Mask and Regenerate: A Classifier-based Approach for Unpaired Sentiment Transformation of Reviews for Electronic Commerce Websites

Shuo Yang
yangshuo@toki.waseda.jp

Abstract

Style transfer is the task of transferring a sentence into the target style while keeping its content. The major challenge is that parallel corpora are not available for various domains. In this paper, we propose a Mask-And-Regenerate approach (MAR). It learns from unpaired sentences by modifying the word-level style attributes. We cautiously integrate the deletion, insertion and substitution operations into our model. This enables our model to automatically apply different edit operations for different sentences. Specifically, we train a multilayer perceptron (MLP) as a style classifier to find out and mask style-characteristic words in the source inputs. Then we learn a language model on non-parallel data sets to score sentences and remove unnecessary masks. Finally, the masked source sentences are input to a Transformer to perform style transfer. The final results show that our proposed model exceeds baselines by about 2 per cent of accuracy for both sentiment and style transfer tasks with comparable or better content retention.

1 Introduction

A text style is a feature that specifies text. The objective of style transfer is to rewrite a given sentence into a target-style domain with the preservation of semantic content. In this paper, we follow the opinion (Fu et al., 2018; Prabhumoye et al., 2018) that textual sentiment should also be treated as styles and conduct experiments to transfer sentiments of sentences collected from three electronic commerce websites. E.g. "*The food here is delicious.*" (Positive) → "*The food here is gross.*" (Negative)

A key issue is that the lack of available parallel data has a considerable impact on the use of supervised learning. It results in the majority of recent studies concentrating on unpaired text transfer approaches (Shen et al., 2017; Luo et al., 2019; Krishna et al., 2020). Compare with related work,

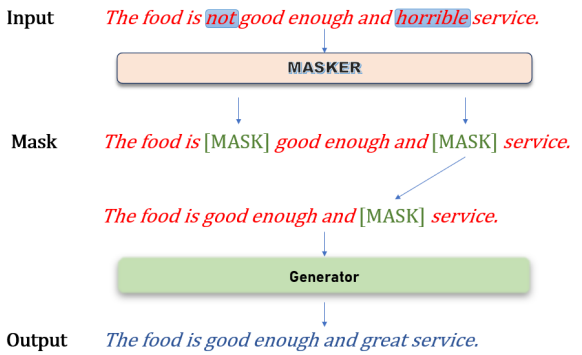

Figure 1: The proposed Mask-and-Regenerate approach. In this example, we transfer a negative sentence to a positive one. The [MASK] of the word 'not' has been removed by a language model.

methods based on word-level operations (Li et al., 2018; Wu et al., 2019a) have become one of the most frequently used approaches because they ensure high content preservation.

The approach we introduce in this paper mainly follows two works, the Delete-Retrieval-Generate (DRG) model (Li et al., 2018) and the Tag-and-Generate model (TAG) (Madaan et al., 2020). The motivation behind the DRG model is to delete style-characteristic words by computing the frequency of occurrence of words, retrieve one similar sentence in the target style corpus and generate a new sentence which is the result of crossing the two sentences. By following the idea of DRG, the TAG model is proposed. The TAG model calculates tf · idf scores (Ramos et al., 2003) to determine style-characteristic words and it includes a Tagger to insert a special symbol '[TAG]' into the input sentences, that will be filled by target-style-characteristic phrases. We identify the following weak points in these models:

1. The hypothesis that the frequency of a word is indicative of style is not always true.

2. Edit operations are not considered equally for all input sentences. Even in the same data set,

Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media (SocialNLP 2022), pages 1 - 10

for parts of sentences, deletion may be the best option to apply, whereas insertion or substitution may be the best for others. For example, we can transfer a sentence from negative to positive by inserting the word 'never' under certain conditions, e.g. "*I will give it up.*" → "*I will never give it up.*" while deletion can also realize a negative to positive transformation, e.g. "*The dipping sauce is too sweet.*" → "*The dipping sauce is sweet.*"

3. Retrieval module might not find suitable sentences. This may result in poor semantic content preservation. The results reported in this paper demonstrate this problem.

To tackle the above problems, we suggest that:

1. We use neural networks instead of statistical methods for the recognition of style-characteristic words. More precisely, we train a style classifier on the two data sets. For each source sentence, we mask each word in it and input it into the classifier. Masks that cause larger variations in the classifier logits correspond to words with higher style contributions. This is based on the fact that if a word is relevant to the style, then masking this word will increase the probability that the source sentence be classified into the wrong style domain. By masking these words, we arguably get a representation of content that is independent of the source style.

2. When multiple possible solutions exist for an input sentence, we propose that the selection of the optimal solution depends on their semantic fluency. For that, we learn a language model (LM) to validate the masks. If a mask-independent content representation already tends to get a low perplexity on the target data set, it means that deletion is a better choice for this sentence than substitution. In this situation, the masks are removed directly.

3. We generate a new sentence without retrieving similar sentences. We do not use any templates that have been summarised from retrieved sentences. As an improvement approach, extracted content representations are input to a Transformer (Vaswani et al., 2017) to rewrite sentences with the target style. The Transformer is designed to fill in the masks with style-characteristic phrases, insert words or retain the original version.

Our main contributions are as follows:

- We propose a novel approach to recognize style-characteristic words. For that, we rely on a neural classifier. To our best knowledge, previous studies of style transfer have not dealt with word recognition using masking models.

- We propose to use an LM to select edit operations (insertion, substitution and deletion) for different inputs. In such a mode, all possible situations for the transformation are covered.

- The results show that our approach outperforms baselines in terms of accuracy with comparable or higher BLEU scores.

2 Related Work

2.1 Style Transfer in Latent Space

Disentangling the style and content is a general idea in unpaired text transfer. Shen et al. (2017) proposed a cross-aligned auto-encoder training method to align transferred samples with target style samples at a shared latent content distribution level across different corpora. Fu et al. (2018) proposed techniques to use adversarial approaches to extract pure content representations and decode them into sentences. Models based on manipulating representations in the latent space (Hu et al., 2017; Prabhumoye et al., 2018) were proposed in the same period. Nevertheless, it is reported that the extraction of style information in a latent space can be very difficult (Elazar and Goldberg, 2018).

2.2 Style Transfer by Modifying Words

In contrast to operations in latent space, recent representative methods are proposed to extract style-independent content representations (Sudhakar et al., 2019; Zhang et al., 2018). Li et al. (2018) presented that a Delete-Retrieve-Generate pipeline also performs well in sentiment transfer tasks. Nevertheless, the retrieving was reported as an unnecessary step (Madaan et al., 2020). Models based on the edit operations show better results (Wu et al., 2019b; Reid and Zhong, 2021). However, the traditional attribute word recognition methods used only focused on word counting. Furthermore, these studies ignored the basis of selecting edit operations.

In this paper, we mainly follow the second approach which assumes the existence of style-characteristic words. We propose a new style-characteristic word recognition method and use a language model to score sentences to determine specific operations.

3 Methodology

We are given a sentence set $X_A = (x_A^{(1)}, ..., x_A^{(M)})$ with the source style A and another sentence set $X_B = (x_B^{(1)}, ..., x_B^{(N)})$ with the target style B. The sentences in these two sets are non-parallel, i.e., $x_A^{(i)}$ does not correspond to $x_B^{(i)}$. The objective is to generate a new set of sentences $\hat{X} = (\hat{x}^{(1)}, ..., \hat{x}^{(M)})$ in the domain of B, where $\hat{x}^{(i)}$ is the result of transferring $x_A^{(i)}$ into style B.

For an overview, we train two independent modules called the Masker and the Generator respectively. The Masker consists of a text MLP and an LM. For an input sentence $x_A^{(i)}$, the Masker masks or deletes style-characteristic words to generate a content representation sequence z_A. The generator is a standard Transformer which is used to insert style-characteristic words into the sequence z_A and replace masks with attribute words of style B.

3.1 Where to Mask?

We propose to use a trained style classifier f_ϕ and an LM to mask words, which is more effective for retaining plain and less style-indicative words. We train the classifier f_ϕ on the two sets to classify sentences to two different styles. The loss function is shown in the Formula (1).

$$\mathcal{L}_{\text{CLS}}(\phi) = -\sum_j \log P(y_j|x_j;\phi) \quad (1)$$

where x_j is the j-th example in a train set and y_j is the style label for x_j.

Inspired by BERT (Devlin et al., 2019), we select a mask-based approach for its reliability and validity. In particular, for a source sentence with k words, $x_A = (w_1, ..., w_k)$, we replace each of them with a special symbol [MASK] and input the masked sentence to the classifier to compute the probability that the classifier classifies this sentence to the target style. We first calculate a distribution $\eta(w_j)$ on sentence x_A to reflect the style contribution of each word w_j.

$$\eta(w_j) = P(B|x_A^{\text{MASK}(j)};\phi) \quad (2)$$

Here, $x_A^{\text{MASK}(j)}$ stands for the sentence x_A with word w_j replaced with a [MASK].

Our objective of this stage is to get the content representation z_A from the input sentence x_A. For that, we mask the word with the highest style contribution in sentence x_A. We repeat this operation until style A cannot be clearly distinguished from the masked sentence by the classifier. Here, we assume that the masked sentence can be regarded as a content representation of the input sentence.

Notice that, if the masking operation cannot extract z_A from x_A, which indicates that there is no obvious style-characteristic word in x_A, then the words in x_A should not be masked. In such a case, the transformation should mainly be performed by insertion. Similarly, if x_A is already judged in the style domain B, it should also not be masked. In this situation, it is possible that x_A is a mistakenly classified sample in the used corpus.

The second step is to tell whether it is necessary to retain masks in z_A. A widespread acknowledgement is that there is not a consistent one-to-one match between each input sentence and each output sentence. For example, an input negative sentence "*I am not really impressed.*", the content representation "*I am [MASK] really impressed.*" can be transferred to "*I am really impressed.*" or "*I am really really impressed.*". The former sounds more natural than the latter.

To make transferred sentences more fluent, we train a 5-gram language model (Heafield, 2011) and use it to score a generated sentence by its probability. If z_A gets a higher score than x_A, then the mask in z_A should not be held anymore. Since we consider insertion as a reverse operation of deletion, the scores computed by the LM are only used to decide whether deletion or substitution should be performed. For a sentence x_A with j words, we compute the probability of it as its score by using Formula (3).

$$P(x_A) = \prod_j P(w_j|w_{j-4}, ..., w_{j-1}), \quad (3)$$

where $P(w_j|w_{j-4}, ..., w_{j-1})$ is approximated by word frequency counting. Here, the LM used was learned on the target style sentence set X_B.

3.2 How to Transfer?

For an input content representation z_A from the Masker, we purpose to learn a mapping function to transfer it into the target style domain instead of retrieving other sentences.

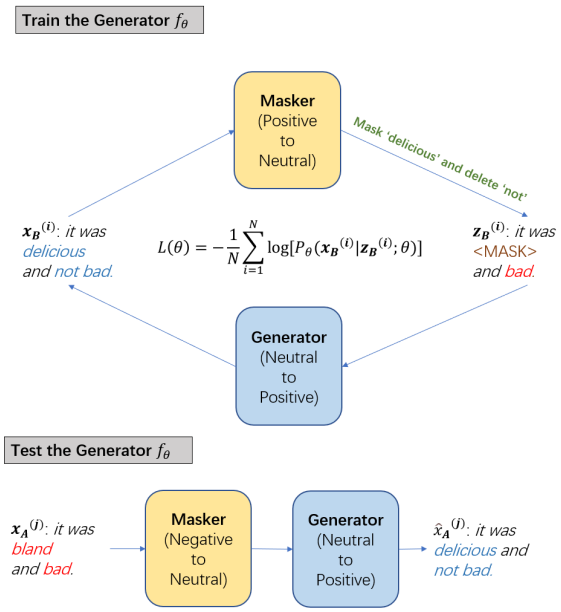

Figure 2: The training and testing stages of the generator. The generator learns to rebuild the original version of x_B from its content representation z_B.

We introduce a reconstruction loss (Luo et al., 2019; Madaan et al., 2020) to train the generator. Specifically, we first generate a content representation z_B of a sampled sentence x_B and treat z_B, x_B as a sentence pair. With the sentence pair, we train a generator f_θ to transfer x_B from its content representation z_B to its original version x_B.

$$\hat{x_B} = f_\theta(z_B), \tag{4}$$

where the generated sentence $\hat{x_B}$ is expected to be the same as x_B.

For a content representation z_A created from sentence x_A, by inference, the trained classifier cannot tell the source style A accurately. Therefore, if we apply f_θ to z_A, the output $\hat{x_A}$ will have the attribute of style B arguably. The loss function of the generator is given in Formula (5).

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log[P(x_B^{(i)}|z_B^{(i)};\theta)] \tag{5}$$

We now give a brief analysis of how these edit operations are respectively used in our model.

The first simple case is when the Masker module does not delete any [MASK] after masking style-characteristic words in every sentence. In this situation, the generator is only trained to fill in the masks. For example, in a sentiment transfer task, the generator learns how to substitute these [MASK] in the content representations z_A with emotional words or phrases. In this case, the transformation is performed by substitution.

For the transfer tasks which are expected to be mainly performed using deletion operations, all of the masks in z_A are deleted. In this case, even if the generator still learns how to fill in the masks, with no masks in the input Z_A, the generator will only learn to copy a sequence to itself. Therefore, the transformation is mainly performed by the Masker.

For the transfer tasks which are expected to be mainly performed by insertion operations, we perform them through an opposite method of the deletion pattern. In training steps, the generator learns how to insert words into z_B to get x_B, with the parallel relation between x_B and z_B. For example, "*That's not bad.*" (x_B) $\rightarrow$ "*That's [MASK] bad.*" $\rightarrow$ "*That's bad.*" (z_B) In practice, when the generator encounters a sentence "*That's bad.*", it will insert the word "*not*" to it automatically.

For other tasks which are in a mixed mode, the above three approaches are performed automatically by the model to find the optimal solution. To summarize, the training process of the generator is shown in Figure 2. Note that the top yellow Masker and the bottom one are in reverse order.

4 Experiments

4.1 Data Sets Used

We test our proposed method on 3 data sets for sentiment transfer and 1 data set for formality transfer. Statistics of the used data sets are shown in Table 1.

Yelp The Yelp data set is a collection of reviews from Yelp users. It is provided by the Yelp Data set Challenge. We use this data set to perform sentiment transfer between these positive and negative business remarks.

Amazon Similar to Yelp, the Amazon data set (He and McAuley, 2016) consists of labelled reviews from Amazon users. We used the latest version provided by (Li et al., 2018).

IMDb The IMDb Movie Review (IMDb) contains positive and negative reviews of movies. We use the version provided by Dai et al. (2019), which is created from previous work (Maas et al., 2011).

GYAFC The Grammarly's Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault, 2018) is a parallel corpus of informal and formal sentences. To demonstrate the situation of unsuper-

Category	Sentiment transfer						Formality transfer	
Data set	**Amazon**		**Yelp**		**IMDb**		**GYAFC**	
	Positive	Negative	Positive	Negative	Positive	Negative	Formal	Informal
Train set	266,041	177,218	277,228	277,769	178,869	187,597	51,967	51,967
Dev. set	2,000	2,000	985	1,015	2,000	2,000	2,247	2,788
Test set	500	500	1,000	1,000	1,000	1,000	1,019	1,332

Table 1: Statistics of the used data sets. 'Dev.' denotes 'development'. The Yelp, Amazon and IMDb data sets are used for sentiment transfer. The GYAFC data set is used for formality transfer.

vised learning, we shuffle all of the used sentences in training.

4.2 Baselines

We select 5 style transfer models as baselines for sentiment transfer comparison and 2 additional models for formality transfer comparison. These 7 baselines can be broadly divided into two categories. The first category consists of a Cross-Align model (Shen et al., 2017) a Style-Transformer (Dai et al., 2019) a DualRL (Luo et al., 2019) model and a DGST (Li et al., 2020) model. These models mainly transfer sentences in a latent space. The second category consists of a DRG (Li et al., 2018) model, a TAG model (Madaan et al., 2020) and an LEWIS model (Reid and Zhong, 2021). These models are mainly based on the substitution of words.

4.3 Automated Evaluation Metric

Transfer accuracy and content preservation are currently the most commonly considered aspects in evaluation. Following standard practice, we consider the following metrics.

Transfer Accuracy Accuracy is considered one of the most important evaluation metrics (Cao et al., 2020; Zhou et al., 2020). It stands for the successful transfer rate. We train a self-attention based convolutional Neural Networks (CNN) as the evaluation classifier f_ω to calculate accuracy. The accuracy is the probability that generated sentences $\hat{X_A}$ are judged to carry the target style B by the trained classifier f_ω. The computation of accuracy is shown in (6).

$$\text{Accuracy} = P(B|\hat{X_A};\omega) \tag{6}$$

Notice that, to avoid an information leakage problem, the evaluation classifier is completely different from the one, i.e., f_ϕ, we used in the training period.

Here, our classifier was able to classify samples with success rates of 83.2%, 98.1%, 97.0% and 84% on the Amazon, Yelp, IMDb and GYAFC datasets, respectively. We understand that the automatic measures via our classifiers may not be convincing enough for the Amazon and GYAFC datasets, whereas quality issues in the two datasets, e.g. misclassification of samples, result that we cannot find a classifier with high accuracy in related work.

Content Preservation BLEU (Papineni et al., 2002) measures the similarity between two sentences at the lexical level. In most recent studies, two BLEU scores are computed: self-BLEU is the BLEU score computed between the input and the output; ref-BLEU is the BLEU score between the output and the human reference sentences (Lample et al., 2019; Sudhakar et al., 2019). We use NLTK (Bird et al., 2009) to calculate them.

4.4 Human Evaluation

Since the use of automatic metrics might be insufficient to evaluate transfer models. To further demonstrate the performance, we select outputs from the two similar models we introduced, i.e., the DAG model and the TAG model, to carry out a human evaluation of the Yelp data set (a popularly used corpus).

We hired 12 paid workers with language knowledge to participate in it. By following (Dai et al., 2019), for each review, we show one input sentence and three transferred samples to a reviewer. Reviewers were asked to separately select the best sentence in terms of three aspects: the degree of the target style, the content preservation and the fluency. We also offer the option "No preference" for concerns about objectivity. Furthermore, we ensure that transferred samples are anonymous to all reviewers in the whole process.

Model	Amazon			Yelp			IMDb	
	ACC.	s-BLEU	r-BLEU	ACC.	s-BLEU	r-BLEU	ACC.	s-BLEU
DRG (Li et al., 2018)	52.2%	57.89 ± 2.19	32.47 ± 12.68	84.1%	32.18 ± 2.05	12.28 ± 1.33	55.8%	55.40 ± 1.79
StyTrans (Dai et al., 2019)	67.8%	82.07 ± 1.56	32.88 ± 2.47	92.1%	52.40 ± 2.14	19.91 ± 2.01	86.6%	66.20 ± 1.55
DGST (Li et al., 2020)	59.2%	83.02 ± 1.25	**42.20 ± 22.37**	88.0%	51.77 ± 2.41	19.05 ± 1.89	70.1%	**70.20** ± 1.42
TAG (Madaan et al., 2020)	79.4%	58.13 ± 1.46	25.95 ± 1.86	88.6%	47.14 ± 2.23	19.76 ± 1.45	N/A	N/A
DIRR (Liu et al., 2021)	62.7%	66.63 ± 2.51	32.68 ± 2.25	91.2%	56.56 ± 1.89	25.60 ± 2.33	83.5%	65.96 ± 1.12
LEWIS (Reid and Zhong, 2021)	71.8%	65.53 ± 1.44	30.61 ± 1.57	89.4%	54.67 ± 1.62	23.85 ± 1.57	N/A	N/A
MAR (Ours)	**80.2%**	**83.42 ± 1.46**	41.21 ± 23.54	**93.9%**	53.32 ± 1.86	22.90 ± 2.01	**87.8%**	66.12 ± 1.33

Table 2: The test results on 3 data sets (sentiment transfer) with 0.95 confidence level. "ACC." stands for Accuracy, "s-BLEU" stands for self-BLEU and "r-BLEU" stands for ref-BLEU. We report the results of baselines by running their official codes or evaluating their official outputs.

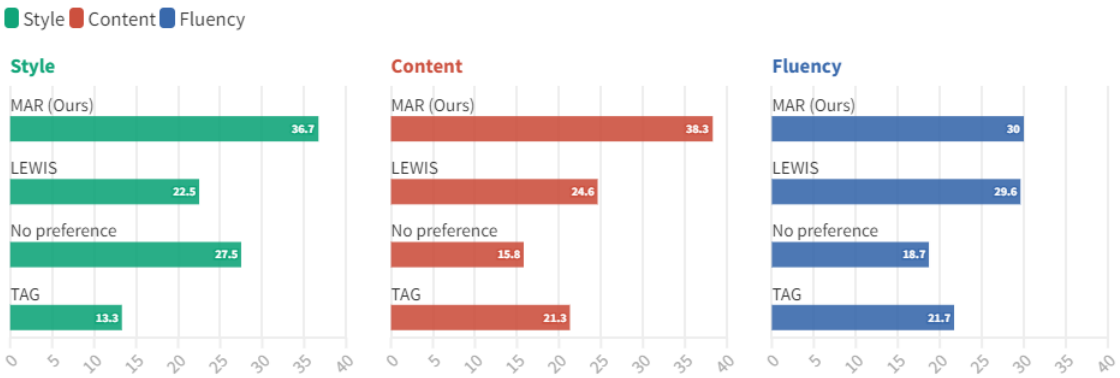

Figure 3: Results of human evaluation of sentences produced by three different models in terms of style, content and fluency. Following standard practice (Dai et al., 2019; Madaan et al., 2020), we randomly selected 100 sentences for evaluation.

4.5 Details

We pre-process the input data to mini-batches with a batch size of 64. All the encoders and decoders in the Transformers used in this paper are made up of a stack of 6 layers. For each layer, it has 8 attention heads and a dimension of 512. The MLP used in training has 4 layers with the same dimension of 512 for each layer. For training steps, the Adam algorithm (Kingma and Ba, 2015) with a learning rate of 0.0001 is employed to update the used models. We use a greedy algorithm to sample words from the probability distribution of the generator logits.

5 Results

5.1 Analysis

Table 2 compares the experimental data obtained on 3 data sets for sentiment transfer. Our proposed model obtains relatively better transfer accuracy than the other 5 models.

For the Amazon data set, our proposed model surpasses the state-of-the-art approach for accuracy and self-BLEU. An interesting aspect is that the DGST model shows a high self-BLEU, but the outputs are far away from the target style domain. We notice that there are no significant differences between the inputs and the outputs with the DGST model. For the Amazon data set, the DGST model merely learns how to copy sentences from inputs to outputs in lots of cases.

For the Yelp data set, our proposed model outperforms the baselines and gets an accuracy of 93.9. In terms of content preservation, our model performs closely to the state-of-the-art model (about 1 per cent) with a self-BLEU of 53.32 and ref-BLEU of 22.90. As all of the models achieved relatively good transfer results on the Yelp data set, we carry out an ablation study and a human evaluation in the next section.

For the IMDb data set, the average sentence length of the IMDb data set is much longer than in the first two data sets, but the number of sentences is much less. In this situation, it is difficult to perfectly train a classifier. This leads to the fact that the Masker in our proposed model tends to mask more words to ensure that the content representation z_A does not contain any emotional words. Theoretically, these operations result in a low self-BLEU. We conclude that our proposed model favours accuracy over self-BLEU scores. Because the IMDb

Yelp	Positive to negative	Negative to positive
Input	it is a cool place , with lots to see and try .	unfortunately , it is the worst .
DRG	it is my waste of time , with lots to try and see .	tender and full of fact that our preference menu is nice and full of flavor .
DGST	it is a sad place , with lots to see and try .	overall , it is the best .
LEWIS	it is a very busy place , with lots to see and try .	cajun food , it is the best !
Ours	it is a horrible place , with nothing to see and try .	wow , it is the best .
Amazon	**Positive to negative**	**Negative to positive**
Input	i won t be buying any more in the future .	because it is definitely not worth full price .
DRG	i won t know how i lived without this in the future .	because it is worth the full price and i am happy with it .
DGST	i won t be buying any more in the future .	because it is definitely not worth full price .
LEWIS	i won t be buying any more in the future . highly recommended .	because it is definitely well made and worth full price .
Ours	i will be buying more in the future .	because it is definitely worth full price .
IMDb	**Positive to negative**	**Negative to positive**
Input	i rate this movie 8/10 .	please , do n't see this movie .
DRG	i rate this movie an admittedly harsh 4/10 .	please , told every one to see this movie .
DGST	i rate this movie 1/10	u , do n't see this "
Ours	i rate this movie 2/10 .	please , you must see this movie .

Table 3: Sentences sampled from sentiment transfer data set. Red text stands for failed style transformation, brown text stands for poor content preservation and blue text stands for suitable transformation.

data set has no human reference, we cannot report a ref-BLEU score in Table 2.

Table 4 shows the result for GYAFC data set. The GYAFC is a formality transfer data set, so it is listed separately. On the GYAFC data set, our proposed model showed strengths in both transfer accuracy and content preservation. However, transfer between formal and informal styles is a very challenging task even for humans. This leads to poor performance of the classifier. Accordingly, all the models we tested in Table 4 do not achieve high accuracy.

Data set	GYAFC		
	ACC.	self-BLEU	ref-BLEU
CrossAlign(Shen et al., 2017)	68.1%	3.77 ± 0.26	2.85 ± 0.20
DualRL(Luo et al., 2019)	72.6%	53.10 ± 1.86	19.27 ± 1.18
StyleTrans(Dai et al., 2019)	74.1%	65.95 ± 1.61	**22.11** ± **1.35**
DGST(Li et al., 2020)	60.5%	62.62 ± 1.21	15.72 ± 1.13
MAR (Ours)	74.6%	70.12 ± 2.12	23.25 ± 1.44

Table 4: The test results on the GYAFC (formality transfer). The confidence level of BLEU is 0.95.

In terms of human evaluation, the results are shown in Figure 3. We analyse that our proposed model shows better results in terms of accuracy and content preservation than the two similar models. In terms of fluency, our proposed model and the TAG model are evenly matched with similar proportions. As we mentioned, the relatively poor fluency of the DRG model might stem from its retrieving module. Comparing these three models, we conclude that our model has the strongest overall performance.

5.2 Case Study

To further demonstrate the superiority of our model, We randomly sampled sentences from the outputs of our model and DRG model for comparison. Table 3 shows that, for particular inputs, the retrieval-based method, i.e., DRG, does not always find a suitable counterpart. When this is the case, the output can largely differ from the original semantics of the input sentence. Redundant words are also introduced. The method based on the transformation in latent space, i.e., DGST, always copies sentences without transferring them into correct style domains.

For the transformation of negative to positive on the IMDb data set, we note that the mask for the word 'do' seems to be redundant. We analyse that the training of the classifier is influenced by the quality of the used data set. In this example, the masking module incorrectly masks a content word. It results in the low self-BLEU in Table 2.

5.3 Additional Study

Following previous work (Dai et al., 2019), we make ablation studies on the Yelp data set to confirm the validity of our model. We inspect the following three aspects:

- Is the special symbol [MASK] necessary?
- How will the results be affected in the absence of a language model in the Masker?
- What is the correlation between human and automatic evaluation?

For the first question, we removed all of the [MASK] in z_A and z_B, and we repeated the above experiments. As shown in Figure 4, the performance of our proposed model without masks shows a lower transfer accuracy and self-BLEU score. Besides, the model without masks is more unstable

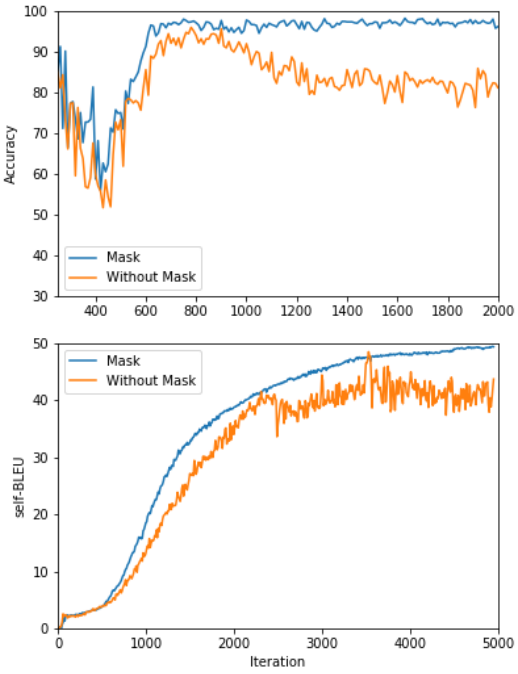

Figure 4: Accuracy and self-BLEU curves of the model during the training phase, with and without masks.

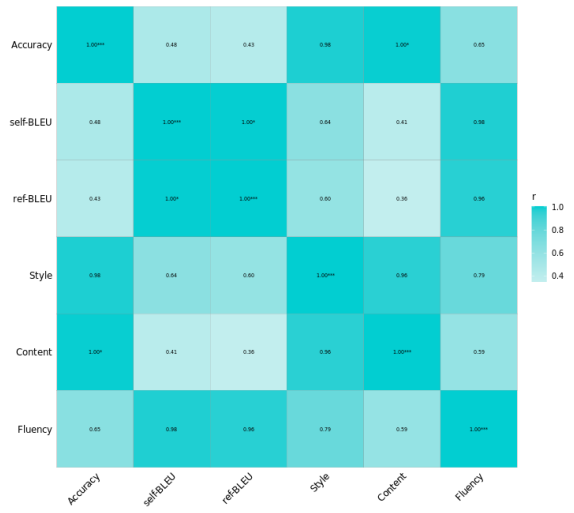

Figure 5: Pearson correlation between different evaluation metrics. Scores marked with * denotes p<0.01.

in performance in the latter stages of training. The mask operation will make the generator easily figure out the positions where the words need to be filled in. Sequences that do not include a mask require the model to make additional judgments about the position, which increases the burden of the model and is likely to lead to text degradation.

For the second question, we removed the used LM and repeated the experiments. It means that the [MASK] will not be removed and the model only learns to do substitution without any insertion or deletion. The results show that the accuracy is not affected (less than one per cent). However, the absence of the LM results in a 4 per cent reduction in BLEU scores. The absence of LM corresponds to the fact that the model cannot perform direct deletion of words. This means that all sentences need to be processed with word substitution, and during word substitution, the generator may insert multiple words for a [MASK], which may be an important cause of the drop in self-BLEU scores.

For the third question, we calculated the Pearson correlation between different evaluation metrics and the results are presented in Figure 5. Overall, positive correlations are observed between all metric combinations. It shows that both automatic evaluation and human evaluation are consistent in sentence evaluation.

Specifically, we observed that: (1) The correlation between "Accuracy" and "Style" is relatively large than the association between "Accuracy" and "Fluency". (2) The BLEU score metrics significantly correlate with the "Content" metric. (3) The "ref-BLEU" and "self-BLEU" metrics show very similar properties. It illustrates that people might have an instinct for copying content words in style transfer tasks.

6 Conclusion

We proposed a novel word substitution based approach called Mask-and-Regenerate for sentiment and style transfer. It can be regarded as a generator in a generative adversarial network to facilitate the training of a detector which can better identify fake comments on electronic commerce platforms.

Due to the lack of available parallel corpora, the original sentences were edited to delete, insert, or substitute words. We carried out a study on the neural-based style-characteristic word recognition and the automatic application of edit operations in the domain of style transfer. For sentiment and formality transfer, the results showed that our proposed model generally outperforms baselines by about 2 per cent in terms of accuracy with comparable or better BLEU scores.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, 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*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

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 *International Conference on Machine Learning*, pages 1587–1596. PMLR.

P. Diederik Kingma and Lei Jimmy Ba. 2015. Adam: A method for stochastic optimization. *international conference on learning representations*.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Empirical Methods in Natural Language Processing*.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

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*, 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)*, Online. Association for Computational Linguistics.

Yixin Liu, Graham Neubig, and John Wieting. 2021. On learning text style transfer with direct rewards. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4262–4273, Online. Association for Computational Linguistics.

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-19*, pages 5116–5122. International Joint Conferences on Artificial Intelligence Organization.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer

through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

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*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Machel Reid and Victor Zhong. 2021. Lewis: Levenshtein editing for unsupervised text style transfer. In *In Findings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

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.

Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019b. A hierarchical reinforced sequence operation method for unsupervised text style transfer. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019*, pages 4873–4883.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *CoRR*, abs/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.