Transductive Learning for Unsupervised Text Style Transfer

Fei Xiao[1,4], **Liang Pang**[1*], **Yanyan Lan**[3*], **Yan Wang**[5], **Huawei Shen**[1,4], **Xueqi Cheng**[2,4]
[1]Data Intelligence System Research Center
and [2]CAS Key Lab of Network Data Science and Technology,
Institute of Computing Technology, Chinese Academy of Sciences
[3]Institute for AI Industry Research, Tsinghua University
[4]University of Chinese Academy of Sciences [5]Tencent AI Lab
{xiaofei19s, pangliang, shenhuawei, cxq}@ict.ac.cn
lanyanyan@tsinghua.edu.cn, brandenwang@tencent.com

Abstract

Unsupervised style transfer models are mainly based on an inductive learning approach, which represents the style as embeddings, decoder parameters, or discriminator parameters and directly applies these general rules to the test cases. However, the lacking of parallel corpus hinders the ability of these inductive learning methods on this task. As a result, it is likely to cause severe inconsistent style expressions, like *the salad is rude*. To tackle this problem, we propose a novel transductive learning approach in this paper, based on a retrieval-based context-aware style representation. Specifically, an attentional encoder-decoder with a retriever framework is utilized. It involves top-K relevant sentences in the target style in the transfer process. In this way, we can learn a context-aware style embedding to alleviate the above inconsistency problem. In this paper, both sparse (BM25) and dense retrieval functions (MIPS) are used, and two objective functions are designed to facilitate joint learning. Experimental results show that our method outperforms several strong baselines. The proposed transductive learning approach is general and effective to the task of unsupervised style transfer, and we will apply it to the other two typical methods in the future.

1 Introduction

Text style transfer is an essential topic of natural language generation, which is widely used in many tasks such as sentiment transfer (Hu et al., 2017; Shen et al., 2017), dialogue generation (Zhou et al., 2018; Niu and Bansal, 2018; Su et al., 2021), and text formalization (Jain et al., 2019). The target is to change the style of the text while retaining style-independent content. As it is usually hard to obtain large parallel corpora with the same content and different styles, unsupervised text style transfer becomes a hot yet challenging research topic in recent years.

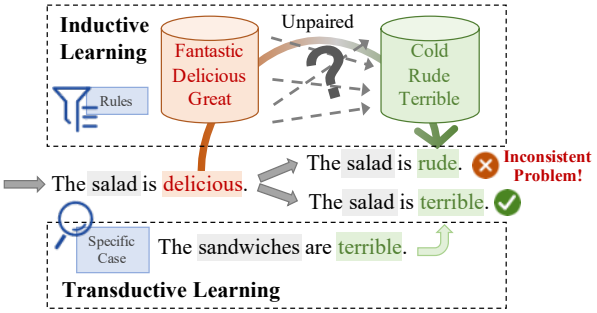

Figure 1: Illustration of the inconsistency problem and the idea of our transductive learning approach.

Most existing methods in this area try to find the general style transfer rules with an inductive learning paradigm, where style is represented as a specific form, e.g., embeddings, decoder parameters, or classifier parameters. For example, embedding based methods (Shen et al., 2017; Lample et al., 2019; John et al., 2019; Dai et al., 2019; Yi et al., 2020; He et al., 2020) utilize a highly generalized style embedding to replace the original sentence style and direct the generation process. Decoder based methods (Prabhumoye et al., 2018; Fu et al., 2018; Luo et al., 2019; Gong et al., 2019; Krishna et al., 2020) use multiple decoders for generation, where each encoder corresponds to an independent style. Classifier based methods (Wang et al., 2019; Liu et al., 2020; Mai et al., 2020) employ the gradient of a pre-trained style classifier to edit the latent representation of the target text.

It has been well accepted that inductive learning methods have the ability to work well when there are numerous supervised labels. However, in the case of unsupervised style transfer, we are just given corpora with different styles without knowing the parallel relation, i.e. supervision label for this

*Corresponding Author

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2510–2521
November 7–11, 2021.

task. As a result, the inductive learning methods fail to produce an accurate style transfer rule, leading to generating some severe inconsistent texts, such as '*the salad is rude*', as shown in Figure 1. The underlying reason for this phenomenon is that a perfect style is usually highly dependent on the context, e.g. '*terrible*' for '*the salad*' and '*rude*' for '*a person*'. Without a large scale of parallel data, it is difficult to learn a general style transfer rule working for various contexts.

Inspired by the idea of transductive learning (Vapnik, 1998) and some successful historical examples, such as Transductive SVM (Joachims et al., 1999), we propose to introduce transductive learning to the area of unsupervised text style transfer. Specifically, transduction learning reasoning from specific cases to specific cases, which avoids learning a general rule to represent the style. For example, once we get a reference sentence '*the sandwiches are terrible*' with negative emotion, a transductive learning method may connect the two sentences by the two kinds of food, e.g. '*salad*' and '*sandwiches*', then use '*terrible*' to express the negative emotion for the food '*salad*'.

From the above discussion, we can see that there are two challenges in applying transductive learning to unsupervised text style transfer: 1) how to find specific samples that are beneficial for the style transfer of the current text; 2) how to use the style expressions in these samples to complete the style transfer process. To tackle these two challenges, we propose a novel **T**ran**S**ductive **S**tyle **T**ransfer (TSST) model. In TSST, a retriever is employed to obtain the required similar samples, which tackles the first challenge. An attention-based encoder-decoder framework is then utilized to combine the specific samples to tackle the second challenge. Specifically, TSST first encodes the original text to a contextual representation and a style-independent embedding. Then either sparse (BM25) or dense retrieval functions (MIPS) are used to find the top-K samples in the target style corpus, which are encoded by the same encoder. After that, a recurrent decoder is utilized to generate transfer text word by word based on the representation of those retrieved samples, contextual representation, and the representations in the last step. To jointly learn the dense retriever, encoder, and decoder, two kinds of objective functions are used in this paper, i.e. retrieval loss and bag-of-words loss.

In summary, our contributions are as follows:

- Facing the inconsistency problem in unsupervised style transfer, we propose a novel transductive learning approach, which avoids learning a general rule but relies on specific samples to complete the style transfer process.
- We design a **T**ran**S**ductive **S**tyle **T**ransfer (TSST) model, which employs a retriever to involve highly related samples to guide the learning of the target style.
- Experiments on two benchmark datasets show that TSST alleviates the inconsistency problem and achieves competitive results against traditional baselines. Our code is available at https://github.com/xiaofei05/TSST.

2 Related Work

Previous unsupervised text style transfer methods can be divided into three categories according to the way they control the text style, e.g. embedding based method, decoder based method, and classifier based method.

The embedding based methods assign a separated embedding for each style to control the style of generated text. Early work tries to disentangle the content and style in the text. They first implicitly eliminate the original style information from the text representation using adversarial training (Shen et al., 2017; Fu et al., 2018; John et al., 2019) or explicitly delete style-related words (Li et al., 2018; Sudhakar et al., 2019; Wu et al., 2019b; Malmi et al., 2020). Then, decode or rewrite the style-independent content with the target style embedding. As a complete disentanglement is unreachable and damaged the fluency of the text, recent approaches (Lample et al., 2019; Dai et al., 2019; Yi et al., 2020; Zhou et al., 2020; He et al., 2020) directly feed original text representation and a separated learned style embedding to a stronger generator, e.g., the attention-based sequence-to-sequence model or Transformer to obtain the style transferred text.

The decoder based methods build a decoder for each style or transfer direction, where the style is implicitly represented as the parameters in the corresponding decoder. The former schema built an independent decoder for each style, which first disentangled the style-irrelevant content from the text and then applied the corresponding decoder to

generate sentences with the target style (Fu et al., 2018; Xu et al., 2018; Prabhumoye et al., 2018; Krishna et al., 2020). The latter built for each transfer direction, which often regarded style transfer as a translation task (Zhang et al., 2018; Gong et al., 2019; Luo et al., 2019; Jin et al., 2019; Wu et al., 2019a; Li et al., 2020). This paradigm reduced the complexity of the learning of style to a certain extent but consumed more resources. It is worth mentioning that the boundaries of embedding controlled and decoder controlled methods are sometimes not very clear, and many studies (Fu et al., 2018; He et al., 2020) consider that they are alternative.

The classifier based methods convert the style by manipulating the latent representation of the text according to a pre-trained classifier. Wang et al. (2019) and Liu et al. (2020) mapped the input sentence into a latent representation, and trained classifiers on this latent space. The latent representation would be edited based on the gradients of the classifier until the predicted style changed; after that, the decoder took the modified representation to generate the desired style sentence. Mai et al. (2020) further expanded this framework to a plug and play scene. Although it has remarkable style accuracy, this method can hardly guarantee content preservation due to the concrete output sharply changes with the latent representation.

We can see that all of these existing methods belong to the inductive learning approach, because they aim to learn a general style transfer rule from the training data, and then apply the rule to the test cases. Due to the lack of parallel corpus for supervision, this inductive learning approach fails to learn an accurate style representation application for various contexts, and may cause some severe inconsistency problems as illustrated before.

3 Transductive Style Transfer

Firstly, we introduce some notations. Consider the unsupervised text style transfer task with M styles, its training set is composed of M single-style subsets $\{D_i\}_{i=1}^{M}$. For an arbitrary input text x in a subset and the target style s_j, the goal of text style transfer is to generate a new sentence y which represents the style s_j while keeping the style-independent content of x as much as possible.

To tackle the aforementioned inconsistency problem, we propose to utilize transductive learning to obtain a context-aware style representation for the style transfer process. Specifically, our proposed transductive style transfer(TSST) model consists of three modules, encoder, retriever, and decoder, as described in Figure 2.

3.1 Encoder

The goal of the encoder is to map the input sentence into hidden representations, to facilitate the following retrieval and generation process. Given the input sentence $x = (w_1, w_2, \ldots, w_n)$, the output of the encoder is a sequence of hidden states,

$$\boldsymbol{H}^{enc} = \text{Encoder}(x), \tag{1}$$

where $\boldsymbol{H}^{enc} = [\boldsymbol{h}_1^{enc}, \boldsymbol{h}_2^{enc}, \ldots, \boldsymbol{h}_n^{enc}]^{\mathrm{T}} \in \mathbb{R}^{n \times d}$, and d is the dimension of the hidden state. Please note that the encoder in our model is very general, and different encoding techniques can be used. Specially, we employ a bidirectional LSTM in our experiments.

3.2 Retriever

The retriever module is introduced to involve the top-K relevant texts in the target style training subset, to facilitate the transductive learning process. In this paper, we adopt both sparse and dense retrieval functions in the retriever.

Sparse Retriever (BM25) BM25 (Robertson and Zaragoza, 2009) is the most famous sparse retrieval function, which has been widely used in information retrieval.

$$\text{BM25}(\boldsymbol{q}, \boldsymbol{d}) = \sum_{w \in \boldsymbol{q}} \frac{\text{IDF}(w) \cdot f(w, \boldsymbol{d}) \cdot (k_1 + 1)}{f(w, \boldsymbol{d}) + k_1 \cdot (1 - b + \frac{b \cdot |\boldsymbol{d}|}{\text{avgdl}})} \tag{2}$$

where k_1 and b are the hyper parameters, $f(w, \mathbf{d})$ represents term frequency of w in document $\mathbf{d}$, $\text{IDF}(w)$ represents inverse document frequency of w, $|\mathbf{d}|$ denotes the document length and avgdl denotes the averaged document length.

After that, the retrieved K texts $\{u_i\}_{i=1}^{K}$ are mapped to latent representations $\boldsymbol{U} = [\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_K}]^{\mathrm{T}} \in \mathbb{R}^{K \times d}$ by the above encoder, where $\boldsymbol{u_i}$ is the final hidden states for u_i.

Dense Retriever (MIPS) A term-based sparse retrieval would have difficulty retrieving such a semantic context, which is essential for style transfer. Recently, dense retrieval methods and their efficient implementation of maximum inner product search (MIPS) (Shrivastava and Li, 2014; Guo et al., 2016; Cai et al., 2021) have been proposed to capture the semantic. For a dense retriever, style-independent

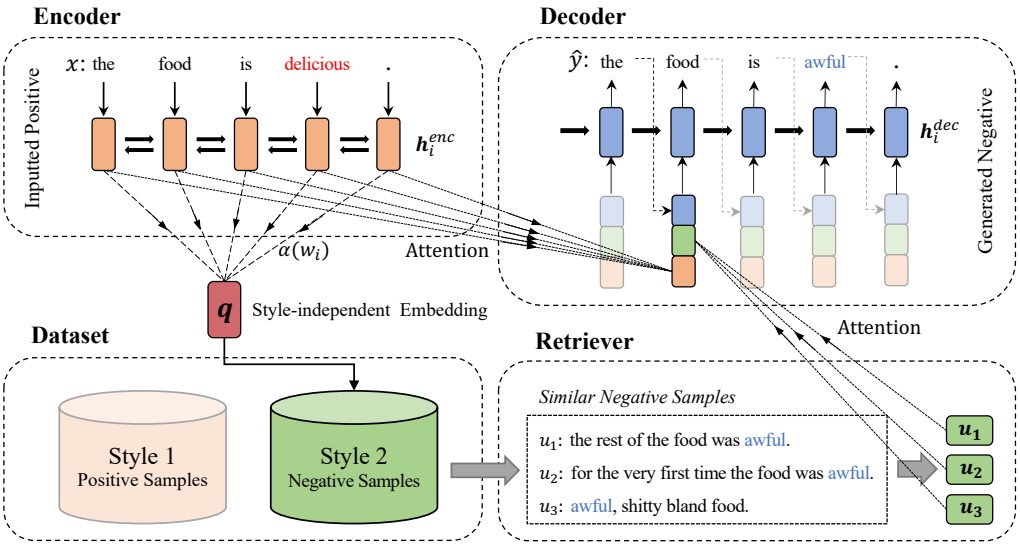

Figure 2: Illustration of our TSST model for binary style transfer as an example.

text embedding is crucial because we rely on this embedding to find similar samples in a different style subset. To this end, the style-independent embedding $\boldsymbol{q}(x)$ of the text x can be represented as a linear combination of the hidden states $\boldsymbol{h}_i^{enc}$:

$$\boldsymbol{q}(x) = \text{Softmax}([\alpha(w_1), \ldots, \alpha(w_n)]) \cdot \boldsymbol{H}^{enc}, \quad (3)$$

where the parameter $\alpha(w_i)$ is the weight for each word w_i. They are initialized as $\alpha(w_i) = 1 - \sum_j |f_{s_j}(w_i) - 1/M|$, where $f_{s_j}(w_i)$ is defined as the count of w_i in the subset of style s_j across its total count in the whole dataset. This initialization assigns a small weight to the discriminative words for each style, whose frequencies in different style subsets vary significantly. Consequently, the embeddings will focus on style-independent words, and help learn the style-independent embeddings.

Based on the text embeddings, the dense retrieval approach is used to retrieve the top-K similar sentences in the target style training subset, where the cosine similarity function is used to measure the similarity. Note that computing text embeddings in the whole training set are time-consuming, so we pre-compute the text embeddings at the beginning and update them after certain training iterations, as inspired by Guu et al. (2020).

After that, the same encoder is employed to obtain the latent representations of the top-K texts, i.e. $\boldsymbol{U} = [\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_K}]^{\mathrm{T}}$, as did in the sparse retriever.

3.3 Decoder

The decoder is used to generated the transferred text word by word. For each step, the inputs of the decoder are composed of three parts: 1) output of the previous step $\hat{y}_{t-1}$, 2) hidden states of x, i.e. $\boldsymbol{c}_t^h$, and 3) latent representations of retrieved samples, i.e. $\boldsymbol{c}_t^u$. The generated text $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m\}$, is obtained by the following equations:

$$\begin{aligned} \boldsymbol{h}_t^{dec} &= \text{Decoder}(\hat{y}_{t-1}, \boldsymbol{c}_t^h, \boldsymbol{c}_t^u), \\ \text{P}(\hat{y}_t | \hat{y}_{<t}, \boldsymbol{H}^{enc}, \boldsymbol{U}; \boldsymbol{\theta}) &= \text{softmax}(\boldsymbol{o}_{\hat{y}_t} \boldsymbol{h}_t^{dec}), \end{aligned} \quad (4)$$

where $\boldsymbol{o}_{\hat{y}_t}$ is the parameters to obtain the predicted probability on the word $\hat{y}_t$, and $\boldsymbol{c}_t^h, \boldsymbol{c}_t^u$ indicates the attended input sentence and retrieved samples, respectively. $\boldsymbol{c}_t^h$ and $\boldsymbol{c}_t^u$ are calculated by two attention modules with different parameters:

$$\begin{aligned} \boldsymbol{c}_t^h &= \text{Attention}(\boldsymbol{h}_{t-1}^{dec}, \boldsymbol{H}^{enc}), \\ \boldsymbol{c}_t^u &= \text{Attention}(\boldsymbol{h}_{t-1}^{dec} \boldsymbol{W}_h, \boldsymbol{U} \boldsymbol{W}_u), \end{aligned} \quad (5)$$

where $\boldsymbol{W}_h \in \mathbb{R}^{d_{dec} \times d_{dec}}$, $\boldsymbol{W}_u \in \mathbb{R}^{d \times d}$ and d_{dec} is the dimension of the decoder's hidden state. The attention module is standard (Bahdanau et al., 2015).

$$\begin{aligned} \text{Attention}(\boldsymbol{h}, \boldsymbol{H}) &= \sum_{j=1}^{n} \frac{exp(e_{ij}) \boldsymbol{H}_j}{\sum_{k=1}^{n} exp(e_{ik})}, \\ e_{ij} &= \boldsymbol{v}^T \tanh(\boldsymbol{W}_d \boldsymbol{h} + \boldsymbol{W}_e \boldsymbol{H}_j), \end{aligned} \quad (6)$$

where $\boldsymbol{v}, \boldsymbol{W}_e, \boldsymbol{W}_d$ are parameters.

Similar to the encoder, various decoding techniques could be used in this step, and we adopt LSTM in our experiments.

3.4 Learning Objectives

Except for the three widely used losses in previous style transfer works, i.e. the reconstruction loss, the cycle reconstruction loss, and the adversarial style loss, we also introduce two more losses related to the retriever, i.e. the retrieval loss and the Bag-of-Word loss. Therefore, for a given input x, its style s_i and the target style s_j, the learning objective function can be represented as:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{cyc} + \mathcal{L}_{adv} + \mathcal{L}_{ret} + \mathcal{L}_{bow}, \quad (7)$$

where $\mathcal{L}_{rec}$, $\mathcal{L}_{cyc}$, $\mathcal{L}_{adv}$, $\mathcal{L}_{ret}$ and $\mathcal{L}_{bow}$ denote the reconstruction loss, the cycle reconstruction loss, the adversarial style loss, the retrieval loss, and the bag-of-words loss, respectively.

Reconstruction Loss According to previous works (Shen et al., 2017; Fu et al., 2018; John et al., 2019), this loss is used to capture the informative features for reconstructing itself:

$$\mathcal{L}_{rec} = -\log P(x|x, s_i). \quad (8)$$

Cycle Reconstruction Loss Cycle consistency is usually included in the loss to improve the preservation of content (Lample et al., 2019; Dai et al., 2019; Yi et al., 2020). For the generated $\hat{y}$, the output of transferring back to the source style should be consistent with x as much as possible:

$$\mathcal{L}_{cyc} = -\log P(x|\hat{y}, s_i), \quad \hat{y} = G(x, s_j), \quad (9)$$

where G is our TSST model.

Adversarial Style Loss If we only use the reconstruction and cycle construction losses, the model merely learns to copy the input to the output. So we employ adversarial training to build style supervision. Specifically, we utilize a classifier with $M+1$ classes as the discriminator C, similar to Dai et al. (2019) and Yi et al. (2020). The first M classes represent the real texts in the datasets, and the $(M+1)$-th class indicates generated fake texts. Since the generated text $\hat{y}$ is expected to be classified as the target style s_j, the adversarial style loss is defined as follows, and the negative gradient of the discriminator is employed to update the model.

$$\mathcal{L}_{adv} = -\log P_C(j|\hat{y}), \quad (10)$$

As for the discriminator C, a loss function $\mathcal{L}_{C_1}$ is usually used in previous works Dai et al. (2019) and Yi et al. (2020). In this paper, we also ask the discriminator to identify the style of the retrieved samples. Therefore, the loss function in our work can be written as:

$$\begin{aligned} \mathcal{L}_C &= \mathcal{L}_{C1} + \mathcal{L}_{C2}, \\ \mathcal{L}_{C1} &= -[\log P_C(i|x) + \log P_C(i|G(x, s_i)) \\ &\quad + \log P_C(M+1|\hat{y})], \\ \mathcal{L}_{C2} &= -[\log P_C(j|Y^{x\to\hat{y}}) + \log P_C(i|Y^{\hat{y}\to x})], \end{aligned} \quad (11)$$

where $Y^{x\to\hat{y}}$ and $Y^{\hat{y}\to x}$ denotes the retrieved samples in the transfer process from x to $\hat{y}$ and from $\hat{y}$ to x, respectively.

To jointly learn the dense retriever with the other parameters and the target style representation from the retrieved samples, we introduce two additional losses to our objective.

Retrieval Loss The retrieval loss is designed to capture the similarity between the style-independent embeddings of the input sentence and the corresponding transferred sentence:

$$\mathcal{L}_{ret} = 1 - \cos(\boldsymbol{q}(x), \boldsymbol{q}(\hat{y})). \quad (12)$$

Bag-of-Words Loss This loss is proposed to encourage the generator to select some new words from the retrieved sentences, making our model pay more attention to the retrieved samples. In this way, the style representation in the target sentence will be well adapted to the context. Let Ω denote a set of new words that appear in the retrieved samples other than the input sentence x, and the bag-of-words loss is defined as:

$$\begin{aligned} \mathcal{L}_{bow} &= \frac{1}{2}(\mathcal{L}_{bow}^{x\to\hat{y}} + \mathcal{L}_{bow}^{\hat{y}\to x}), \\ \mathcal{L}_{bow}^{x\to\hat{y}} &= \frac{1}{\|\Omega\|}\sum\nolimits_{i=1}^{\|\hat{y}\|}\sum\nolimits_{w\in\Omega}\log p_{\hat{y}_i}(w), \\ \mathcal{L}_{bow}^{\hat{y}\to x} &= \frac{1}{\|\Omega'\|}\sum\nolimits_{i=1}^{\|x\|}\sum\nolimits_{w\in\Omega'}\log p_{x_i}(w). \end{aligned} \quad (13)$$

Note that the Bag-of-Words loss is applied to both directions in the cycle reconstruction process.

3.5 Discussion

Please note that there are also some other works to involve a retrieval module to enhance the unsupervised style transfer, e.g. Li et al. (2018); Sudhakar et al. (2019); Jin et al. (2019). The differences between our TSSM and these works are listed as follows. 1) Our TSST model uses the retrieved samples to directly control the style, instead of using them as an external knowledge or pseudo-parallel corpus. 2) The retrieval module in our

Dataset	Styles	Train	Dev	Test
Yelp	Positive	266,041	2,000	500
	Negative	177,218	2,000	500
GYAFC	Formal	51,967	2,247	1,019
	Informal	51,967	2,788	1,332

Table 1: Statistics of Yelp and GYAFC datasets.

TSST can be trained in an end-to-end way (together with the encoder and decoder), to improve the style-independent text retrieval. 3) The retrieved samples influence the decoder word by word. In this way, the information of the retrieved samples will be fully exploited to learn a good context-aware style representation. It may bring no benefit if we just use the retrieved samples without these modifications, as shown in Sudhakar et al. (2019).

4 Experiment

In this section, we conduct experiments to study how well and why the proposed TSST model alleviates the inconsistency problem. Furthermore, a detailed ablation study is demonstrated to show each objective function's contribution to the overall performance.

4.1 Datasets

The experiments are conducted on two well-known transfer tasks, sentiment transfer and formality transfer. The statistics of each datasets are shown in Table 1.

Yelp Dataset[1] is widely used as the benchmark for sentiment transfer. It is collected from restaurant and business reviews, with each text marked as positive or negative. The same pre-processing as (Li et al., 2018) is used in our experiment, and human references are also provided for the test set.

GYAFC Dataset[2] denotes the Grammarly's Yahoo Answers Formality Corpus released by Rao and Tetreault (2018), a typical benchmark for formality transfer. The GYAFC dataset contains formal and informal sentences in two different domains, *Entertainment & Music* and *Family & Relationships*. In this paper, we use the latter one because it is more popular in this area.

4.2 Setups

Our baselines cover three different kinds of inductive learning approaches, as described in Section 2. For the embedding based method, we choose **CrossAlign** (Shen et al., 2017), **StyTrans** (Dai et al., 2019), **PFST** (He et al., 2020) and **StyIns** (Yi et al., 2020). For the generator based method, **MultiDecoder** (Fu et al., 2018) and **DualRL** (Luo et al., 2019) are selected for comparisons. Then, **Revision** (Liu et al., 2020) is considered as the representative of the discriminator based methods. At last, we also compare our model with previous methods involved in retrieval, **DRG** (Li et al., 2018), **IMaT** (Jin et al., 2019), **B-GST** and **G-GST** (Sudhakar et al., 2019). All baselines (including generated results) are directly taken or implemented from their public source codes, so the detailed settings are omitted in our paper.

For our proposed TSST model, we employ the LSTM as our encoder and decoder to ensure the fairness of the experiment compared with previous methods. Following Yi et al. (2020), we pre-train a forward LSTM language model in each dataset and use its parameters to initialize our encoder and decoder. Similar to Yi et al. (2020), the discriminator is a CNN-based classifier with Spectral Normalization (Miyato et al., 2018), with the same word embeddings as the encoder. The word embedding size, hidden state size, and the number of retrieved samples K are set to 256, 512, and 5, respectively. We exclude the trivial candidates the same as the input sentence in the retriever. The embeddings of all sentences for dense retrieval are updated every 200 steps. To demonstrate the effectiveness of the sparse and dense retrievers, we compare them with a random sampling retriever, and the corresponding TSST model is denoted as TSST-random.

4.3 Evaluation Metrics

Previous works mainly focus on evaluating the style transfer methods from the following three aspects, i.e. style transfer accuracy, content preservation, and sentence fluency. Consequently, different automatic evaluation measures such as accuracy, *self*-BLEU, *ref*-BLEU, and perplexity(PPL) are also used in the evaluation. However, all of these metrics cannot well evaluate how well a model alleviates the consistency problem, as we introduced before. So we introduce an additional human evaluation in our experiment.

Automatic Evaluation To evaluate the style transfer accuracy, we first finetune a pre-trained BERT-based (Devlin et al., 2019) classifier on each dataset. The two classifiers achieve 98.6% and 89.9% accuracy on the test set of Yelp and GYAFC,

[1] http://bit.ly/2LHMUsl.

[2] https://github.com/raosudha89/GYAFC-corpus.

Model	Yelp					GYAFC				
	Acc↑	s-BLEU↑	r-BLEU↑	PPL↓	GM↑	Acc↑	s-BLEU↑	r-BLEU↑	PPL↓	GM↑
CrossAlign (Shen et al., 2017)	78.7	16.65	8.11	66	7.09	61.6	2.21	3.25	**37**	3.32
DRG (Li et al., 2018)	88.1	36.75	16.66	100	10.4	58.2	31.57	21.88	103	9.65
MultiDec (Fu et al., 2018)	45.4	40.07	15.07	188	8.51	24.5	16.08	11.95	151	5.54
B-GST (Sudhakar et al., 2019)	82.4	30.82	16.32	156	9.52	-	-	-	-	-
G-GST (Sudhakar et al., 2019)	61.1	45.72	22.08	257	10.27	-	-	-	-	-
DualRL (Luo et al., 2019)	87.9	58.90	28.77	105	13.37	55.5	52.80	43.69	159	12.61
StyTrans (Dai et al., 2019)	86.0	59.46	27.32	154	12.90	60.3	61.15	43.95	168	13.33
IMaT (Jin et al., 2019)	**93.9**	16.92	11.26	**14**	9.08	-	-	-	-	-
Revision (Liu et al., 2020)	90.6	13.23	7.93	21	7.45	39.6	27.64	20.83	66	8.59
PFST (He et al., 2020)	84.6	48.90	23.72	67	12.35	63.9	28.68	21.53	40	10.17
StyIns (Yi et al., 2020)	90.9	53.10	26.09	110	12.80	69.9	61.87	47.80	140	14.32
TSST-random	88.8	**59.64**	28.44	117	13.34	**76.5**	62.84	50.13	108	15.06
TSST-sparse	90.7	59.03	28.71	108	13.46	75.3	**64.03**	50.39	101	**15.15**
TSST-dense	91.8	59.34	**28.89**	108	**13.54**	74.1	63.70	**50.49**	103	15.06

Table 2: Automatic evaluation results on Yelp and GYAFC dataset. Acc denotes the accuracy of generated samples judged by the pre-trained classifier. s-BLEU and r-BLEU stands for *self*-BLEU and *ref*-BLEU, respectively. **Bold** denotes the best value in terms of each metric.

Model	Yelp				GYAFC			
	Sty	Cont	Flu	Cons	Sty	Cont	Flu	Cons
DualRL	3.70	4.34	**4.29**	1.38	1.84	2.92	3.11	0.37
StyTrans	2.98	4.18	3.54	1.04	2.43	3.27	3.53	0.49
PFST	3.21	3.69	4.18	1.26	1.97	2.28	3.41	0.31
StyIns	3.43	4.12	3.82	1.27	1.36	1.62	2.21	0.11
TSST-random	3.64	4.34	3.84	0.99	**2.84**	3.55	3.59	0.51
TSST-sparse	3.52	**4.42**	4.02	1.33	2.74	3.48	3.67	**0.75**
TSST-dense	**3.85**	4.40	4.09	**1.46**	2.70	**3.59**	**3.69**	0.74

Table 3: Human evaluation results on Yelp and GYAFC dataset. The Krippendorff's alpha of human rating on two datasets is 0.76 and 0.73, respectively, indicating acceptable inter-annotator agreement.

respectively. Then these classfiers are used to predict the style label of the generated transferred sentences, and the classification accuracy acts as the style transfer accuracy. Both *self*-BLEU and *ref*-BLEU are used for content preservation evaluations. The former is the BLEU score between transferred sentences and source sentences, while the latter is between transferred sentences and human references. Following Dai et al. (2019) and Yi et al. (2020), we train a 5-gram language model KenLM (Heafield, 2011) for each style to measure the lanuguage fluency by the perplexity (PPL) of transferred sentence. In addition, we report the geometric mean (GM) of Acc, *self*-BLEU, *ref*-BLEU and $\frac{1}{\log \text{PPL}}$ as the overall performance.

Human Evaluation We recruit three annotators who have high-level language skills for human evaluation. We choose four models with the highest GM scores and three types of TSST models in this experiment. Following previous works (Dai et al., 2019; Yi et al., 2020; Liu et al., 2020), we randomly selected 100 generated sentences (50 for each style) in the test set and the annotators are required to score sentences from 1 to 5, in terms of each aspect, i.e. style transfer accuracy (Sty), content preservation (Cont), and sentence fluency (Flu), where 1 is the lowest and 5 is the highest. In addition, they need to evaluate the consistency of the style words and other contexts in each generated sentence. To make it more clear, consistency (Cons) focuses on judging whether the modified parts are consistent with the retained content, while the fluency only focuses on the grammatical errors. The consistency is rated from 0 to 2, where 0, 1, and 2 stands for inconsistent, unsure, and consistent, respectively.

4.4 Experimental Results

Automatic Evaluation As listed in Table 2, we can see that our transductive learning models significantly improve the overall performance on both datasets, as compared with the inductive learning baselines. As for the four specific evaluation measures, our model achieves better results in terms of three of them, i.e. Acc for style transfer accuracy, s-BLEU, and r-BLEU for content preservation, and all our models are able to generate sentences with relatively low perplexity. Though previous models achieve the best on a single metric, a significant drawback can be found on another metric. For the TSST-random, although target-style relevant words in the random samples are not consistent with the input content, the few modifications and the use of target-style strong relevant words will still make the random baseline achieve the high automatic metrics. Thus we need human evaluation.

Human Evaluation Results are shown in Table 3. Firstly, the comparison results are consistent with the automatic evaluation results in terms of both style accuracy, content preservation, and

Model	Negative to Positive	Positive to Negative
input	it 's just too expensive for what you get.	mustard beef ribs are a must.
DualRL	it 's just fun for what you get.	mustard beef ribs are a mess.
StyTrans	it 's just too expensive for what you get.	mustard beef ribs are a negative.
StyIns	it 's just very fun for what you get.	mustard beef ribs were a total.
PFST	it 's just right for what you get.	gross.
Revision	it 's too expensive for what you get and it 's always good.	the beef ribs are a must do not have a bbq beef.
TSST-dense	it 's just really **reasonable** for what you get.	mustard beef ribs are **a joke**.
Retrieved Samples	u_1 - the price is **reasonable** for what you get.	u_1 - garlic mashed potatoes were **a joke**.
	u_2 - prices are very **reasonable** for what you get.	u_2 - all olive gardens are **a joke**.
	u_3 - prices are **reasonable** for the qualiy of what you get.	u_3 - the ribs are over cooked.
	u_4 - good food , a little expensive for what you get.	u_4 - our food was barely edible.
	u_5 - definitely a good price for what you get.	u_5 - cold grits are n't a treat.

Table 4: Case Study. Red denotes *positive* expressions, blue denotes *negative* expressions, and **bold** denotes the expression taken from retrieved samples.

fluency, indicating the reliability of our human evaluation. More importantly, our TSST-sparse and TSST-dense models achieve the highest consistency score, as compared to other baselines, showing the superiority of our transductive learning approach in tackling the inconsistency problem. In contrast, the consistency will become worse if the retrieved samples are not related to the input, as shown in TSST-random, which further demonstrate the soundness of our approach. Comparing TSST-sparse and TSST-dense, we can see that joint learning retriever yields better consistency results, which is accordant with previous studies.

Case Study To better understand what transductive learning bring to text style transfer task, Table 4 shows some transferred examples from the Yelp test set. For the first example to transfer from negative to positive, DualRL and StyIns are able to capture the style transfer but with inappropriate expressions, e.g. '*fun*' or '*very fun*'. StyTrans and Revision fail to do the style transfer, either just copy or use negative expressions. Our TSST-dense model produces perfect results by using 'reasonable' as the transferred style expression, learned from the retrieved examples, as shown in the table. Similar results are observed for the second example to transfer positive to negative. More importantly, although there is no retrieved examples containing exactly the same phrase '*mustard beef ribs*', our model still capture the negative pattern like '*[food] is a joke*' to complete the style transfer process.

Ablation Study To study the role of each loss function in the objective function, we remove them one at a time and train the model from scratch. Due to the high cost of the human evaluation, we only report the automatic results on the Yelp dataset, as shown in Table 5. We can see that each loss contributes to the performance of the model, and their combination performs the best.

Model	Acc↑	s-BLEU↑	r-BLEU↑	PPL↓	GM↑
TSST-dense	91.8	59.34	**28.89**	**108**	**13.54**
$-\mathcal{L}_{rec}$	52.7	0.40	0.37	N/A	1.01
$-\mathcal{L}_{adv}$	81.3	18.96	8.48	N/A	6.67
$-\mathcal{L}_{cyc}$	89.0	52.07	24.71	134	12.37
$-\mathcal{L}_{ret}$	89.6	59.36	28.65	110	13.42
$-\mathcal{L}_{bow}$	89.3	55.26	26.43	121	12.84
$-\mathcal{L}_{C1}$	**95.6**	7.6	3.9	N/A	4.47
$-\mathcal{L}_{C2}$	87.4	**61.41**	28.58	112	13.43

Table 5: Ablation study on the Yelp dataset, where '$-$' means removing one of the loss terms in the objective functions, and N/A is a very large value.

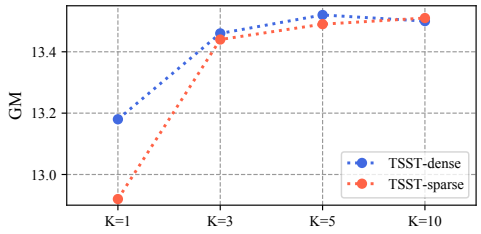

Figure 3: The influence of the number of retrieved samples K on the Yelp dataset.

We also conduct experiments to explore the influence of different numbers of retrieved samples, as shown in Figure 3. Specifically, we test four different values, i.e. $K = 1, 3, 5, 10$. We can see that the overall performance of the model gradually increases with the increase of K, and become stable for 3, 5, and 10. For the sake of both effectiveness and efficiency, we set $K = 5$ in our experiments.

5 Conclusions and Future Work

Previous style transfer models are mainly based on inductive learning approach, thus suffer from the inconsistent style expression problem with lack of the parallel corpus as supervisions. To tackle this problem, we propose a novel transductive learning approach for unsupervised text style transfer. The key idea of our TSST model is to learn context-aware style expressions via retrieved samples from

the target style datasets. Experimental results on two typical style transfer tasks show that TSST significantly improves the performances in terms of both automatic and human evaluation.

Our proposed transductive learning approach is very general, and this work mainly focus on the embedding-based methods. In future, we plan to extend our approach to other methods, such as decoder-based and discriminator-based methods. Moreover, we will try more powerful retrieval methods, such as DPR (Karpukhin et al., 2020).

Ethical Considerations

We honor and support the ACL code of Ethics. The paper focuses on style transfer, which aims to change the style of the text while preserving the semantic content. We recognize the style transfer methods may be misused to generate misinformation, e.g. fake customer reviews. However, the style transfer methods can also provide strong support for mitigating harmful biases in online information, e.g. the transfer from offensive to non-offensive (Nogueira dos Santos et al., 2018; Tran et al., 2020) and from biased to neutral (Pryzant et al., 2020). Overall, it is still meaningful to continue research into this work on the basis of predecessors. Simultaneously, the datasets we used in this paper are all from previously published works and do not involve privacy or ethical issues.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) under Grants No. 61906180, No. 61773362 and No. 91746301, National Key R&D Program of China under Grants 2020AAA0105200 and the Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202033).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. 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)*, pages 7307–7318, Online. Association for Computational Linguistics.

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

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

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670. AAAI Press.

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

Ruiqi Guo, Sanjiv Kumar, Krzysztof Choromanski, and David Simcha. 2016. Quantization based fast inner product search. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 482–490. JMLR.org.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

Parag Jain, Abhijit Mishra, Amar Prakash Azad, and Karthik Sankaranarayanan. 2019. Unsupervised controllable text formalization. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6554–6561. AAAI Press.

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)*, pages 3097–3109, Hong Kong, China. Association for Computational Linguistics.

Thorsten Joachims et al. 1999. Transductive inference for text classification using support vector machines. In *Icml*, volume 99, pages 200–209.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Xiao Li, Guanyi Chen, Chenghua Lin, and Ruizhe Li. 2020. DGST: a dual-generator network for text style transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7131–7136, Online. Association for Computational Linguistics.

Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. 2020. Revision in continuous space: Unsupervised text style transfer without adversarial learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8376–8383. AAAI Press.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5116–5122. ijcai.org.

Florian Mai, Nikolaos Pappas, Ivan Montero, Noah A. Smith, and James Henderson. 2020. Plug and play autoencoders for conditional text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6076–6092, Online. Association for Computational Linguistics.

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

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social

media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. Association for Computational Linguistics.

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

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Information Retrieval*, 3(4):333–389.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6830–6841.

Anshumali Shrivastava and Ping Li. 2014. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2321–2329.

Yixuan Su, Wang Yan, Deng Cai, Simon Baker, Anna Korhonen, and Nigel Collier. 2021. Prototype-to-style: Dialogue generation with style-aware editing on retrieval memory. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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

Minh Tran, Yipeng Zhang, and Mohammad Soleymani. 2020. Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2107–2114, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. Wiley-Interscience.

Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11034–11044.

Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019a. A hierarchical reinforced sequence operation method for unsupervised text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4873–4883, Florence, Italy. Association for Computational Linguistics.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019b. " mask and infill": Applying masked language model to sentiment transfer. *arXiv preprint arXiv:1908.08039*.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia. Association for Computational Linguistics.

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

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *arXiv preprint arXiv:1808.07894*.

Chulun Zhou, Liangyu Chen, Jiachen Liu, Xinyan Xiao, Jinsong Su, Sheng Guo, and Hua Wu. 2020. Exploring contextual word-level style relevance for unsupervised style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7135–7144, Online. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the*

Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 730–739. AAAI Press.