sentences.

**Putting them together**, the final joint training loss $\mathcal{L}$ is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \lambda_b\mathcal{L}_{\text{BOW}} + \lambda_s \sum_{j=1}^{k} \mathcal{L}_{\text{Attr},s_j}, \tag{11}$$

where $\lambda_b$ and $\lambda_s$ are balancing hyper-parameters. It should be noted that $\mathcal{L}_{\text{BOW}}$ and $\mathcal{L}_{\text{Attr},s_j}$ also act as regularizers that prevent the encoder from being trapped into a KL vanishing state (Bowman et al. 2016; Kingma et al. 2016; Yang et al. 2017; Shen et al. 2018; Alemi et al. 2018).

## 2.2 Text Style Transfer

Given the original sentence $x$, the inference process of style transfer is performed in the continuous space. We revise its representation $z$ by gradient-based optimization as follows:

$$\hat{z} = z - \eta(\sum_{j=1}^{k} \nabla_z\mathcal{L}_{\text{Attr},s_j} + \lambda_c\nabla_z\mathcal{L}_{\text{BOW}}), \tag{12}$$

where $\eta$ is the step size and $\lambda_c$ is the trade-off parameter to balance the content preservation and style transfer strength.

[1] Some attributes can be obtained directly without using classifiers, such as the length $\hat{s}_j$ of $\hat{x}$.

We iterate such optimization to find the $z^*$ until the output confidence score of attribute predictors $p(s_j|z)$ is greater than a threshold $\beta$ or reach the maximum number of rounds $T$. The target $x^*$ is obtained by decoding $z^*$ with a beam search (Och and Ney 2004). An example procedure is shown in Figure 1.

## 3 Experiments

In this section, we evaluate the proposed method on three publicly available datasets of sentiment transfer and gender style transfer tasks. Then we conduct several experiments on text sentiment transfer tasks and simultaneously control other fine-grained attributes such as length and keyword presence.

### 3.1 Text Sentiment Transfer

**Data** We use two datasets, Yelp restaurant reviews and Amazon product reviews (He and McAuley 2016)[2], which are commonly used in prior works too (Shen et al. 2017; Fu et al. 2018; Li et al. 2018; Prabhumoye et al. 2018). Following their experimental settings, we use the same pre-processing steps and similar experimental configurations.

**Metrics** There are three criteria for a good style transfer (Li et al. 2018; Prabhumoye et al. 2018). Concretely, the generated sentences should: 1) have the desired attributes; 2) be fluent; 3) preserve the attribute-independent content of the original sentence as much as possible. For the first and second criteria, we follow previous works (Shen et al. 2017; Fu et al. 2018; Li et al. 2018; Prabhumoye et al. 2018) in using model-based evaluation. We measure whether the style is successfully transferred according to the prediction of a pre-trained bidirectional LSTM classifier (Schuster and Paliwal 1997), and measure the language quality by the perplexity (PPL) of the generated sentences with a pre-trained language model. Following previous works, we use the trigram Kneser-Ney smoothed language model (Kneser and Ney 1995) trained on the respective dataset. Since it is hard to measure the content preservation, we follow previous works and report two metrics: 1) Word overlap, which counts the unigram word overlap rate of the original sentence $x$ and the generated sentence $\hat{x}$, computed by $\frac{\text{count}(w_x \cap w_{\hat{x}})}{\text{count}(w_x \cup w_{\hat{x}})}$; 2) As argued in (Melnyk et al. 2017; Li et al. 2018), almost all of the nouns in sentences are attribute-independent content and should be kept in style transfer task, we also calculate the percentage of nouns (e.g., as detected by a POS tagger) in the original sentence appearing in the generated sentence (denoted as Noun%). There are 1000 human annotated sentences as the ground truth of the transferred sentences in (Li et al. 2018). We also take them as references and report the bi-gram BLEU scores (Papineni et al. 2002).

**Baselines** We compare our method with several previous state-of-the-art methods (Shen et al. 2017; Fu et al. 2018; Li et al. 2018; Prabhumoye et al. 2018). We report the results of the human-written sentences as a strong baseline. The results of not making any changes to the original sentences (denoted as Original) are also reported.

**Results** Table 1 shows the evaluation results on two datasets. It should be noted that a good style transfer method should perform well on all metrics as we discussed above. If we only use the original sentence as the output without any modifications, we can still get good performance on both language fluency (PPL) and content retention (Overlap, Noun%) as shown in the first row of Table 1. Therefore, we highlight the metrics where the performances of the models are poor with underline. We find that StyleEmbedding and MultiDecoder achieve high content retention (Overlap, BLEU, and Noun%), but their fluency (PPL) and transfer accuracy are significantly worse than our method. Though the fluency of CrossAligned is better than StyleEmbedding and MultiDecoder, it does not perform well in both content preservation and sentiment transfer. On the contrary, BST achieves high fluency and transfer accuracy, while the content is poorly preserved. Ours (style-strengthen) performs better than BST and CrossAligned on all metrics for these two tasks.

Because the methods proposed in (Li et al. 2018) (except for RetrievalOnly) are based on prior knowledge, which directly revises few words in the original sentence in the discrete space, they can easily achieve high content retention but do not guarantee fluency and accuracy. As shown in the results, their fluency and the transfer accuracy are bad compared to our method. The method RetrievalOnly retrieves the human-written sentence as output, thus this method can achieve high transfer accuracy and fluency, but its content retention is worse than our method. Our methods revise the original sentence in a continuous space, which does well in fluency, content preservation, and transfer accuracy. In addition, our methods can control the trade-off between the transfer accuracy and content preservation.

**Human Evaluation** We conduct human evaluations to further verify the performance of our methods on two datasets further. Following previous works (Li et al. 2018; Fu et al. 2018), we randomly select 50 original sentences and ask 7 evaluators[3] to evaluate the sentences generated by different methods. Each generated sentence is rated on the scale of 1 to 5 in terms of transfer accuracy, preservation of content, and language fluency. The results are shown in Table 3. It can be seen that our models perform well on all metrics and significantly outperform all baselines on the percentage success rate (Suc%) for two datasets.

### 3.2 Text Gender Style Transfer

We use the same dataset[4] as in (Prabhumoye et al. 2018), which contains reviews from Yelp annotated with two sexes (they only consider male or female due to the absence of corpora with other gender annotations). Following (Prabhumoye et al. 2018), we use the same pre-processing steps and

[2]These datasets can be download at http://bit.ly/2LHMUsl.

[3]All evaluators have Bachelor or higher degree. They are independent of the authors' research group.

[4]This dataset can be download at http://tts.speech.cs.cmu.edu/style_models/gender_classifier.tar.

| Methods | Accuracy↑ | PPL↓ | Overlap↑ | Noun%↑ | BLEU↑ |
|---|---|---|---|---|---|
| Original | 0.1 | 22.9 | 100.0 | 100.0 | 42.4 |
| Human | 91.8 | 76.9 | 47.2 | 78.5 | 100.0 |
| Delete, Retrieve, & Generate (Li et al. 2018): | | | | | |
| TemplateBased | 81.3 | 183.6 | 55.6 | **83.3** | 28.9 |
| DeleteOnly | 85.8 | 81.4 | 49.5 | 74.9 | 24.7 |
| DeleteAndRetrieve | 89.5 | 96.1 | 49.4 | 74.0 | 24.9 |
| RetrievalOnly | **98.4** | 25.7 | 15.8 | 39.6 | 4.7 |
| StyleEmbedding (Fu et al. 2018) | 7.2 | 93.9 | **75.4** | 74.2 | **31.9** |
| MultiDecoder (Fu et al. 2018) | 48.8 | 166.5 | 51.5 | 52.2 | 23.1 |
| BTS (Prabhumoye et al. 2018) | 94.8 | 32.8 | 21.5 | 23.5 | 6.8 |
| CrossAligned (Shen et al. 2017) | 73.6 | 72.0 | 41.1 | 42.9 | 18.4 |
| Ours (content-strengthen) | 88.2 | 26.5 | 46.6 | 77.4 | 21.8 |
| Ours (style-content balance) | 92.3 | **18.3** | 38.9 | 69.3 | 18.8 |
| Ours (style-strengthen) | 95.7 | 20.6 | 39.7 | 61.5 | 17.9 |
| Methods | Accuracy↑ | PPL↓ | Overlap↑ | Noun%↑ | BLEU↑ |
| Original | 23.4 | 24.4 | 100.0 | 100.0 | 57.2 |
| Human | 88.1 | 62.9 | 60.5 | 85.0 | 100.0 |
| Delete, Retrieve, & Generate (Li et al. 2018): | | | | | |
| TemplateBased | 69.6 | 108.9 | 73.3 | 87.9 | 42.8 |
| DeleteOnly | 51.6 | 49.3 | **74.4** | **95.1** | **44.7** |
| DeleteAndRetrieve | 55.2 | 48.2 | 69.1 | 92.6 | 41.8 |
| RetrievalOnly | 87.2 | 28.7 | 21.0 | 44.5 | 6.7 |
| StyleEmbedding (Fu et al. 2018) | 40.5 | 87.7 | 42.2 | 41.8 | 22.1 |
| MultiDecoder (Fu et al. 2018) | 66.5 | 80.8 | 30.6 | 30.4 | 14.3 |
| BTS (Prabhumoye et al. 2018) | 82.6 | 25.3 | 24.7 | 22.5 | 9.2 |
| CrossAligned (Shen et al. 2017) | 69.6 | 18.3 | 19.3 | 20.4 | 5.0 |
| Ours (content-strengthen) | 81.9 | 35.0 | 37.7 | 76.0 | 11.5 |
| Ours (style-content balance) | 85.1 | 21.8 | 49.3 | 49.8 | 21.5 |
| Ours (style-strengthen) | **90.0** | **15.9** | 39.5 | 41.4 | 16.3 |

Table 1: Evaluation results of the sentiment transfer tasks on Yelp (Top) and Amazon (Bottom). The notation ↑ means the higher the better, while ↓ means the lower the better. For our models, we report different results (denoted as Ours (content-strengthen), Ours (style-content balance), and Ours (style-strengthen)) corresponding to different choices of hyper-parameters ($\lambda_c$ and $\beta$), which demonstrates our models' ability to control the trade-off between attribute transfer and content preservation. For each evaluation criterion, we bold the best values (except for Human and Original). The accuracies of the classifier on the test set of Yelp and Amazon are 98.2% and 84.0%. Note that a good model should perform well on all metrics, we further highlight the metrics where the performances of the models are poor with underline.

similar experimental configurations. We directly compare our method against BST (Prabhumoye et al. 2018) which has been shown to outperform the previous approach (Shen et al. 2017) on this task. We use the same metrics described in Section Text Sentiment Transfer except for the BLEU score because this dataset does not provide the human annotated sentences. The implementation of BST is based on their source code[5]. The results are shown in Table 4. We can see that our methods outperform BST (Prabhumoye et al. 2018) on all metrics.

### 3.3 Multiple Fine-Grained Attributes Control

To verify our method can also achieve multiple fine-grained attributes control, we take the attributes length, keyword presence, and sentiment as the case study in this experiment. We use the same dataset, Yelp, and the same metrics used in Section Text Sentiment Transfer. For the attribute of length, we design two tasks: 1) We hope that the target sentence can add some relevant content to the original sentence, and increase its length by twice (denoted as Length⇑); 2) We hope that the target sentence can compress the content of the original sentence and reduce its length by half (denoted as Length⇓). For evaluation, we measure the percentage of the length of the generated sentences to the length of the original sentences (denoted as Len%). For the attribute of keyword presence, we hope that the target sentence can contain a pre-defined keyword and retain the content of the original sentence as much as possible (denoted as Keywords). In our experiments, we define a keyword as a noun that is semantically most relevant (computed by the cosine distance of pre-trained word embeddings) to the original sentence but do not appear in the original sentence. The percentage of the generated sentences contain the pre-defined keyword (denoted as Key%) is reported.

[5]https://github.com/shrimai/Style-Transfer-Through-Back-Translation

| Sentiment transfer from **negative** to **positive** (Yelp) | |
|---|---|
| Original | we sit down and we got some really slow and lazy service . |
| Human | the service was quick and responsive . |
| CrossAligned | we went down and we were a good , friendly food . |
| MultiDecoder | we sit down and we got some really and fast food . |
| DeleteAndRetrieve | we got very nice place to sit down and we got some service . |
| BackTranslation | we got and i and it is very nice and friendly staff . |
| Ours (content-strengthen) | we sat down and got some really good service and friendly people . |
| Ours (style-content balance) | we sat down the street and had some really nice and fast service . |
| Ours (style-strengthen) | we really sit down and the service and food were great . |
| Sentiment transfer from **positive** to **negative** (Yelp) | |
| Original | i love this place , the service is always great ! |
| Human | hate this place , service was bad . |
| CrossAligned | i know this place , the food is just a horrible ! |
| MultiDecoder | i love this place , the service is always great ! |
| DeleteAndRetrieve | i did not like the homework of lasagna , not like it , . |
| BackTranslation | i wish i have been back , this place is a empty ! |
| Ours (content-strengthen) | however , this place is the worst i have ever been to . |
| Ours (style-content balance) | i do n't know why i love this place , but the service is horrible . |
| Ours (style-strengthen) | i do n't know why this place has the worst customer service ever . |

Table 2: Samples of the sentiment transfer task from ours and baselines on Yelp. The Original denotes the input sentence, and the Human denotes the human annotated sentence. The samples of the sentiment transfer from negative to positive and positive to negative are shown in top and bottom, respectively.

| | Yelp | | | | Amazon | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc | Gra | Con | Suc% | Acc | Gra | Con | Suc% |
| Human | 4.1 | 4.4 | 3.6 | 78 | 3.5 | 4.3 | 3.9 | 60 |
| CrossAligned (Shen et al. 2017) | 3.3 | 2.9 | 2.6 | 22 | 3.0 | 3.3 | 1.6 | 6 |
| MultiDecoder (Fu et al. 2018) | 2.4 | 3.0 | 3.1 | 12 | 2.3 | 2.7 | 2.5 | 6 |
| BTS (Prabhumoye et al. 2018) | **3.9** | 3.7 | 1.8 | 26 | 2.8 | 3.3 | 1.8 | 8 |
| DeleteAndRetrieve (Li et al. 2018) | 3.8 | 3.6 | **3.5** | 54 | 2.4 | 3.5 | **3.8** | 28 |
| Ours (content-strengthen) | 3.6 | 4.1 | 3.1 | 66 | 3.4 | 4.0 | 2.8 | 42 |
| Ours (style-content balance) | 3.7 | **4.3** | 3.2 | **72** | 3.7 | 4.0 | 2.4 | 40 |
| Ours (style-strengthen) | 3.8 | 4.1 | 3.0 | 60 | **3.8** | **4.5** | 2.5 | **50** |

Table 3: Human evaluation results of the sentiment transfer tasks on Yelp and Amazon. We show average human ratings for transfer accuracy (Acc), preservation of content (Con), and fluency of sentences (Gra) on 1 to 5 score. "Suc%" denotes the overall percentage success rate. Similar to previous works, we consider a generated output "successful" if it is rated no less than 3 on all three criteria (Att, Con, Gra).

| Methods | Accuracy↑ | PPL↓ | Overlap↑ | Noun%↑ |
|---|---|---|---|---|
| Orginal | 21.9 | 183.4 | 100.0 | 100.0 |
| BTS (Prabhumoye et al. 2018) | 60.3 | 145.0 | 37.9 | 35.3 |
| Ours (content-strengthen) | 70.6 | 98.2 | 46.8 | **69.6** |
| Ours (style-content balance) | 71.3 | 87.8 | **51.8** | 57.5 |
| Ours (style-strengthen) | **79.9** | **78.9** | 46.4 | 53.8 |

Table 4: Evaluation results of the gender transfer task on Yelp. For our models, we report different results corresponding to different choices of hyper-parameters ($\lambda_c$ and $\beta$) to demonstrate our models' ability to control the trade-off between attribute transfer and content preservation. The accuracy of the classifier on the test set is 83.1%.

The results are shown in Table 5. For a single fine-grained attribute, it can be observed that Keywords achieves 92.3 Key% score, Length⇑ and Length⇓ achieve 208.8 and 40.8 Len% scores respectively. At the same time, the fluency and content retention scores are still high. These results demonstrate the proposed method can control such fine-grained attributes. When we further control the sentiment attribute, we can see that Sentiment + Keywords achieves 91.6% accuracy,

| Methods | Accuracy↑ | PPL↓ | Overlap↑ | Noun%↑ | Len% | Key%↑ |
|---|---|---|---|---|---|---|
| Original | 0.1 | 22.9 | 100.0 | 100.0 | 100.0 | 7.8 |
| Keywords | 16.7 | 43.9 | **39.2** | 56.0 | 98.1 | **92.3** |
| Sentiment + Keywords | 91.6 | 52.6 | 24.5 | 42.4 | 106.0 | 78.3 |
| Length⇑ | 0.2 | 29.8 | 25.0 | 48.3 | **208.8** | 5.9 |
| Sentiment + Length⇑ | **97.7** | 25.4 | 21.4 | 51.7 | 189.5 | 9.2 |
| Keywords + Length⇑ | 25.6 | 44.5 | 29.8 | **61.8** | 165.0 | 83.2 |
| Sentiment + Keywords + Length⇑ | 93.0 | 51.8 | 18.8 | 50.0 | 183.7 | 66.6 |
| Length⇓ | 0.2 | 31.3 | 30.7 | 25.2 | **40.8** | 6.3 |
| Sentiment + Length⇓ | 95.1 | **23.0** | 29.1 | 38.1 | 66.9 | 6.7 |
| Keywords + Length⇓ | 21.4 | 87.0 | 28.4 | 38.9 | 61.6 | 83.7 |
| Sentiment + Keywords + Length⇓ | 87.6 | 123.8 | 16.3 | 23.7 | 60.9 | 63.0 |

Table 5: Results of fine-grained Attributes control on the Yelp. Different rows correspond to the set of attributes being controlled by the model.

while the accuracy of Sentiment + Length⇑ and Sentiment + Length⇓ is 97.7% and 95.1% respectively. Meanwhile, their rest scores have not declined significantly. When simultaneously controlling all these attributes, Sentiment + Keywords + Length⇑ achieves 93.0% accuracy, 183.7 Len% score, and 66.6 Key% score, while Sentiment + Keywords + Length⇓ achieves 87.6% accuracy, 60.9 Len% score, and 63.0 Key% score. Since it is more difficult to reduce sentence length than to increase sentence length while controlling other attributes, the fluency of Sentiment + Keywords + Length⇓ is worse than Sentiment + Keywords + Length⇑. These results indicate that our proposed method can control multiple attributes simultaneously.

## 4 Related Works

We have witnessed an increasing interest in text style transfer under the setting of non-parallel data. Most such methods explicitly disentangle the content and the attribute. One line of research leverages the auto-encoder framework to encode the original sentence into an attribute-independent content representation with adversarial learning, which is then fed into the decoder with a style vector to output the transferred sentence. In (Hu et al. 2017; Shen et al. 2017; Prabhumoye et al. 2018), adversarial learning is utilized to ensure that the output sentence has the desired style. In order to disentangle the content and the attribute, (Hu et al. 2017) enforces the output sentence to reconstruct the content representation, while (Fu et al. 2018; Zhao, Zhao, and Eskenazi 2017; John et al. 2019) apply adversarial learning to discourage encoding style information into the content representation. (Shen et al. 2017) utilizes adversarial learning to align the generated sentences from one style to the data domain of the other style. In (Yang et al. 2018), the authors extend the cross-align method (Shen et al. 2017) by employing a language model as the discriminator, which can provide a more stable and more informative training signal for adversarial learning.

However, as argued in (Li et al. 2018; Guillaume Lample 2019), it is often easy to fool the discriminator without actually learning the representations that are disentangled. Unlike the methods mentioned above that disentangle the content and the attribute with adversarial learning, another line of research (Prabhumoye et al. 2018; Logeswaran, Lee, and Bengio 2018; Guillaume Lample 2019) applies back-translation (Wintner et al. 2016) to rephrase a sentence while reducing the stylistic properties and encourage content compatibility. Besides, the authors in (Li et al. 2018) directly mask out the words associated with the original style of the sentence to obtain the attribute-independent content text. Instead of revising the sentence in the discrete space with prior knowledge as in (Li et al. 2018), our method maps the discrete sentence into a continuous representation space and revises the continuous representation with the gradient provided by the predictors. This method does not explicitly disentangle the content and the attribute and avoids the training difficulties caused by the use of adversarial learning in the previous methods. Similar ideas have been proposed in (Mueller, Gifford, and Jaakkola 2017; Luo et al. 2018) for sentence revision and neural architecture search. As pointed out in (Shen et al. 2017), the model proposed in (Mueller, Gifford, and Jaakkola 2017) does not necessarily enforce content preservation, while our method employs a content predictor to enhance content preservation. Furthermore, unlike most previous methods that only control a single binary attribute (e.g., positive and negative sentiments), our approach can further control multiple fine-grained attributes such as sentence length and the existence of specific words. Note that controlling such fine-grained attributes has already been studied in the previous works for other tasks (Post and Vilar 2018; Makino et al. 2019), which only serves as a case study to demonstrate the generality of our method.

## 5 Conclusion and Future Work

In this paper, we propose a new framework for unsupervised text style transfer which revises the original sentences in a continuous space based on gradient optimization in the inference stage. Compared with previous adversarial learning based methods, our method is easy to train, interpretable, and more controllable. Extensive experiments on three popular text style transfer tasks show that our approach outperforms five previous state-of-the-art methods. Furthermore, experimental results demonstrate that the proposed method can simultaneously manipulate multiple fine-grained attributes such as sentence length and the presence of specific words.

In future work, we plan to explore control over other fine-

grained attributes. In addition, it would be interesting to extend the proposed approach to other natural language generation tasks, such as dialogue and headline generation.

## Acknowledgment

This work is supported by National Key R&D Program of China under contract No. 2017YFB1002201 and supported by National Natural Science Fund for Distinguished Young Scholar (Grant No. 61625204) and partially supported by the Key Program of National Science Foundation of China (Grant No. 61836006).

## References

Alemi, A.; Poole, B.; Fischer, I.; Dillon, J.; Saurous, R. A.; and Murphy, K. 2018. Fixing a broken elbo. In *ICML*.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Jozefowicz, R.; and Bengio, S. 2016. Generating sentences from a continuous space. In *CoNLL*.

Fu, Z.; Tan, X.; Peng, N.; Zhao, D.; and Yan, R. 2018. Style transfer in text: Exploration and evaluation. In *AAAI*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

Goyal, A. G. A. P.; Sordoni, A.; Côté, M.-A.; Ke, N. R.; and Bengio, Y. 2017. Z-forcing: Training stochastic recurrent networks. In *NIPS*.

Guillaume Lample, Sandeep Subramanian, E. M. S. L. D. M. R. Y.-L. B. 2019. Multiple attribute text rewriting. In *ICLR*.

He, R., and McAuley, J. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward controlled generation of text. In *ICML*.

John, V.; Mou, L.; Bahuleyan, H.; and Vechtomova, O. 2019. Disentangled representation learning for text style transfer. In *AAAI*.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. In *ICLR*.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved variational inference with inverse autoregressive flow. In *NIPS*.

Kneser, R., and Ney, H. 1995. Improved backing-off for m-gram language modeling. In *ICASSP*.

Kusner, M. J., and Hernández-Lobato, J. M. 2016. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*.

Li, J.; Jia, R.; He, H.; and Liang, P. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL-HLT*.

Logeswaran, L.; Lee, H.; and Bengio, S. 2018. Content preserving text generation with attribute controls. In *NeurIPS*.

Luo, R.; Tian, F.; Qin, T.; Chen, E.; and Liu, T.-Y. 2018. Neural architecture optimization. In *NeurIPS*.

Makino, T.; Iwakura, T.; Takamura, H.; and Okumura, M. 2019. Global optimization under length constraint for neural text summarization. In *ACL*.

Melnyk, I.; Santos, C. N. d.; Wadhawan, K.; Padhi, I.; and Kumar, A. 2017. Improved neural text attribute transfer with non-parallel data. In *NIPS (Workshop)*.

Mueller, J.; Gifford, D.; and Jaakkola, T. 2017. Sequence to better sequence: continuous revision of combinatorial structures. In *ICML*.

Och, F. J., and Ney, H. 2004. The alignment template approach to statistical machine translation. *Computational linguistics*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Post, M., and Vilar, D. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *NAACL*.

Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; and Black, A. W. 2018. Style transfer through back-translation. In *ACL*.

Roberts, A.; Engel, J.; Raffel, C.; Hawthorne, C.; and Eck, D. 2018. A hierarchical latent vector model for learning long-term structure in music. In *ICML*.

Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*.

Semeniuta, S.; Severyn, A.; and Barth, E. 2017. A hybrid convolutional variational autoencoder for text generation. In *EMNLP*.

Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. 2017. Style transfer from non-parallel text by cross-alignment. In *NIPS*.

Shen, X.; Su, H.; Niu, S.; and Demberg, V. 2018. Improving variational encoder-decoders in dialogue generation. In *AAAI*.

Sutton, R. S.; McAllester, D. A.; Singh, S. P.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*.

Wintner, S.; Mirkin, S.; Specia, L.; Rabinovich, E.; and Patel, R. N. 2016. Personalized machine translation: Preserving original author traits. In *EACL*.

Yang, Z.; Hu, Z.; Salakhutdinov, R.; and Berg-Kirkpatrick, T. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *ICML*.

Yang, Z.; Hu, Z.; Dyer, C.; Xing, E. P.; and Berg-Kirkpatrick, T. 2018. Unsupervised text style transfer using language models as discriminators. In *NeurIPS*.

Zhang, Y.; Gan, Z.; Fan, K.; Chen, Z.; Henao, R.; Shen, D.; and Carin, L. 2017. Adversarial feature matching for text generation. In *ICML*.

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*.