"Slow Service" ⇏ "Great Food": Enhancing Content Preservation in Unsupervised Text Style Transfer

Anonymous ACL submission

Abstract

Text style transfer aims to change the style (e.g., sentiment, politeness) of a sentence while preserving its content. A common solution is the prototype editing approach, where stylistic tokens are deleted in the "mask" stage and then the masked sentences are infilled with the target style tokens in the "infill" stage. Despite their success, these approaches still severely suffer from the content preservation problem. By closely inspecting the errors made by existing approaches, we identify two common types of issues: 1) many content-related tokens are masked and 2) irrelevant words associated with the target style are infilled. Our paper aims to enhance content preservation by tackling each of them. In the "mask" stage, we utilize a BERT-based keyword extraction model that incorporates syntactic information to prevent content-related tokens from being masked. In the "infill" stage, we create a psuedo-parallel dataset and train a T5 model to infill the masked sentences without introducing irrelevant content. Empirical results show that our method outperforms the state-of-the-art baselines by a large margin in terms of content preservation, while maintaining comparable transfer effectiveness and language quality.

1 Introduction

potential risks

There has been a recent surge of interest in text style transfer, with the aim of altering the text style (e.g., sentiment, politeness, formality) of a sentence while preserving its content. For example, a sentiment transfer model may transfer a positive-sentiment sentence from "This is the best book I've read ever!" to "This is the worst book I've read ever!". As another example, "what happened to my personal station?" may be transferred to "could you please let me know what happened to my personal station?" for a more polite expression. Text style transfer has been shown to be useful in many downstream applications, such as author obfuscation (Shetty et al., 2018), data augmentation (Xie et al., 2020; Kaushik et al., 2019), text simplification (Xu et al., 2015), and writing assistance (Heidorn, 2000).

Unsupervised style transfer has been extensively explored since parallel data are difficult to obtain. One intuitive and promising solution is the prototype editing approach (Li et al., 2018; Wu et al., 2019; Reid and Zhong, 2021), where the "mask" and "infill" steps are sequentially applied. In the "mask" stage, stylistic tokens are identified and deleted by frequency-ratio based methods (e.g., TF-IDF) and/or attention-based methods, resulting in a content-only masked sentence. In the "infill" stage, the masked sentence is infilled by adding new style markers through template-based methods (e.g., Li et al. (2018)) or masked language models (e.g., BERT) (Wu et al., 2019; Malmi et al., 2020).

While these models have shown their power to transfer the input text to the target style with high transfer effectiveness, most of them, if not all, suffer from the content preservation issue. As shown in Table 1, despite the style has been transferred successfully, the content is changed too (e.g., "service" → "food").

In this paper, we propose a novel approach to enhance **content preservation** for *unsupervised text style transfer*. We first summarize two important observations of common errors made by the existing models:

- In the "mask" stage, content-related tokens may be removed (e.g., cases (a), (c), (d), (e) in Table 1);
- In the "infill" stage, irrelevant words with strong styles may be generated (e.g., (a), (b), (d), (e) in Table 1).

To preserve content-related tokens in the "mask" stage, we highlight the central component of the

Transfer Type	Source Sentences	Transferred Sentences
(a) Negative → Positive:	we sit down and we got some really slow and lazy service.	we sit down and we got some really good food and loved it.
(b) Positive → Negative:	the taste is awesome.	the taste is not good and the service is slow.
(c) Factual → Romantic:	a man and a woman show their tatooed hearts on their wrists.	a man and a woman show their loved hearts on their anniversary.
(d) Male → Female:	the locker room is clean.	the locker room is cute.
(e) Toxic → Civil:	as stupid and arrogant as his boss.	as warm hearted as his boss.

Table 1: Error analysis of existing state-of-the-art models. Tokens masked are in red, and new tokens generated are in blue. Tokens underlined are either content-related tokens removed or irrelevant words generated.

sentence and prevent them to be masked. Specifically, we utilize a BERT-based keyword extraction model which incorporates syntactic information (e.g., dependency parsing) to identify content-related tokens. In dependency parsing, the head word of a constituent was the central organizing word of a larger constituent (e.g., the primary noun in a noun phrase, or verb in a verb phrase) (Jurafsky, 2000), and therefore, should be more likely to remain unmasked. Lastly, we apply an attention network to determine what tokens should be masked. In the style classification task, an attention score could be interpreted as to what extent a token has style attribute.

In the "infill" stage, existing state-of-the-art approaches typically fine-tune a large pre-trained masked language model (e.g., BERT) on the target style corpus and treat it as a fill-in-the-mask problem (Wu et al., 2019; Malmi et al., 2020; Reid and Zhong, 2021). While such language models can generate fluent sentences of the target style well, they often introduce tokens irrelevant to the source sentence, which results in the change of content. To prevent irrelevant words generation in the "infill" stage, we create a psuedo-parallel dataset and train a large pre-trained language model — T5 (Raffel et al., 2020) to specifically learn to generate from a masked sentence to a target style sentence without introducing unnecessary and irrelevant content.

To summarize, we make the following contributions to enhance content preservation in unsupervised text style transfer:

- In the "mask" stage, we utilize a BERT-based keyword extraction model and leverage dependency parsing information to preserve content-related tokens.
- In the "infill" stage, we propose to create a psuedo-parallel dataset in a self-supervised manner, and explicitly learn to recover the masked sentences in the target style without adding irrelevant content.
- While existing papers demonstrate good performance, most of them report results on 2-3 datasets only, which raises concerns on the generalizability of their models. In our experiment, we conduct experiments on five diverse benchmark datasets, and consistently demonstrate that our approach outperforms the state-of-the-art models in content preservation, while maintaining satisfactory transfer effectiveness and language quality.

2 Proposed Model

2.1 Problem Formulation

In this paper, we formulate the unsupervised text style transfer as follows: for two non-parallel corpora $\mathbf{X} = \{x_1, x_2, ..., x_m\}$ and $\mathbf{Y} = \{y_1, y_2, ..., y_n\}$ with styles S_x and S_y respectively, the task aims at training a style transfer model G that generates samples $\hat{\mathbf{X}} = \{\hat{x}_1, \hat{x}_2, ..., \hat{x}_m\}$ in the target style S_y, conditioned on samples in $\mathbf{X}$.

2.2 Model Overview

Figure 1 illustrates our proposed model architecture. Following Li et al. (2018); Wu et al. (2019), we assume that style is localized to certain tokens in a sentence and we can delete those tokens to obtain a style-free corrupted sentence.

At the **training** stage, we first build a style removal model G_d to obtain corrupted sentences $\mathbf{Y}_c$ from $\mathbf{Y}$, the collection of sentences in the target corpus.[1] Such corrupted sentences $\mathbf{Y}_c$ are considered style-free under our aforementioned assumption, and ideally there is no loss of content. Second,

[1] "Corrupted sentences" and "masked sentences" are used interchangeably.

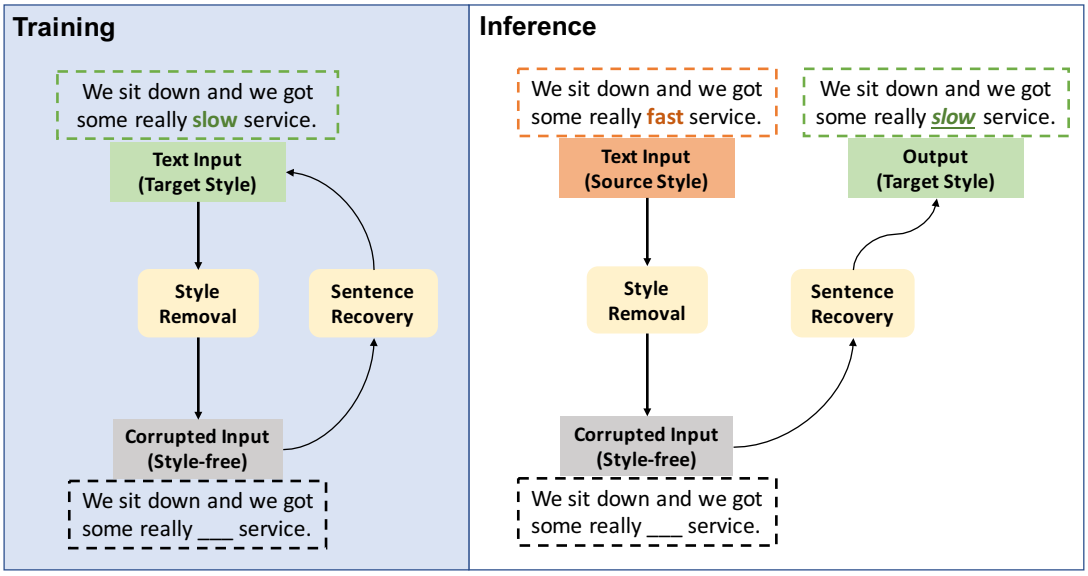

Figure 1: Overview of the model architecture.

we train a sentence recovery model G_r to recover the original sentences $\mathbf{Y}$ from the corrupted sentences $\mathbf{Y}_c$. Such a sentence recovery model G_r is expected to recover the style-free corrupted sentences $\mathbf{Y}_c$ to the original sentences $\mathbf{Y}$ in the target style S_y, and very importantly, without introducing irrelevant content.

After training, we have a style removal model G_d and a sentence recovery model G_r. Now at the **inference** stage, we apply the style removal model G_d on the source style sentences $\mathbf{X}$ and obtain style-free corrupted sentences $\mathbf{X}_c$. Then, we produce the final output $\hat{\mathbf{X}}$ using the sentence recovery model G_r, which is trained to recover corrupted sentences to the target style S_y.

In Sections 2.3 and 2.4, we will introduce the details of the style removal model G_d and the sentence recovery model G_r.

2.3 The Style Removal Model

Existing models typically make use of frequency-ratio based methods (e.g., TF-IDF) and/or attention based methods to remove the stylistic tokens. However, they achieve mediocre performance as many content-related and style-free tokens are masked too. Section 2.3.1 explains how content-related tokens are preserved and Section 2.3.2 shows how the style-related tokens are masked.

2.3.1 Keyword Extraction

To preserve the relevant content, we explicitly utilize a keyword extraction model, which incorporates syntactic information (e.g., dependency parsing) to highlight the content-related tokens and prevent them from being removed.

With a source style sentence $x = \{t_1, t_2, ..., t_k\}$, where t_i is the i-th token, the model extracts content-related keywords in three steps:

(a) **Embedding**: we use BERT embeddings[2] to represent all of the keywords $e_{t_1}, e_{t_1}, ..., e_{t_k}$ and the entire sentence e_x in a high-dimensional vector space;

(b) **Dependency Parsing**: we construct a dependency tree that captures word-level relations with the Stanford dependency parser (Manning et al., 2014). From the dependency tree, we obtain the depth d_i and the outdegree o_i for each word token t_i. In dependency parsing, the head word of a constituent was the central organizing word of a larger constituent (Jurafsky, 2000). The more central the words are (higher depth or larger outdegree), the more likely it contains meaningful content and therefore, the less likely they should be masked.

(c) **Ranking**: all candidates are ranked to represent the keywords of the sentence:

$$r_{t_i} = \alpha \cdot \cos(e_{t_i}, e_x) + \beta \cdot d_i + \gamma \cdot o_i$$

To alleviate the redundant keywords issue, we follow Bennani-Smires et al. (2018) to use Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) for diversified candidates by optimizing keyword informativeness with dissimilarity among selected candidates.

[2]We use "bert-base-uncased" in https://huggingface.co/docs/transformers/model_doc/bert.

Finally, we select candidates over a threshold $thres$ and prevent them from being masked. Empirically, we take $\alpha = 0.8$, $\beta = 0.1$, $\gamma = 0.1$, and $thres = 0.74$, based on the results of the validation dataset in the Yelp dataset.

2.3.2 Attention

After the keywords have been extracted, we train an attention-based classifier to identify the style-related tokens. Given a sentence x of k tokens $\{t_1, t_2, ..., t_k\}$, we encode the sentence and concatenate the forward and the backward hidden states for each word with a bidirectional LSTM. The hidden states are represented as $\mathbf{H} = (h_1, h_2, ..., h_k)$. We apply an attention network, which produces attention weights a and weighted hidden states c. Finally, the hidden states c is transformed to the probability distribution y with a softmax layer:

$$a = softmax(\mathbf{w} \cdot tanh(\mathbf{W}\mathbf{H}^T))$$

$$c = a \cdot \mathbf{H}$$

$$y = softmax(\mathbf{W}' \cdot c)$$

where $\mathbf{w}$, $\mathbf{W}$, $\mathbf{W}'$ are learnable parameters.

After training, the attention-based classifier can be used to extract attention weights, which capture the style information of each word. For simplicity, we follow Wu et al. (2019) and set the averaged attention value in a sentence as the threshold. Words with attention weights higher than the threshold are viewed as style markers. Note that the content-related keywords identified in Section 2.3.1 are preserved and **not** marked as style markers.

2.4 The Sentence Recovery Model

With style-free corrupted sentences $\mathbf{X}_c$, we focus on recovering them in the target style S_y. Here, we introduce to solve the problem by creating a pseudo-parallel training dataset and training a model G_r for sentence recovery explicitly. Recall that in Section 2.3, we obtain corrupted sentences $\mathbf{Y}_c$ given the original sentences $\mathbf{Y}$. Therefore, if we take them in a reverse direction, we then have a parallel training dataset to learn from (i.e., $\mathbf{Y}_c$ –> $\mathbf{Y}$).

We select T5 (Raffel et al., 2020), a strong pre-trained text-to-text model, as the base architecture, and fine-tune it on the constructed psuedo-parallel dataset. After being trained, the model is expected to take as input a corrupted style-free input $\mathbf{Y}_c$ and generate sentences in the target style without introducing additional irrelevant content. Finally, we apply the trained T5 model on corrupted input $\mathbf{X}_c$ and generate the final output $\hat{X}$, which is expected to be of the target style S_y.

Intuition: As demonstrated by Wu et al. (2019); Malmi et al. (2020), it is an intuitive idea to treat this as a fill-in-the-mask problem, and generate sentences by a fine-tuned masked language model. However, such masked language models (e.g., BERT) are designed to predict tokens for a "mask" and generate sentences with the highest sentence probability. Despite that they are able to generate fluent sentences in the target style, they may introduce tokens that are irrelevant to the source sentence (e.g., case (b) in Table 1) and therefore, may potentially change the content. Here, what we expect is not a generic model for generating a fluent sentence, but rather a specialized model that works only for *sentence recovery without introducing irrelevant content*. Therefore, we construct a pseudo-parallel training dataset and train the model explicitly for such a task. After training on such a dataset, the T5 model is expected to learn specifically to generate sentences in the target style without introducing additional and irrelevant information.

Dataset	Style	Train	Valid	Test
Yelp	Positive	270K	2K	500
	Negative	180K	2K	500
Amazon	Positive	277K	985	500
	Negative	278K	1015	500
Captions	Romantic	6K	300	-
	Humorous	6K	300	-
	Factual	-	-	300
Politeness	Polite	219K	28K	-
	Impolite	199K	24K	800
Detoxification	Toxic	150K	5K	10K
	Non-toxic	150K	5K	-

Table 2: Dataset statistics for style transfer tasks.

3 Empirical Evaluation

In this section, we empirically evaluate the performance of our proposed approach (denoted as "STEC"[3]) and a set of baseline models.

3.1 Datasets

Sentiment Transfer: We use the Yelp dataset and the Amazon dataset (Li et al., 2018), which are

[3]short for "Style Transfer with Enhanced Content"

business reviews on Yelp and product reviews on Amazon respectively. Each of the dataset consists of two non-parallel corpora with positive and negative sentiments. Each example is labeled as having either positive or negative sentiment.

Captions: The Captions dataset (Gan et al., 2017; Li et al., 2018) has image captions labeled as being factual, romantic or humorous. We focus on the task of converting factual sentences into romantic and humorous ones.

Politeness: The Politeness dataset (Madaan et al., 2020) is produced by filtering through the Enron Email corpus (Klimt and Yang, 2004). We aim to transform the tone of a sentence from impolite to polite.

Detoxification: We employed the largest publicly available toxicity detection dataset to date from "Jigsaw Unintended Bias in Toxicity Classification" Kaggle challenge.[4] We follow Dale et al. (2021) to obtain non-parallel data, and focus on transferring from toxic to non-toxic.

Dataset statistics are presented in Table 2. For the Yelp, Amazon and Captions datasets, human annotated solutions are also provided for measuring content preservation.

3.2 Baselines

We compare our proposed approach with the following competitive baseline models:

1. CAE: it achieves style transfer from style transfer from nonparallel text by cross alignment (Shen et al., 2017).
2. DRG (Li et al., 2018): this is the first successful work that brought the prototype editing methods to attention. We compare against the full method — delete-retrieve-generate.
3. Mask and Infill (MI) (Wu et al., 2019): the styles are first separated from content by masking the positions of sentimental tokens with a fusion model. Then, a masked language model is retrofitted to predict words/phrases conditioned on the context and the target sentiment.
4. Tag and Generate (TAG) (Madaan et al., 2020): it first tags tokens with the original style and/or adds new tags inside a sentence. Then, it conditionally generates the target sentence from style-agnostic tagged source sentence.
5. NAST (Huang et al., 2021): it first predicts word alignments conditioned on the source sentence, and then generates the transferred sentence with a non-autoregressive decoder. We report results by the model building upon StyTrans (Dai et al., 2019).
6. RACoLN (Lee et al., 2021): it implicitly removes style at the token level using reverse attention, and fuses content information to style representation using conditional layer normalization.

3.3 Evaluation

Following prior work (Madaan et al., 2020; Reid and Zhong, 2021), we evaluate all model outputs along three dimensions: transfer effectiveness, content preservation and language quality. *Transfer effectiveness* refers to whether the transferred sentences reveal the target style property. *Content preservation* captures how a sentence maintains its content throughout the transfer process. *Language quality* measures whether the generated sentences are grammatical, fluent and readable.

3.3.1 Automatic Evaluation

Effectiveness: We follow Reid and Zhong (2021) and train a RoBERTa-base classifier on the training data for the respective dataset. Our evaluation classifier achieves accuracy of 98.0% on Yelp, 84.2% on Amazon, 79.6% on Captions, 88.3% on Politeness, and AUC-ROC of 0.97 on Detoxification. We measure the percentage of the generated sentences classified to be in the target domain by the classifier.

Content Preservation: The standard metric for measuring content preservation is BLEU-self (BL-s) (Papineni et al., 2002) which is compared with respect to the original sentences. However, BLEU scores can measure syntactic content preservation only. In addition, to measure semantic content preservation, we report BERTScore-self (BS-s) (Zhang et al., 2019) against the source sentences. Besides, we report BLEU-reference (BL-r) and BERTScore-reference (BS-r) using the human reference sentences on the Yelp, Amazon and Captions datasets (Li et al., 2018).

Language Quality: We adopt GRUEN (Zhu and Bhat, 2020) to evaluate the language quality.

3.3.2 Human Evaluation

In addition to automatic evaluation, we validate the generated outputs with human evaluation. With

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

each model except CAE, we randomly sample 100 outputs from each dataset.[5] Given the target style and the original sentence, two annotators (graduate students who are specialized in NLP) are asked to evaluate the model generated sentence with a score range from 1 (Very Bad) to 5 (Very Good) on style transfer accuracy, content preservation, and language quality.

3.4 Results

The automatic evaluation results based on best-found hyperparameters are summarized in Table 3. We observe a significant improvement in content preservation scores across various datasets (specifically in the Captions dataset and the Detoxification dataset), highlighting the ability of our model to retain content better than the baseline models. Alongside, we observe comparable performance of our model on transfer effectiveness and language quality.

As for the human evaluation, we report the average scores from the 2 annotators in Table 4. We observe that the result mainly conforms with the automatic evaluation. Our model received the highest score on the content evaluation metric, while maintaining comparable score on transfer effectiveness and language quality. Both automatic and human evaluation depict the strength of our proposed model in preserving content.

Among all the baselines, TAG has the best performance consistently in both automatic evaluation and human evaluation, in particular, on the Politeness dataset. This is expected as the "tagger" component is designed to find place for insertion of polite expressions inside a sentence.

For the two state-of-the-art papers that tackles content preservation — RACoLN and NAST, though they perform well on some datasets, the models are not robust across different datasets. Comparably, our approach has consistently good performance and therefore, demonstrates its superior generalizability.

3.5 Ablation Study

We compare with the following ablations of STEC and show the results in Figure 2:

1. no-parsing: we exclude the dependency parsing information and use BERT embeddings only to preserve the keywords.

[5]We excluded CAE for human evaluation because it performs poorly as determined by the automatic evaluation.

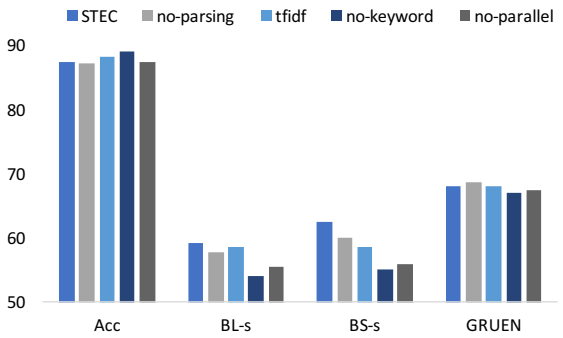

Figure 2: Ablation study. Plots show average results across all five datasets. We scale GRUEN by 100 times for better visualization.

2. tfidf: instead of using the attention network for masking the style-related works, we follow (Li et al., 2018) to use the TF-IDF to mask the style-related words.

3. no-keyword: we exclude the entire keyword extraction model and use the attention network directly to mask the style-related words.

4. no-parallel: instead of constructing a psuedo-parallel dataset and train the T5 model in the "infill" stage, we treat it as a fill-in-the-mask problem and solve it by a fine-tuned masked language model.

We observe that our approach performs better than all ablations in terms of content preservation, and all ablations have comparable performance for transfer effectiveness and language quality. Compared with no-keyword and no-parallel, we conclude that each of the proposed model (i.e., Section 2.3 and Section 2.4) contributes to content preservation well respectively. Besides, by comparing no-keyword and no-parsing, we demonstrate that dependency parsing information can help preserve the content too. In addition, the performance drop by tfidf indicates that an attention network works better in masking stylistic tokens.

3.6 Case Study

Examples of the transferred results by our model are presented in Table 5. We find that our proposed keyword extraction model can preserve the content-related words well. Besides it, we also observe that the T5 model is able to recover the corrupted sentences in the target style without introducing irrelevant content.

	Yelp						Amazon					
	Acc	**BL-s**	**BL-r**	**BS-s**	**BS-r**	**GR**	**Acc**	**BL-s**	**BL-r**	**BS-s**	**BS-r**	**GR**
CAE	73.6	20.2	7.7	33.6	22.9	0.69	**78.0**	2.6	1.7	9.8	6.9	0.51
DRG	88.5	36.7	14.5	48.5	33.3	0.72	51.2	57.1	29.9	66.9	46.2	0.62
MI	90.5	41.7	15.3	49.8	36.0	0.75	74.5	60.0	28.5	61.2	44.7	0.62
TAG	85.8	47.1	19.7	57.9	37.2	**0.78**	66.4	**68.7**	34.8	**69.5**	48.2	**0.66**
NAST	89.4	59.0	21.0	55.8	45.9	0.72	64.1	55.8	27.9	61.7	39.9	0.59
RACoLN	**91.3**	58.9	20.0	62.1	42.1	0.75	69.1	31.9	20.1	36.9	31.1	0.63
STEC	89.2	**60.2**	**22.7**	**68.9**	**48.6**	0.75	68.2	67.1	**36.5**	68.8	**50.9**	**0.66**

(a) Sentiment transfer.

	Captions						Politeness				Detoxification			
	Acc	**BL-s**	**BL-r**	**BS-s**	**BS-r**	**GR**	**Acc**	**BL-s**	**BS-s**	**GR**	**Acc**	**BL-s**	**BS-s**	**GR**
CAE	89.7	2.1	1.6	11.2	6.7	0.51	**99.4**	7.0	30.7	0.71	92.3	13.4	22.9	0.52
DRG	**95.7**	31.8	11.8	40.2	28.4	0.58	90.3	11.8	41.4	0.69	95.6	38.5	42.7	0.58
MI	92.0	42.2	13.3	44.6	31.2	0.64	91.3	55.7	62.9	**0.72**	95.6	38.9	45.1	0.62
TAG	93.2	51.0	15.6	50.2	36.4	0.65	84.8	**70.4**	**71.6**	0.71	92.1	35.1	39.2	0.54
NAST	94.4	44.1	13.3	44.1	32.0	0.64	88.8	65.1	66.7	0.70	93.7	40.1	44.9	0.56
RACoLN	91.2	48.1	13.8	47.7	32.1	**0.67**	87.5	49.9	54.6	0.71	92.9	36.6	40.3	0.52
STEC	92.5	**55.6**	**17.9**	**54.8**	**38.5**	0.65	88.9	68.7	71.1	0.71	**96.6**	**46.0**	**49.1**	**0.63**

(b) Style transfer on more difficult forms.

Table 3: Automatic evaluation results on sentiment transfer. Best results are in bold. Acc: Accuracy; BL-s: BLEU-self; BL-r: BLEU-reference; BS-s: BERTScore-self; BS-f: BERTScore-reference; GR: GRUEN.

	Yelp			Amazon			Captions			Politeness			Detoxification		
	Eff.	**CP**	**LQ**	**Eff.**	**CP**	**LQ**	**Eff.**	**CP**	**LQ**	**Eff.**	**CP**	**LQ**	**Eff.**	**CP**	**LQ**
DRG	4.0	3.7	3.5	3.7	3.0	3.3	2.8	2.7	3.0	4.0	4.1	3.7	4.0	3.0	3.3
MI	4.0	3.6	3.7	3.6	2.9	3.4	2.6	3.2	3.1	4.1	4.1	4.0	**4.2**	2.7	3.0
TAG	3.8	4.0	**3.8**	**3.8**	3.4	**3.8**	3.1	**3.5**	**3.5**	**4.4**	**4.5**	**4.3**	3.9	2.6	3.4
NAST	4.2	4.1	3.7	3.4	2.7	3.0	2.2	2.4	2.9	3.9	3.9	3.8	3.9	3.1	3.1
RACoLN	**4.3**	4.3	3.5	3.1	2.4	3.1	2.4	2.1	2.8	3.7	3.6	3.8	3.6	2.4	2.8
STEC	4.1	**4.6**	3.7	3.6	**3.9**	3.6	**3.2**	**3.5**	3.2	4.2	4.2	4.0	**4.2**	**3.7**	**3.4**

Table 4: Human evaluation results. Best results are in bold. Eff.: Transfer Effectiveness; CP: Content Preservation; LQ: Language Quality.

3.7 Efficiency Comparison

We implemented all models in Python 3.7 and conducted all the experiments on a computer with twenty 2.9 GHz Intel Core i7 CPUs and one GeForce GTX 1080 Ti GPU.

We report the average training time on five datasets. CAE: 7.8 hours; DRG: 4.7 hours; MI: 3.6 hours; TAG: 15.2 hours; NAST: 8.1 hours; RACoLN: 17.2 hours; STEC: 6.6 hours. We observe that our model requires relatively low training cost, compared to the baselines.

4 Related Work

Textual style transfer, the task of changing the style of an input sentence while preserving its content, has recently received increasing attention (Jin et al., 2021). To date, a wide range of solutions have been proposed to solve the task of textual style transfer, such as latent representation disentanglement (Shen et al., 2017; Fu et al., 2018; Riley et al., 2021; Nangi et al., 2021), prototype editing (Li et al., 2018; Wu et al., 2019; Malmi et al., 2020; Madaan et al., 2020; Reid and Zhong, 2021), and others (Gong et al., 2019; Jin et al., 2019; Goyal et al., 2021; Liu et al., 2021).

Many recent works have reported good performance on several aspects of style transfer, including sentiment (Li et al., 2018; Gong et al., 2019), formality (Rao and Tetreault, 2018), simplicity (Van den Bercken et al., 2019; Cao et al., 2020), politeness (Madaan et al., 2020), gender (Prabhumoye et al., 2018), authorship (Jhamtani et al., 2017; Carlson et al., 2018). However, content preservation still remains as a major challenge and yet to be

Transfer Type	Source Sentences	Transferred Sentences
(a) Negative → Positive:	we sit down and we got some really slow and lazy service.	we sit down and we got some really great service.
(b) Positive → Negative:	the taste is awesome.	the taste is really bad.
(c) Factual → Humorous:	the group of hikers is resting in front of a mountain.	the group of hikers is being pulled in front of a mountain.
(d) Factual → Romantic:	several young people celebrate by clapping and cheering.	several young people celebrate their lovely friendship by clapping and cheering.
(e) Impolite → Polite:	yes go ahead and remove it	could you please go ahead and remove it
(f) Toxic → Civil:	suggesting that people change their commute times is stupid.	suggesting that people change their commute times is useless.

Table 5: Case study: style transfer results by our proposed model. Tokens masked are in red, and new tokens generated are in blue.

solved (Jin et al., 2021; Lee et al., 2021; Huang et al., 2021).

To enhance content preservation, researchers have made some recent progress. For instance, Lee et al. (2021) propose to implicitly remove style at the token level using reverse attention, and fuse content information to style representation using conditional layer normalization. Besides it, Huang et al. (2021) propose a non-autoregressive generator, which can serve as an alternative generator for other established models. It explicitly models word alignments to suppress irrelevant words, exploits the word-level transfer between different styles, and is shown to improve content preservation for cycle-loss-based models. In addition, Gong et al. (2020) propose to encode rich syntactic and semantic information with a graph neural network and show its ability on sentiment transfer.

Our work differs from them in the following aspects:

1. Existing approaches for enhancing content preservation falls in the category of latent representation disentanglement approach, while, to the best of our knowledge, we have proposed the first model to enhance content preserve in the category of prototype editing.

2. Existing approaches rely on the assumption that latent representation can implicitly partially retain both content and style information. However, this assumption lacks justification and remains challengeable (Jin et al., 2021).

3. Most importantly, existing approaches primarily work on sentiment transfer, while limiting the generalizability of their models. Comparably, our proposed approach is demonstrated to work well on five diverse types of styles (i.e., sentiment, politeness, romance, humor, detoxification) and thus, shows more robustness and better generalizability.

5 Conclusion

In this paper, we identify two common types of errors on content preservation by existing style transfer models. To solve them, we propose to utilize a keyword extraction model to preserve the content-related tokens in the "mask" stage, and to leverage the self-supervision scheme to create a psuedo-parallel dataset in the "infill" stage. With the two core components, our model is able to enhance content preservation while keeping the outputs with target style. Both automatic and human evaluation shows that our model has the best ability in preserving content and is strong in other evaluation measures too.

Limitation and Future work: 1) we rely on the assumption that style is localized to certain tokens in a sentence and we can delete those tokens to obtain a style-free corrupted sentence. However, this assumption is not always true, especially for more complicated styles (e.g., from modern English to Shakespearean English). 2) In some styles, there are few words associated with the source target, which makes the "mask" model difficult to work well. For instance, in the Politeness dataset, "send me the data" is not a polite expression, but there are no impolite words associated either.

Reproducibility: Our code and pre-trained models will be made publicly available upon acceptance of the paper. The link is omitted for anonymity during review, but we will provide access to the reviewers and the program chair upon request.

Ethical Considerations

Risks in deployment: A text style transfer model can pose potential harm when used with a malicious intention. It can lead to a situation where one deliberately distorts a sentence for his or her own benefit. If one intentionally changes the style of another person with the proposed model structure, the generated output can be exploited to create misinformation or fake news. Given the limited scope of the present study, we call for attention to these aspects by way of well-designed experiments before deployment.

Regulatory standpoint on the present study: Institutional Review Board (IRB) gave us clear feedback on what is considered human research and thus subject to IRB review. Analyses relying on user-generated content do not constitute human-subject research, and are thus not the purview of the IRB, as long as 1) the data analyzed are posted on public fora and were not the result of direct interaction from the researchers with the people posting, 2) there are no private identifiers or personally identifiable information associated with the data, and 3) the research is not correlating different public sources of data to infer private data.[6] All of these conditions apply to the present study.

Risks in annotation: The data we use in this paper were posted on publicly accessible websites, and do not contain any personally identifiable information (i.e., no real names, email addresses, IP addresses, etc.). The annotators were graduate assistants in the lab receiving research credit for their annotation and were blind to the systems they were annotating. They were warned about the toxic content before they read the data, and were informed that they could quit the task at any time if they were uncomfortable with the content. The annotators in our study were evaluating the quality of the generated sentences only.

[6] This position is in line with Title 45 of the Code of Federal Regulations, Part 46 (45 CFR 46), which defines human research.

References

Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. 2018. Simple unsupervised keyphrase extraction using sentence embeddings. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society Open Science*, 5(10):171920.

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

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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 (NAACL-HLT)*.

Hongyu Gong, Linfeng Song, and Suma Bhat. 2020. Rich syntactic and semantic information helps unsupervised text style transfer. In *Proceedings of the 13th International Conference on Natural Language Generation*.

Navita Goyal, Balaji Vasan Srinivasan, N Anandhavelu, and Abhilasha Sancheti. 2021. Multi-style transfer with discriminative feedback on disjoint corpus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

George Heidorn. 2000. Intelligent writing assistance. *A handbook of natural language processing: Techniques and applications for the processing of language as text*, 8.

Fei Huang, Zikai Chen, Chen Henry Wu, Qihan Guo, Xiaoyan Zhu, and Minlie Huang. 2021. Nast: A non-autoregressive generator with word alignment for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2021. Deep learning for text style transfer: A survey. *Computational Linguistics*, pages 1–51.

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

Dan Jurafsky. 2000. *Speech & language processing*. Pearson Education India.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Bryan Klimt and Yiming Yang. 2004. Introducing the enron corpus. In *CEAS*.

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

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018*. Association for Computational Linguistics (ACL).

Yixin Liu, Graham Neubig, and John Wieting. 2021. On learning text style transfer with direct rewards. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*.

Sharmila Reddy Nangi, Niyati Chhaya, Sopan Khosla, Nikhil Kaushik, and Harshit Nyati. 2021. Counterfactuals to control latent disentangled text representations for 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)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL)*.

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 (ACL)*.

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

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Machel Reid and Victor Zhong. 2021. Lewis: Levenshtein editing for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David C Uthus, and Zarana Parekh. 2021. Textsettr: Few-shot text style extraction and tunable targeted restyling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

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

Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. {A4NT}: Author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium (USENIX Security 18)*.

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. "mask and infill": Applying masked language model to sentiment transfer. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Empirical Methods in Natural Language Processing: Findings (Findings of EMNLP)*.