Unsupervised Text Style Transfer via An Enhanced Operation Pipeline

Wanhui Qian[1,2], Jinzhu Yang[1,2], Fuqing Zhu[1,⋆], Yipeng Su[1], and Songlin Hu[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, China
{qianwanhui,yangjinzhu,zhufuqing,suyipeng,husonglin}@iie.ac.cn

Abstract. Unsupervised text style transfer aims to change the style attribute of the given unpaired texts while preserving the style-independent semantic content. In order to preserve the content, some methods directly remove the style-related words in texts. The remaining content, together with target stylized words, are fused to produce target samples with transferred style. In such a mechanism, two main challenges should be well addressed. First, due to the style-related words are not given explicitly in the original dataset, a detection algorithm is required to recognize the words in an unsupervised paradigm. Second, the compatibility between the remaining content and target stylized words should be guaranteed to produce valid samples. In this paper, we propose a multi-stage method following the working pipeline – *Detection*, *Matching*, and *Generation*. In the *Detection* stage, the style-related words are recognized by an effective joint method and replaced by mask tokens. Then, in the *Matching* stage, the contexts of the masks are employed as queries to retrieve target stylized tokens from candidates. Finally, in the *Generation* stage, the masked texts and retrieved style tokens are transformed to the target results by attentive decoding. On two public sentimental style datasets, experimental results demonstrate that our proposed method addresses the challenges mentioned above and achieves competitive performance compared with several state-of-the-art methods.

Keywords: Unsupervised Learning · Natural Language Generation · Sentiment Transfer.

1 Introduction

Text style transfer task focuses on alternating the style attribute of given texts while preserving the original semantic content and has drawn much attention from natural language generation community, especially with the inspiration of transfer learning. Due to the ability to modify attributes of texts in a fine-grained manner, text style transfer has potential applications on some specific tasks, such as dialogue system [19], authorship obfuscation [25]. However, pair-wised style-to-style translation corpora are hard to obtain. Therefore, searching an effective

⋆ *Corresponding author*

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

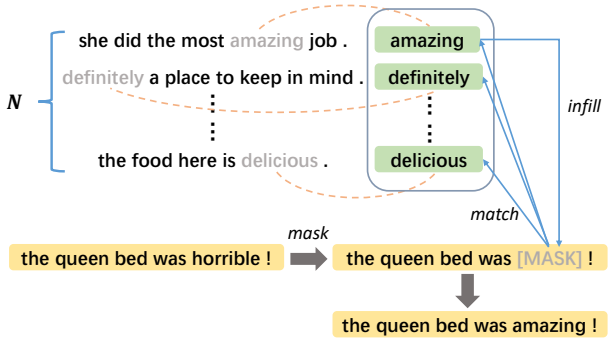

Fig. 1. An illustration of the context matching operation. The context of the '[MASK]' looks up the table of attribute markers (for simplicity, attribute markers are all set as words) extracted from N samples sentences. The word '*amazing*' is selected to infill the mask slot.

unsupervised method to conduct style transfer on unpaired texts has become a primary research direction.

Majority of the existing methods focus on separating style information and content apart at the first step. According to the separation manner, these methods fall into two groups – implicit disentanglement and explicit disentanglement. Specifically, implicit disentanglement methods [24, 6, 17] usually leverage the adversarial training strategy [7] to formulate the consistency of content distribution disentangled with various styles. Explicit disentanglement methods [34, 32, 31] recognize the *attribute markers* (words or phrases in the sentence, indicating a particular attribute[3]) in the original sentences[4], and replace the attribute markers with the expected ones. Compared to the implicit disentanglement operation, explicit replacement improves the model interpretability and enhances the ability of content preservation. However, the existing explicit disentanglement methods have apparent limitations in handling two problems.

First, selecting an effective attribute marker detection algorithm is critical for subsequent processing. Frequency-based [16] and Attention-based methods [32] are designed for detecting attribute markers. However, due to the inherent flaws, both of the two methods can hardly maintain the detection precision at a relatively high level. Second, the compatibility of the target attribute markers and style-independent content is the guarantee for producing valid sentences. For example, if we want to transfer the sentiment style of "*I love the movie*" from positive to negative, the word '*love*' should be replaced with '*hate*', but not '*disappoint*'. *DeleteAndRetrieve* [16] achieves the style transfer by exchanging the attribute markers between two sentences with similar content but different

[3] The concept of *attribute marker* is borrowed from the work in [16], we use this term to indicate the style information.

[4] In this work, the 'text' and 'sentence' terms are exchangeable.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

styles. In actual situations, the conditions are too harsh to be well satisfied. Wu et al. (2019) [31] fine-tune a BERT [5] to infer the attribute markers to infill the masks. However, the one mask-one word infilling manner of BERT restricts the flexibility of selecting attribute markers.

To address the above limitations, we propose a multi-stage method following the enhanced pipeline – *Detection*, *Matching* and *Generation*. In the *Detection* stage, motivated by the fused masking strategy in [31], we design an effective joint method to recognize the attribute markers for alleviating the detection problem. In the *Matching* stage which is illustrated in Fig. 1, the contexts of the masks are employed as queries to retrieve compatible attribute markers from a set of sampled attribute tokens. In the last *Generation* stage, the masked texts and retrieved attribute markers are fed into a decoder to produce the transferred samples. To ensure the style is correctly transferred, we introduce an auxiliary style classifier pre-trained on the non-parallel corpus.

We conduct experiments on public two sentiment style datasets. Three aspects indicators are utilized to evaluate the transferred samples, i.e., target style accuracy, content preservation and language fluency. The main contributions of this paper are summarized as follows:

- A multi-stage (including *Detection*, *Matching* and *Generation*) method is proposed for unsupervised text style transfer task, transferring the given unpaired texts by explicitly manipulation.
- In the *Detection* and *Matching* stages, the attribute marker detection and content-marker compatibility problems are well alleviated by the joint detection method and the context-matching operation.
- Experimental results demonstrate that the proposed model achieves competitive performance compared with several state-of-the-art methods. Besides, the model gains an excellent trade-off between style accuracy and content preservation.

2 Related Work

Recently, unsupervised text style transfer task has attracted broad interest. Shen et al. (2017); Fu et al. (2018); and John et al. (2019) [24, 6, 9] assume that the style and content can be separated via generative models (e.g. GAN [7], VAE [12]). Follow the above assumption, Yang et al. (2018) [33] adopt a pre-trained language model to improve the fluency of the generated sentences. Considering the content preservation, Prabhumoye et al. (2018) [22] design a dual language translation model. The semantic content in the source and target sentences of translation remains unchanged. Logeswaran et al. (2018) [17] discard the language translation process and back-translate the transferred sentence to the original sentence directly. Despite the developments of those methods, some work [15, 4, 29] suspects the efficiency of the separation in latent space and proposes methods with end-to-end translation fashion. Some work attempts to pre-build a parallel dataset from the original non-parallel corpus. Zhang et al. (2018a) [35] initialize the dataset by the unsupervised word mapping techniques [14, 1] which

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

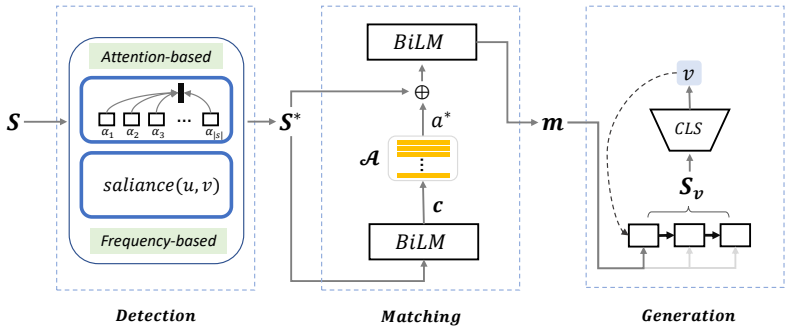

Fig. 2. Overview structure of the proposed model. First, the original sentence $\boldsymbol{s}$ is fed into the *Detection* module, the attribute markers are recognize and replaced by masks. Then the remaining content $\boldsymbol{s}^*$ is passed to the *Matching* module. The bidirectional language model BiLM extracts the contexts $\boldsymbol{c}$ from $\boldsymbol{s}^*$, and perform a matching operation between mask context in $\boldsymbol{c}$ and a sampled attribute token set $\mathcal{A}$. The retrieved results a^* and $\boldsymbol{s}^*$ are combined and encoded into memories $\boldsymbol{m}$. Finally, the *Generation* module decodes the $\boldsymbol{m}$ to the target sentence $\boldsymbol{s}_v$. Additionally, a pre-trained style classifier CLS is appended to the decoder to strengthen the style control ability of the decoder.

are widely used in the unsupervised machine translation task. Jin et al. (2019) [8] construct sentence pairs according to the distance measurement between sentences. More straightforwardly, Luo et al. (2019) [18] utilize the transferred results of [16] as the initial target samples.

In our work, the style of sentences is transferred through word-level manipulation. The content of the original sentences is preserved effectively, and the manipulation process is interpretable. Previous methods based on explicit word manipulation usually improve the word detection process or the generation process. Zhang et al. (2018b)[34] propose a self-attention method to detect attribute words. Li et al. (2018) [16] present a frequency-based method and four generating strategies. [32] adopt a similar attention-based method with that in [34] and propose a cycle-reinforcement learning algorithm. Wu et al. (2019) [31] fuse the detection methods in [16] and [32], and generate target sentences with a fine-tuned BERT [5]. Sudhakar et al. (2019) [26] improve each operation step in [16] and gain better performance. Compared to the above methods, our model overcomes the inherent weakness of previous detection methods and a well-designed joint method is utilized to locate attribute markers. Additionally, our method retrieves target attribute markers by a matching operation, and the final results are generated through attentive decoding.

3 Proposed Method

In this paper, we employ the corpus with a style set $\mathcal{V}$. Each collection $\mathcal{D}_v$ represents the sentences with the style $v \in \mathcal{V}$. Given any source style v_{src} and any

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

target style v_{tgt} ($v_{\text{src}} \neq v_{\text{tgt}}$), the goal of the style transfer task is to learn a projection function $f_{\text{src}\to\text{tgt}}$ achieving label transfer of sentences in $\mathcal{D}_{v_{\text{src}}}$ from v_{src} to v_{tgt} while preserving the style-independent semantic content. In this section, we describe our style transfer method from the perspective of the corresponding working pipeline and learning algorithm. The overview architecture of the proposed model is illustrated in Fig. 2, which contains three modules (i.e., *Detection*, *Matching*, and *Generation*).

3.1 Detection Module

We first introduce two existing methods for detecting attribute markers – Frequency-based method [16] and Attention-based method [32]. Then, a joint detection method is proposed for better identifying attribute markers.

Frequency-based Method If the frequency of an given n-gram u appears in $\mathcal{D}_v$ is much higher than that in other datasets, u has a higher probability of being a v-style attribute marker. Specifically, for a given n-gram u and a chosen style v from the style set $\mathcal{V}$, a quantity $s(u, v)$ called *salience* is defined for the statement as:

$$s(u,v) = \frac{\text{count}(u, \mathcal{D}_v) + \lambda}{\sum_{v' \in \mathcal{V}, v' \neq v} \text{count}(u, \mathcal{D}_{v'}) + \lambda}, \tag{1}$$

where $\text{count}(u, \mathcal{D}_v)$ is the number of times that an n-gram u appears in $\mathcal{D}_v$, and λ is the smoothing parameter. If $s(u, v)$ is larger than a predefined threshold, u will be identified as an attribute marker for the style v.

Attention-based Method To apply this method, a pre-trained attention-based LSTM classifier is required. Given a sentence consists of a sequence of tokens $\boldsymbol{s} = [t_1, t_2, \ldots, t_{|\boldsymbol{s}|}]$ with style $v \in \mathcal{V}$, a LSTM module first encodes the tokens into a sequence of hidden states $\boldsymbol{h} = [h_1, h_2, \ldots, h_{|\boldsymbol{s}|}]$. Then, an attention operation is conducted between v and $\boldsymbol{h}$ to obtain a sequence of normalized weights $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_{|\boldsymbol{s}|}]$:

$$\boldsymbol{\alpha} = \text{softmax}(\text{attention}(v, \boldsymbol{h})), \tag{2}$$

Finally, an accumulated feature vector $\boldsymbol{\alpha} \cdot \boldsymbol{h}$ is utilized to predict the style label v.

In such a process, the weights $\boldsymbol{\alpha}$ can be regarded as the contributions of corresponding tokens for style prediction. Therefore, we use α_i denoting the score of style for token t_i, $i \in \{1, 2, \ldots, |\boldsymbol{s}|\}$. To identify the attribute markers, we define the attribute identifier $\hat{\alpha}_i$ as follows:

$$\hat{\alpha}_i = \begin{cases} 0, & \text{if } \alpha_i < \bar{\alpha}; \\ 1, & \text{if } \alpha_i \geq \bar{\alpha}, \end{cases} \tag{3}$$

where the $\bar{\alpha}_i$ is the mean of weights $\boldsymbol{\alpha}$. If the identifier $\hat{\alpha}_i$ is equals to 1, the corresponding token will be recognized as a single-token attribute marker, otherwise, the token will be ignored.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

Joint Detection Method Even though the above two methods implement the extraction of attribute markers, they both have some limitations. Occasionally, as mentioned in [31], the Frequency-based method recognizes style-independent grams as attribute markers. This misdetection ruins both style transfer and content preservation. As for the Attention-based method, the LSTM module affects the precision of detecting the attribute markers. Because the last output of the LSTM accumulates the full sequential information, resulting in that the classifier tends to pay more attention to the end of outputs. To alleviate the inherent problems, we design a joint heuristic method combining the Frequency-based method and Attention-based method.

Given style v, the joint method assign a score α_u^* for a gram u according to following equation:

$$\alpha_u^* = \alpha_u * \log s(u, v). \tag{4}$$

The criterion of decision making has the same formulation with Eq.3. We explain the formulation of the joint detection method (Eq.4) from three points:

1. The 'misdetection' problem in Frequency-based method is alleviated by multiplying the score α_u acquired from Attention-based method.
2. The LSTM module in Attention-based method is removed to prevent the 'focusing-on-last' problem.
3. The logarithm over $s(u, v)$ prevents the extremely large values affecting the detection accuracy.

Once the attribute markers detected, we replace them with mask tokens (one mask for one attribute marker).

3.2 Matching Module

In this module, we construct compatible attribute markers through matching operation. The overall process consists of determining queries, constructing candidates, and matching.

In our implementation, the queries are the contexts of corresponding masks. Ubiquitous bidirectional language models BiLM (such as BiLSTM [21], Transformer Encoder [5]) are effective tools for extracting contexts. Given a masked sentence, $\boldsymbol{s}^* = [t_1, t_2, \ldots, t_{|\boldsymbol{s}^*|}]$. Without loss of generality, we/ suppose that the k-th token in $\boldsymbol{s}^*$ is a mask[5]. The context information $\boldsymbol{c} = [c_1, c_2, \ldots, c_{|\boldsymbol{s}^*|}]$ is obtained by:

$$\boldsymbol{c} = \text{BiLM}(\boldsymbol{s}^*). \tag{5}$$

Obviously, the k-th element c_k is the context for mask t_k. Now the query c_k is prepared, the next problem is how to construct the candidates to be matched.

By investigating the corpus of interest, we observe that the frequencies of attribute tokens (Words that consist attribute markers) have the long-tail phenomena. Most of the attribute tokens appear only a few times, and the vast

[5] To simplify the demonstration, only one mask t_k is considered. The strategy of multi-masks is the parallel situation for single mask.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

majority of sentences share a small number of attribute tokens. Therefore, we randomly samples N sentences and extract the attribute tokens from the sentences as the candidates, represented by $\mathcal{A} = [a_1, a_2, \ldots, a_{|\mathcal{A}|}]$.

The matching operation indicates an attention between c_i and $\mathcal{A}$, which produces the a compatible attribute marker a^*:

$$\boldsymbol{\beta} = \text{softmax}(\text{attention}(c_i, \mathcal{A})), \tag{6}$$

$$a^* = \boldsymbol{\beta} \cdot \mathcal{A}, \tag{7}$$

where the $\boldsymbol{\beta}$ is normalized attention weights. The attentive result a^* is the weighted sum of candidates in $\mathcal{A}$, which can be viewed as a composition of attribute tokens. A new sentence representation $\hat{\boldsymbol{s}} = [t_1, t_2, \ldots, a^*, \ldots, t_{|\boldsymbol{s}^*|}]$ is acquired by replacing t_k with a^* in $\boldsymbol{s}^*$. At last, we encode $\hat{\boldsymbol{s}}$ to an external memories $\boldsymbol{m}$ for further processing:

$$\boldsymbol{m} = \text{BiLM}(\hat{\boldsymbol{s}}). \tag{8}$$

3.3 Generation Module

Compared to the vanilla auto-encoder framework, decoding with attention enables to stabilize the generation of sentences. The generation module adopts the attention structure from [2].

A sentence $\boldsymbol{s}_v$ with style v is generated by recurrent decoding:

$$\boldsymbol{s}_v = \text{Decoder}(I_v, \boldsymbol{m}), \tag{9}$$

where the given style indicator I_v for style v is set as the initial hidden state of the recurrent decoder, and $\boldsymbol{m}$ is the external memory from the matching module to be attended.

Previous methods for text style transfer usually adopt vanilla auto-encoder as the backbone. The attention mechanism is disabled because the sentence reconstruction tends to degenerate to a copy operation. Inspired by the solution in [13], we add noises to original sentences to prevent the corruption. Each token in content has an equivalent probability p_{noise} to be removed, replaced or appended with a sampled token.

3.4 Learning Algorithm

The learning process has two steps. The first step attempts to reconstruct the original sentence $\boldsymbol{s}$ with style v_{src}. The second step evaluates the style accuracy of the transferred samples $\boldsymbol{s}_{\text{tsf}}$ with target style v_{tgt}. Two steps produce two losses – reconstruction loss and classification loss, respectively.

Reconstruction Loss Due to the non-parallel nature of datasets, we follow a self-transfer routine $\boldsymbol{s} \to \boldsymbol{s}^* \to \boldsymbol{s}$. To recover the original sentence $\boldsymbol{s}$, the attribute token candidates $\mathcal{A}_{\text{src}}$ is aggregated with those extracted from sampled sentences in $\mathcal{D}_{v_{\text{src}}}$ ($\boldsymbol{s}$ is included in the sampled sentences). At generation step, the style indicator I_v is set as $I_{v_{\text{src}}}$. The reconstruction loss is formulated as:

$$\mathcal{L}_{\text{rec}} = -\log p(\boldsymbol{s}|\boldsymbol{s}^*, \mathcal{A}_{\text{src}}, I_{v_{\text{src}}}). \tag{10}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

Classification Loss As previous methods done, we append an auxiliary pre-trained style classifier to the end of decoder. The classifier aims to enhance the control over style transferring. The prediction routine is $\boldsymbol{s} \rightarrow \boldsymbol{s}^* \rightarrow \boldsymbol{s}_{\text{tsf}} \rightarrow p(v_{\text{tgt}}|\boldsymbol{s}_{\text{tsf}})$, the transferred sample $\boldsymbol{s}_{\text{tsf}}$ is expected to possess the target style v_{tgt}. To enable the style transfer, the attribute token candidates $\mathcal{A}_{\text{tgt}}$ consists of those extracted from sampled sentences in $\mathcal{D}_{v_{\text{tgt}}}$. The generation indicator is set as $I_{v_{\text{tgt}}}$. The classification loss is:

$$\mathcal{L}_{\text{cls}} = -\log p(v|\boldsymbol{s}^*, \mathcal{A}_{\text{tgt}}, I_{v_{\text{tgt}}}). \tag{11}$$

To resolve the discreteness problem of texts, we adopt the same strategy in [31].

The final optimization target is:

$$\min_{\phi} \mathcal{L} = \mathcal{L}_{\text{rec}} + \eta \mathcal{L}_{\text{cls}}, \tag{12}$$

where ϕ represents the trainable parameter set of the model, and η is a predefined parameter for scaling the classification loss.

4 Experiments

In this section, we first describe the experimental settings, including datasets, baselines, evaluation metrics and experimental details. Then we show the experimental results and analysis, where some comparisons will be provided to demonstrate the effectiveness of the proposed method.

4.1 Datasets

This paper evaluates the performance of the proposed method on two public sentimental style datasets released by [16], i.e., *Yelp* and *Amazon*, including reviews with binary polarity. The above datasets are pre-processed and divided into three sets for training, developing and testing. Additionally, crowd-workers are hired on Amazon Mechanical Turk to write references for all testing sentences [16], ensuring each of the references hold opposite sentiment and similar content with the original sentence. The references can be regarded as the standard gold outputs to evaluate the performances of the proposed model.

4.2 Baselines

In this paper, the following state-of-the-art methods are employed as baselines for comparison, including **CrossAE** [24], three strategies (**Template**, **DeleteOnly**, **Del&Retr**) in [16], **CycleS2S** [32], **C-BERT** [31], **DualRL** [18], **PTO** [30].

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

4.3 Evaluation Metrics

Automatic Evaluation Following the work [24, 16], we estimate the **style accuracy** (ACC) of the transferred sentences with a pre-trained classifier of fastText[6] [10]. The validation accuracy on development set of *Yelp* and *Amazon* achieves 97% and 80.4%, respectively. Similar to the machine translation task, the BLEU [20] score between generated result and the human reference is the measurement of **content preservation**. In our implementation, the BLEU score is calculated through the moses script[7]. To evaluate the **fluency** of sentences, we pre-train two language models on *Yelp* and *Amazon* datasets by fine-tuning two distinct GPT-2[8] [23]. The perplexity (PPL) of transferred sentences indicates the fluency rate.

Human Evaluation For either *Yelp* or *Amazon*, we sample and annotate 100 transferred sentences (50 for each sentiment) randomly. Without any knowledge about the model which produces the sentences, three annotators are required to evaluate every sentence from the aspect of style control, content preservation, and language fluency. For each target sentence, the annotator should give answers for three questions: 1) Does the sentence hold the correct style? 2) Is the content preserved in the target sentence? 3) Is the expression fluent? For any question with answer 'yes', the target sentence is labeled with '1', otherwise labeled with '0'.

4.4 Experimental Details

During the data pre-processing, the sentence length on *Yelp* and *Amazon* datasets is limited to 23 and 30, respectively. To alleviate the word sparsity problem, we set the word as 'unk' if the corresponding frequency is below 5. The noise rate p_{noise} is set to 0.05 to stabilize the training process. In terms of model setting, the style detection module recognizes n-grams with up to 4 tokens, the smoothing parameter λ is 1.0. The BiLM and Decoder are implemented as recurrent networks, which both adopt GRU [3] as the recurrent unit. The dimension of word embedding and the hidden size of GRU are both set to 512. All the attention operations in the proposed model are employed in the *Scaled Dot-Product* schema [28]. At training stage, the scale coefficient η and batch size is set to 0.05 and 100 respectively. Then, the training is done after 30K iterations, optimized by an Adam [11] optimizer with a fixed learning rate 0.0003.

4.5 Experimental Results and Analysis

The automatic evaluation results are shown in Table 1. To avoid the margin problem described in [27], each of the results is an averaged value from 5 single

[6] https://github.com/facebookresearch/fastText
[7] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl
[8] https://github.com/huggingface/pytorch-transformers

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

Table 1. Automatic evaluation results on the *Yelp* and *Amazon* datasets. ↑ denotes the higher the better and ↓ denotes the lower the better.

Models	*Yelp*			*Amazon*		
	Accuracy(%)↑	BLEU↑	Perplexity↓	Accuracy(%)↑	BLEU↑	Perplexity↓
CrossAE [24]	84.4	5.4	53.3	60.7	1.7	45.2
Template [16]	84.2	22.2	111.7	69.8	31.8	95.9
DeleteOnly [16]	85.6	15.1	68.8	47.2	27.9	69.9
Del&Retr [16]	89.1	15.5	46.4	47.9	27.9	54.8
CycleS2S [32]	52.8	18.5	161.7	49.9	14.2	N/A
C-BERT [31]	95.2	26.0	54.9	88.4	33.9	114.0
PTO [30]	86.3	**29.3**	46.0	45.9	**36.6**	76.6
DualRL [18]	88.4	27.5	48.2	45.8	34.8	**43.9**
Ours	**95.5**	25.8	**41.4**	**94.0**	23.8	47.4

Table 2. Human evaluation results (%) on the *Yelp* and *Amazon* datasets. The evaluation includes three aspects: style accuracy (denoted as Sty), content preservation (denoted as Con), and fluency (denoted as Flu). Each cell indicates the proportion of sentences that passed the human test.

Models	*Yelp*			*Amazon*		
	Sty	Con	Flu	Sty	Con	Flu
Del&Retr	19.3	25.8	36.8	7.0	29.1	26.8
C-BERT	32.8	**43.1**	41.3	**16.8**	**33.1**	30.1
Ours	**34.6**	42.8	**45.5**	10.1	21.1	**36.1**

running models initialized with different random seeds. Compared to state-of-the-art systems, the proposed model achieves a competitive performance and get a more excellent style accuracy on both *Yelp* and *Amazon*. However, some methods, such as C-BERT, DualRL, PTO, preserve more style-independent content than the proposed model. The reason maybe that the detection module in the proposed model tends to boost the recall rate of recognizing attribute markers. As a result, more tokens in original sentences are removed, then the less contents are preserved. In terms of language fluency (i.e., Perplexity), the proposed model is superior to most baselines. Because the matching module guarantees the compatibility of target attribute markers and contents, the attentive generation module keeps the stability of the generation process.

Table 2 shows the human evaluation results of the two well-performed models (both have the similar training process of our model), Del&Retr and C-BERT, in automatic evaluation. We find that the results on *Yelp* are generally consistent with that in automatic evaluation. However, the most confusing part is the style accuracy of the proposed model on *Amazon* (i.e., our model performs 6.7 percentage points lower than C-BERT. However in the automatic evaluation, our

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

Table 3. Examples of generated sentences from CrossAE [24], Del&Retr [16], PTO [30] and our model. Words with different colors have different meanings, specifically: blue → sentiment words; green → correct transferred part; red → errors (i.e. sentiment error, grammar error, content changed and etc).

	Yelp: negative → positive	*Yelp*: positive → negative
Source	i ca n't believe how inconsiderate this pharmacy is .	portions are very generous and food is fantastically flavorful .
CrossAE	i do n't know this store is great .	people are huge and the food was dry and dry .
Del&Retr	this pharmacy is a great place to go with .	portions are bland and food is fantastically not at all flavorful .
PTO	i delightfully n't believe how great this is .	portions are very bland and food is not flavorful .
Ours	i always believe how good this pharmacy is .	portions are very weak and food is deeply bland .

	Amazon: negative → positive	*Amazon*: positive → negative
Source	it crashed for no reason , saves got corrupted .	exactly what i need for my phone and at the best price possible .
CrossAE	it works for me for # years, etc .	i don t believe the price for # months and i am using it .
Del&Retr	it works flawlessly , works , and does easy to use and clean , saves got .	was really excited to get this for my phone and at the best price possible .
PTO	it crashed for no reason , got delicious .	exactly what i need for my phone and at the worst price possible .
Ours	it worked great for no reason , got perfect results .	not what i needed for my game and at the same price possible .

model perform 5.6 percentage points higher than C-BERT.) By investigating the *Amazon* dataset, we find that the reason is the imbalance phenomena in positive and negative product reviews. For example, the word 'game' appears 14,301 times in negative training set, while it appears only 217 times in positive set. Therefore, the detection module tends to recognize the 'game' as a negative attribute marker. This phenomenon interferes the detection of attribute markers severely, and more style-independent content is misidentified. We believe that leveraging external knowledge could alleviate the above imbalance problem, this is a potential candidate for further exploring.

Besides the formal evaluations, some transferred sentences are presented in Table 3 for further qualitative analysis. The grammar of sentences generated by our model is generally correct. The semantic is more consistent than some baselines. The presented results demonstrates that the cooperation of the *Detection* and *Matching* indeed make a stable improvement of our model. We observe the results in *Yelp* and *Amazon* datasets, respectively. Most of the models struggle in transferring sentiment of sentences on *Amazon*. Our model has transferred the the last example (*Amazon*: positive → negative) by replacing the word 'phone' with 'game'. This phenomena is consistent with the observation of imbalance problem mentioned in human evaluation part.

4.6 Ablation Study

To estimate the influence of different components on the overall performance, we remove the components individually and check the model performance on *Yelp* dataset. The results are reported in Table 4.

First, we replace the Joint-Detection method with Frequency-based method. As a result, the style accuracy reduces drastically (18.9% below the full model) while the fluency is increased. The Frequency-based method tends to recognize style-independent tokens as attribute markers, and the undetected attribute markers deeply affect the style accuracy. If we replace the Joint-Detection with Attention-based method, the style accuracy and language fluency decrease slightly.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

Table 4. Automatic evaluation results of ablation study on *Yelp* dataset. 'Joint → Freq' indicates replacing Joint Detection with Frequency-based method. Similarly, 'Joint → Attn' indicates replacing Joint Detection with the Attention-based method.

Models	Accuracy(%)	BLEU	Perplexity
Joint → Freq	76.6	25.8	35.4
Joint → Attn	94.6	25.8	44.0
- match	95.0	24.8	41.8
- noise	88.0	26.4	42.4
- $\mathcal{L}_{\text{cls}}$	86.1	26.4	40.6
Full model	95.5	25.8	41.4

The above results show that the proposed Joint-Detection method is superior to the Frequency-based and Attention-based methods in terms of the detection accuracy.

Then, we discard the matching operation in the matching module. The rest model is similar to the *DeleteOnly* in [16], which infers the removed attribute markers based on style-independent content. The overall performance reduces in all three aspects. This result further supports our claim that the matching operation tends to select compatible attribute markers.

Finally, two training tricks – denoising mechanism and classification loss, are taken into consideration. Without noises, the model can preserve more content as the BLEU score increases. However, the style accuracy dropped by 7.5% at the same time. Moreover, due to the lack of denoising ability, the model cannot generate sentences smoothly (the corresponding perplexity rises). If we remove the classification loss $\mathcal{L}_{\text{cls}}$, the style accuracy decreases from 95.5% to 86.1%. Therefore, the classification loss is critical for boosting the style transfer strength. Due to the model would corrupt if the $\mathcal{L}_{\text{rec}}$ is disabled, we ignore the ablation study on $\mathcal{L}_{\text{rec}}$.

In summary, the studies on the detection and matching modules have proved that the issues of detection accuracy and compatibility could be resolved. The studies on denoising mechanism and classification loss demonstrate their crucial rule in improving the style transfer strength.

5 Conclusion

In this paper, we propose a multi-stage method to address the detection accuracy and compatibility issues for unsupervised text style transfer. The joint detection method is designed to combine the Frequency-based and Attention-based methods for recognizing attribute markers. The matching operation is presented to seek the compatible tokens for retrieving the target style information. Both the automatic and human evaluation results show that the proposed model achieves competitive performance compared with several state-of-the-art systems. The

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

ablation study confirms that the designed joint detection method enhances the style transfer strength, and the matching operation improves the fluency of generated sentences. In *Amazon*, we observe the data imbalance problem which severely reduces the model performance. Therefore, achieving unsupervised text style transfer in imbalanced scenario is the topic for our future exploration.

Acknowledgement

This research is supported in part by the Beijing Municipal Science and Technology Project under Grant Z191100007119008.

References

1. Artetxe, M., Labaka, G., Agirre, E.: Learning bilingual word embeddings with (almost) no bilingual data. In: Proc. ACL. pp. 451–462 (2017)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proc. ICLR (2015)
3. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proc. EMNLP. pp. 1724–1734 (2014)
4. Dai, N., Liang, J., Qiu, X., Huang, X.: Style transformer: Unpaired text style transfer without disentangled latent representation. In: Proc. ACL. pp. 5997–6007 (2019)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proc. NAACL. pp. 4171–4186 (2019)
6. Fu, Z., Tan, X., Peng, N., Zhao, D., Yan, R.: Style transfer in text: Exploration and evaluation. In: Proc. AAAI. pp. 663–670 (2018)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proc. NeurIPS. pp. 2672–2680 (2014)
8. Jin, Z., Jin, D., Mueller, J., Matthews, N., Santus, E.: Unsupervised text style transfer via iterative matching and translation. In: Proc. EMNLP. pp. 3097–3109 (2019)
9. John, V., Mou, L., Bahuleyan, H., Vechtomova, O.: Disentangled representation learning for non-parallel text style transfer. In: Proc. ACL. pp. 424–434 (2019)
10. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proc. EACL. pp. 427–431 (2017)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. ICLR (2015)
12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Proc. ICLR (2014)
13. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only. In: Proc. ICLR (2018)
14. Lample, G., Conneau, A., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. In: Proc. ICLR (2018)
15. Lample, G., Subramanian, S., Smith, E.M., Denoyer, L., Ranzato, M., Boureau, Y.: Multiple-attribute text rewriting. In: Proc. ICLR (2019)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9

16. Li, J., Jia, R., He, H., Liang, P.: Delete, retrieve, generate: a simple approach to sentiment and style transfer. In: Proc. NAACL. pp. 1865–1874 (2018)
17. Logeswaran, L., Lee, H., Bengio, S.: Content preserving text generation with attribute controls. In: Proc. NeurIPS. pp. 5103–5113 (2018)
18. Luo, F., Li, P., Zhou, J., Yang, P., Chang, B., Sun, X., Sui, Z.: A dual reinforcement learning framework for unsupervised text style transfer. In: Proc. IJCAI. pp. 5116–5122 (2019)
19. Oraby, S., Reed, L., Tandon, S., S., S.T., Lukin, S.M., Walker, M.A.: Controlling personality-based stylistic variation with neural natural language generators. In: Proc. SIGDIAL. pp. 180–190 (2018)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proc. ACL. pp. 311–318 (2002)
21. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proc. NAACL. pp. 2227–2237 (2018)
22. Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., Black, A.W.: Style transfer through back-translation. In: Proc. ACL. pp. 866–876 (2018)
23. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
24. Shen, T., Lei, T., Barzilay, R., Jaakkola, T.: Style transfer from non-parallel text by cross-alignment. In: Proc. NeurIPS. pp. 6830–6841 (2017)
25. Shetty, R., Schiele, B., Fritz, M.: A4NT: author attribute anonymity by adversarial training of neural machine translation. In: Proc. USENIX. pp. 1633–1650 (2018)
26. Sudhakar, A., Upadhyay, B., Maheswaran, A.: "transforming" delete, retrieve, generate approach for controlled text style transfer. In: Proc. EMNLP. pp. 3260–3270 (2019)
27. Tikhonov, A., Shibaev, V., Nagaev, A., Nugmanova, A., Yamshchikov, I.P.: Style transfer for texts: Retrain, report errors, compare with rewrites. In: Proc. EMNLP. pp. 3927–3936 (2019)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proc. NeurIPS. pp. 5998–6008 (2017)
29. Wang, K., Hua, H., Wan, X.: Controllable unsupervised text attribute transfer via editing entangled latent representation. arXiv preprint arXiv:1905.12926 (2019)
30. Wu, C., Ren, X., Luo, F., Sun, X.: A hierarchical reinforced sequence operation method for unsupervised text style transfer. In: Proc. ACL. pp. 4873–4883 (2019)
31. Wu, X., Zhang, T., Zang, L., Han, J., Hu, S.: Mask and infill: Applying masked language model for sentiment transfer. In: Proc. IJCAI. pp. 5271–5277 (2019)
32. Xu, J., Sun, X., Zeng, Q., Zhang, X., Ren, X., Wang, H., Li, W.: Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In: Proc. ACL. pp. 979–988 (2018)
33. Yang, Z., Hu, Z., Dyer, C., Xing, E.P., Berg-Kirkpatrick, T.: Unsupervised text style transfer using language models as discriminators. In: Proc. NIPS. pp. 7287–7298 (2018)
34. Zhang, Y., Xu, J., Yang, P., Sun, X.: Learning sentiment memories for sentiment modification without parallel data. In: Proc. EMNLP. pp. 1103–1108 (2018)
35. Zhang, Z., Ren, S., Liu, S., Wang, J., Chen, P., Li, M., Zhou, M., Chen, E.: Style transfer as unsupervised machine translation. arXiv preprint arXiv:1808.07894 (2018)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_9