Mask and Infill: Applying Masked Language Model to Sentiment Transfer

Xing Wu[1,2,3] , **Tao Zhang**[1,2] , **Liangjun Zang**[1] , **Jizhong Han**[1] and **Songlin Hu**[1*]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Baidu Inc., Beijing, China

wuxing03@baidu.com,{zhangtao,zangliangjun,hanjizhong,husonglin}@iie.ac.cn

Abstract

This paper focuses on the task of sentiment transfer on non-parallel text, which modifies sentiment attributes (e.g., positive or negative) of sentences while preserving their attribute-independent content. Due to the limited capability of RNN-based encoder-decoder structure to capture deep and long-range dependencies among words, previous works can hardly generate satisfactory sentences from scratch. When humans convert the sentiment attribute of a sentence, a simple but effective approach is to only replace the original sentimental tokens in the sentence with target sentimental expressions, instead of building a new sentence from scratch. Such a process is very similar to the task of Text Infilling or Cloze, which could be handled by a deep bidirectional Masked Language Model (e.g. BERT). So we propose a two step approach "Mask and Infill". In the *mask* step, we separate style from content by masking the positions of sentimental tokens. In the *infill* step, we retrofit MLM to Attribute Conditional MLM, to infill the masked positions by predicting words or phrases conditioned on the context[1]and target sentiment. We evaluate our model on two review datasets with quantitative, qualitative, and human evaluations. Experimental results demonstrate that our models improve state-of-the-art performance.

1 Introduction

The goal of sentiment transfer [Huang and Belongie, 2017; Logeswaran *et al.*, 2018] is to change the sentiment attribute of text while keeping its semantic content unchanged, with broad applications of review sentiment transformation, news rewriting, and so on. For example, we could convert a positive sentence "I highly recommend this movie" to a negative one "I regret watching this movie". Lacking supervised parallel data, i.e. pairs of sentences with the same content but different attributes, makes it hard to change the sentiment of

---

[*]Corresponding Author
[1]In this paper, *content* and *context* are equivalent, *style*, *attribute* and *label* are equivalent.
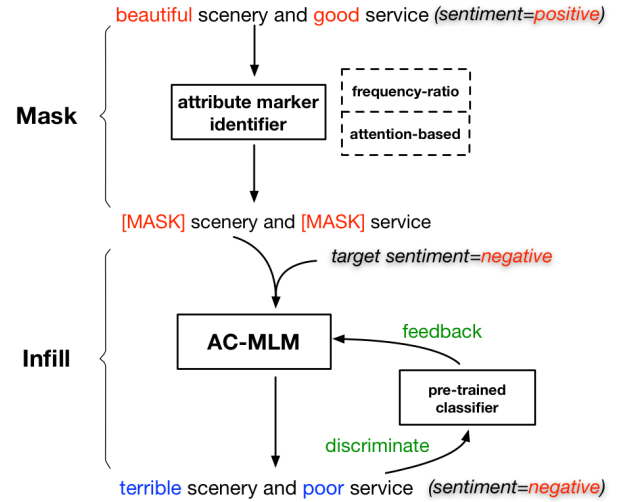

Figure 1: Process of our approach. In the mask stage, we explicitly identify and mask the sentiment tokens in a sentence. In the infill stage, we fill the masked positions with new expressions conditioned on their context and the target sentiment.

text without loss of its semantic content. Recently, several models have been proposed to learn sentiment transfer from non-parallel text, notably [Shen *et al.*, 2017; Fu *et al.*, 2018; Xu *et al.*, 2018; Li *et al.*, 2018]. Some of them [Shen *et al.*, 2017; Prabhumoye *et al.*, 2018; Fu *et al.*, 2018] try to learn the disentangled representation of content and attribute of a sentence in a hidden space, while the others [Xu *et al.*, 2018; Li *et al.*, 2018] explicitly separate style from content in feature-based ways and encode them into hidden representations respectively. Afterwards, they utilize a RNN decoder (e.g. LSTM [Hochreiter and Schmidhuber, 1997; Sutskever *et al.*, 2014; Cho *et al.*, 2014]) to generate a new sentence conditioned on the hidden representations of original content and target sentiment attribute. On the one hand, a shallow RNN encoder structure has limited ability to produce high-quality hidden representation. On the other hand, usage of RNN decoder restricts prediction ability to a short range, which often fails to produce long realistic sentences.

In this work, we leverage an important observation. When people convert a sentence with a particular sentiment into one

with a different sentiment, we do not have to create a new sentence from scratch. Instead, we simply replace the sentimental tokens of the sentence with other expressions indicative of a different sentiment. For example, given a negative review "*terrible* scenery and *poor* service", a straightforward positive review is "*beautiful* scenery and *good* service". Such a process is very similar to the task of text infilling [Zhu *et al.*, 2019] or Cloze [Taylor, 1953]. Given the template "___ scenery and ___ service", we human feel easy to fill the right words because we have enough prior linguistic knowledge to predict the missing words from their contexts.

We transform the problem of sentiment transfer to the task of text infilling by a pre-trained Masked Language Model (MLM). We utilize a pre-trained deep bidirectional language model (corresponding to human linguistic knowledge) to infill the masked positions. Our approach comprises two stages: Mask and Infill (Figure 1). In the mask stage, we explicitly identify and mask the sentiment tokens in a given sentence. In the infill stage, we fill the masked positions with new expressions conditioned on their context and the target sentiment.

There exist two feature-based methods for identifying sentiment attribute markers (i.e. words or phrases that are indicative of a particular sentiment attribute). Attention-based method [Xu *et al.*, 2018] trains a self-attention[Bahdanau *et al.*, 2015] sentiment classifier, where the learned attention weights can be used as signals to identify sentiment markers, and the tokens with weights higher than the average are regarded as sentiment attribute markers. The attention-based method prefers isolated single words to n-gram phrases, which restricts MLM to produce sentences of diverse expressions. Frequency-ratio method [Li *et al.*, 2018] constructs an attribute marker dictionary of n-gram phrases for each sentiment attribute and looks up the dictionary to identify sentiment phrases in sentences. Frequency-ratio method is inclined to mask longer phrases, and its performance heavily relies on the quality of attribute marker dictionaries. Thus, we explore a simple fused method to utilize the merits of both methods. Specifically, we combine the attention-based classifier with the frequency-ratio method, filtering out false attribute markers and discovering new single sentiment words. When the frequency-ratio method fails to identify any attribute marker or recognize too many ones (with insufficient content left), we utilize the attention-based method directly.

MLM predicts the masked tokens only conditioned on their context, considering no attribute information. To fill the masked positions in a sentence with expressions compatible to a particular sentiment, we retrofit MLM to Attribute Conditional Masked Language Model (AC-MLM) by integrating attribute embeddings with the original input for MLM. Furthermore, we introduce a pre-trained sentiment classifier to constrain AC-MLM, which ensures the generated sentences compatible with the target sentiment. To deal with the discrete nature of language generation, we utilize soft-sampling [Hu *et al.*, 2017] to guide the optimization of AC-MLM, where soft-sampling is used to back-propagate gradients through the sampling process by using an approximation of the sampled word vector.

We evaluate our models on two review datasets *Yelp* and *Amazon* by quantitative, qualitative, and human evaluations.

Experimental results show that our method achieves state-of-the-art results on accuracy, BLEU[Papineni *et al.*, 2002] and human evaluations.

Our contributions are summarized as follows:

- We propose a two-stage "Mask and Infill" approach to sentiment transfer task, capable of identifying both simple and complex sentiment markers and producing high-quality sentences.

- Experimental results show that our approach outperforms most state-of-the-art models in terms of both BLEU and accuracy scores.

- We retrofit MLM to AC-MLM for labeled sentence generation. To the best of our knowledge, it is the first work to apply a pre-trained masked language model to labeled sentence generation task.

2 Related Work

2.1 Non-parallel Style Transfer

[Shen *et al.*, 2017; Fu *et al.*, 2018; Xu *et al.*, 2018; Li *et al.*, 2018; Yang *et al.*, 2018] are most relevant to our work. [Shen *et al.*, 2017] assumed a shared latent content distribution across different text corpora, and leverages refined alignment of latent representations. [Fu *et al.*, 2018] learned a representation that only contains the content information by multi-task learning and adversarial training.[Xu *et al.*, 2018] proposed a cycled reinforcement learning method by collaboration between a neutralization module and an emotionalization module. [Li *et al.*, 2018] separated attribute and content by deleting attribute markers, and attempted to reconstruct the sentence using the content and the retrieved target attribute markers with an RNN. [Yang *et al.*, 2018] used a target domain language model as the discriminator to guide the generator to produce sentences. Unlike these models, which adopt RNN as encoder and decoder, we utilize pre-trained MLM to capture longer linguistic structure and better language representation.

2.2 Fine-tuning on Pre-trained Language Model

Language model pre-training has attracted widespread attention, and fine-tuning on pre-trained language models has proven to be effective in improving many downstream natural language processing tasks. [Dai and Le, 2015] improved Sequence Learning with recurrent networks by pre-training on unlabeled data. [Radford *et al.*, 2018] improved the performance largely on many sentence-level tasks from the GLUE benchmarks [Wang *et al.*, 2018]. [Howard and Ruder, 2018] proposed Universal Language Model Fine-tuning (ULMFiT), which was a general transfer learning method for fine-tuning a language model. [Radford *et al.*, 2018] demonstrated that by generative pre-training language model on unlabeled text from diverse corpora, large gains could be achieved on a diverse range of tasks. By introducing a deep bidirectional masked language model, BERT [Devlin *et al.*, 2018] obtained new state-of-the-art results on a broad range of tasks. Unlike previous works fine-tuning pre-trained language models to perform discriminative tasks, we aim to apply a pre-trained masked language model on generative task.
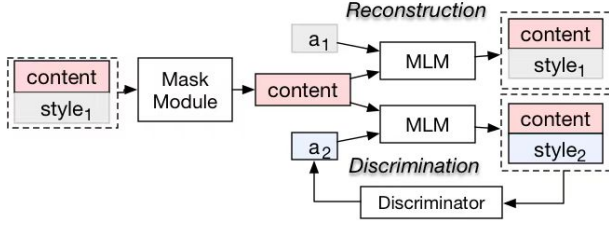
Figure 2: The overall model architecture consists of two modules. The Mask module is the fusion-method introduced in Section 3.2. The Infill module consists of two parts, reconstruction and discrimination, corresponding to Equation 10.

3 Approach

In this section, we introduce our proposed approach. An overview is presented in Section 3.1. The mask process is introduced in Section 3.2. The details of the infill process are shown in Section 3.3.

3.1 Overview

Given a corpus of labeled sentences: $\mathcal{D} = \{(S_1, a_1), ..., (S_m, a_m)\}$, where S_i is a sentence and a_i is the corresponding attribute. We define $\mathcal{A}$ as the set of all possible attributes, $\mathcal{D}_a = \{S : (S, a) \in \mathcal{D}\}$ as the set of sentences with attribute a.

Our goal is to learn a model that takes as input $(S, \hat{a})$, where S is a sentence with original sentiment attribute and $\hat{a}$ is target sentiment attribute, outputs a sentence $\hat{S}$ that retains the semantic content of S while exhibiting $\hat{a}$.

The proposed approach consists of two parts: a mask module and an infill module, as shown in Figure 2. The mask module combines the strength of two methods, utilizing a candidate phrase vocabulary to find the potential attribute markers in sentences, then performs mask operation. Then the masked sentences are sent into the infill module, which infills masked positions via AC-MLM. During the second stage, for better attribute compatibility of generated sentences, we introduce a pre-trained classifier as the discriminator. Due to the discrete nature of language generation, we draw support from soft-sampling to guide the optimization of our model. The implement details are shown in Algorithm 1.

3.2 Identify and Mask Attribute Marker

We first introduce the frequency-ratio method, then the attention-base method. At last, we propose a fusion-method by combing the strength of these two methods.

Frequency-ratio Method

We define for any a in $\mathcal{A}$, let $count(u, \mathcal{D}_a)$ denotes the number of times an n-gram u appear in $\mathcal{D}_a$, we score the salience of u with respect to a by its smoothed frequency-ratio :

$$s_c(u, a) = \frac{count(u, \mathcal{D}_a) + \lambda}{\left(\sum_{a' \in \mathcal{A}, a' \neq a} count(u, \mathcal{D}_{a'}) \right) + \lambda} \quad (1)$$

where λ is the smoothing parameter. We declare u to be a candidate attribute marker for a if $s_c(u, a)$ is above a specified threshold γ_c, these candidate attribute markers constitute a candidate attribute marker vocabulary $\mathcal{V}_a$.

Attention-based Method

Each word in a sentence contributes differently to the label or attribute of the whole sentence, we train an attention-based classifier to extract the attribute relative extent. Given a sequence S of N tokens, $< t_1, t_2, ..., t_N >$, we adopt a bidirectional LSTM to encode the sentence and concatenate the forward hidden state and the backward hidden state for each word, and obtain:

$$\mathbf{H} = (h_1, h_2, \cdots, h_N) \quad (2)$$

where N is the length of given sentence. The attention mechanism produces a vector of attention weights $\mathbf{a}$ and a vector of weighted hidden states $\mathbf{c}$. Finally a softmax layer is followed to transform $\mathbf{c}$ into probability distribution $\mathbf{y}$:

$$\mathbf{a} = softmax(\mathbf{w} \cdot tanh(\mathbf{WH}^T)) \quad (3)$$
$$\mathbf{c} = \mathbf{a} \cdot \mathbf{H} \quad (4)$$
$$\mathbf{y} = \mathbf{softmax}(\mathbf{W}' \cdot \mathbf{c}) \quad (5)$$

where $\mathbf{w}, \mathbf{W}, \mathbf{W}'$ are projection parameters. After being well-trained, the attention-based classifier can be used to extract attention weights, which capture the sentiment information of each word. For simplicity, following [Xu *et al.*, 2018], we set the averaged attention value in a sentence as the threshold, words with attention weights higher than the threshold are viewed as attribute markers.

Fusion-Method

However, there are fake attribute markers in the candidate vocabulary[2]. So we use a pre-trained classifier, i.e. the attention-based classifier, to calculate the probability p of being real for each candidate attribute marker with Equation 5 and update the salience score:

$$s(u, a) = s_c(u, a) * p \quad (6)$$

We filter out fake attribute markers with $s(u, a)$ below threshold γ, obtaining a high-quality attribute marker vocabulary $\mathcal{V}$ for attribute a.

After the vocabulary constructed, we identify and mask as many attribute markers as possible by matching the sentences, because within a sentence style words are usually less than content words. Lastly, if frequency-ratio method identifies no attribute marker or too many attribute markers (with insufficient content left, especially for short sentences), we re-process it with the attention-based method[3] by identifying the positions with high weights.

3.3 Text Infilling with MLM

Attribute Conditional Masked Language Model

Unlike traditional language models which predict tokens based on previously generated tokens, Masked Language Model (MLM) predicts the masked tokens according to their

---

[2]In the original positive vocabulary of Yelp dataset, "he did a" scores as high as 26.0, yet it is not a positive attribute marker.

[3]Note that we use the attention-based classifier twice: the first time, it is used as discriminator to improve the quality of candidate attribute **vocabulary**, the second time, it is used to extract attention weights for a **sentence**.

context, rather than reconstructing the entire input. For the task of sentiment transfer, infilled words or phrases should be consistent to a target attribute. Therefore, we propose Attribute Conditional Masked Language Model (AC-MLM) to infill masked positions conditioned on both their context and a target label, by adding attribute embeddings into original input embeddings. Let $\hat{a} \in \mathcal{A}$ be a target attribute, $S = \langle t_1, t_2, ..., t_N \rangle$ a sentence of N tokens, $M = \{t_{i_1}, ..., t_{i_m}\}$ the set of masked tokens, $\overline{S} = S \setminus M$ the set of content tokens.

Reconstruction
Lacking parallel sentiment sentence pairs, so we train the AC-MLM to reconstruct the original sentence S conditioned on the content $\overline{S}$ and original attribute a. The AC-MLM is optimized to minimize the reconstruction error of original masked words:

$$\mathcal{L}_{rec} = -\sum_{a \in \mathcal{A}, t_i \in M} log p(t_i | \overline{S}, a) \quad (7)$$

Well-trained AC-MLM model can take the sentence content $\overline{S}$ and a target attribute $\hat{a}$ as input, and output $p(\cdot | \hat{a}, \overline{S})$ for each masked position. After all masked positions are infilled, we get the transferred sentence:

$$\hat{S} = \text{AC-MLM}(\overline{S}, \hat{a}) \quad (8)$$

Specifically, we use the pre-trained BERT as our MLM, and substitute segmentation embeddings (which is useless when training single sentence) with attribute embeddings.

Pre-trained Classifier Constraint
To further improve the transfer accuracy, we introduce a pre-trained sentiment classifier to the AC-MLM, i.e. AC-MLM-SS, "-SS" indicates with soft-sampling which will be introduced later.

Discrimination Given the content $\overline{S}$ and a target attribute $\hat{a}$, when a sentence is successfully transferred, the label predicted by classifier should be consistent to the target attribute. So the AC-MLM-SS is further optimized to minimize the discrimination discrepancy:

$$\mathcal{L}_{acc} = -log p(\hat{a} | \hat{S}) \quad (9)$$

where $\hat{S}$ indicates transferred sentence generated by AC-MLM.

By combining Equation 7 and Equation 9 we obtain the objective function of AC-MLM-SS :

$$min_\theta \mathcal{L} = \mathcal{L}_{rec} + \eta \mathcal{L}_{acc} \quad (10)$$

where θ is AC-MLM-SS's parameters and η is a balancing parameter.

To propagate gradients from the discriminator through the discrete samples, we adopt soft-sampling.

Soft-sampling Soft-sampling back-propagates gradients by using an approximation of the sampled word vector. The approximation replaces the sampled token t_i (represented as a one-hot vector) at each step with the probability vector

$$\hat{t}_i \sim softmax(\mathbf{o}_t / \tau) \quad (11)$$

which is differentiable w.r.t the AC-MLM's parameters. The resulting soft generated sentence is fed into the discriminator to measure the fitness to the target attribute.

Algorithm 1 Implementation of "Mask and Infill" approach.

1: Pre-train attention-based classifier *Cls* (Eq.2-5)
2: Construct attribute marker vocabulary $\mathcal{V}$ (Eq.1,6)
3: **for** every sentence S in $\mathcal{D}$ **do**
4:     Mask attribute markers within S by looking up $\mathcal{V}$, getting $\overline{S}$
5:     **if** $\overline{S}$ is too short or $\overline{S}$ is the same as S **then**
6:         Re-mask with attention weights calculated by *Cls* (Eq.3)
7:     **end if**
8: **end for**
9: **for** each iteration i=1,2,...,M **do**
10:     Sample a masked sentence $\overline{S}$ with attribute a
11:     Reconstruct S with $\overline{S}$ and a, calculating $\mathcal{L}_{rec}$ based (Eq.7)
12:     $\hat{a}$ = the target attribute
13:     Construct $\hat{S}$ (Eq.8)
14:     calculating $\mathcal{L}_{acc}$ (Eq.9)
15:     Update model parameters θ
16: **end for**

---

4 Experiment

In this section, we evaluate our method on two review datasets Yelp and Amazon. Firstly, we introduce datasets, baseline models, training details, and evaluation metrics. Secondly, we compare our approach to several state-of-the-art systems. Finally, we present our experimental results and analyze each component in detail.

4.1 Datasets

We experiment our methods on two review datasets from [Li *et al.*, 2018]: Yelp and Amazon [He and McAuley, 2016], each of which is randomly split into training, validation and testing sets. Examples in YELP are sentences from business reviews on Yelp, and examples in AMAZON are sentences from product reviews on Amazon. Each example is labeled as having either positive or negative sentiment. The statistics of datasets are shown in Table 1.

4.2 Baselines

We compared our methods to existing relevant models: CrossAligned [Shen *et al.*, 2017], StyleEmbedding and MultiDecoder [Fu *et al.*, 2018], CycledReinforce [Xu *et al.*, 2018][4], TemplateBased, RetrievalOnly, DeleteOnly and DeleteAndRetrieval [Li *et al.*, 2018], LM+Classifer [Yang *et al.*, 2018].

4.3 Experiment Details

Mask step

For the frequency-ratio method, we consider n-gram up to 4-gram and set the smoothing parameter λ to 1. Other hyperparameters are following [Li *et al.*, 2018]. For the attention-based method, we train the attention-based classifier for 10 epochs with accuracy 97% for Yelp and 82% for Amazon. For the fusion-method, we consider a sentence with content less than 5 tokens as insufficient content, and re-process it with the attention-based method.

---

[4]https://github.com/lancopku/Unpaired-Sentiment-Translation

Dataset	Attributes	Train	Dev	Test
YELP	Positive	270K	2000	500
	Negative	180K	2000	500
AMAZON	Positive	277K	985	500
	Negative	278K	1015	500

Table 1: Dataset statistics.

Infill step

We use pre-trained BERT$_{base}$ as MLM and substitute segment embedding layer with attribute embedding layer. The input size is kept compatible with original BERT and relevant hyperparameters can be found in [Devlin *et al.*, 2018]. The pre-trained discriminator is a CNN-based classifier [Kim, 2014] with convolutional filters of size 3, 4, 5 and use Word-Piece embeddings. The hyperparameters in Equation 10 and 11 are selected by a grid-search method using the validation set. We fine-tune BERT to AC-MLM for 10 epochs, and further train 6 epochs to apply discriminator constraint.

4.4 Evaluation

Automatic Evaluation

We compute automatic evaluation metrics by employing automatic evaluation tools from [Li *et al.*, 2018] and [Yang *et al.*, 2018]. Accuracy score assesses whether outputs have the desired attribute. BLEU score is computed between the outputs and the human references. A high BLEU score primar-

[5]https://github.com/lijuncen/Sentiment-and-Style-Transfer

[6][Yang *et al.*, 2018] only evaluated their models on Yelp. For fair comparison, we only evaluate our models on Yelp too.

	YELP		AMAZON	
	ACC (%)	BLEU	ACC(%)	BLEU
CrossAligned	73.7	3.1	74.1	0.4
StyleEmbedding	8.7	11.8	43.3	10.0
MultiDecoder	47.6	7.1	68.3	5.0
CycledReinforce	85.2	9.9	77.3	0.1
TemplateBased	81.7	11.8	68.7	27.1
RetrievalOnly	95.4	0.4	70.3	0.9
DeleteOnly	85.7	7.5	45.6	24.6
DeleteAndRetrieval	88.7	8.4	48.0	22.8
w/frequency-ratio				
AC-MLM	55.0	12.7	28.7	31.0
AC-MLM-SS	90.5	11.6	75.7	26.0
w/attention-based				
AC-MLM	41.5	**15.9**	31.2	**32.1**
AC-MLM-SS	97.3	14.1	75.9	28.5
w/fusion-method				
AC-MLM	43.5	15.3	42.9	30.7
AC-MLM-SS	**97.3**	14.4	**84.5**	28.5

Table 2: Automatic evaluation performed by tools from [Li *et al.*, 2018][5]. "ACC" indicates accuracy, "BLEU" measures content similarity between the outputs and the human references. "AC-MLM", represents attribute conditional masked language model. "w/" represents "with"."-SS" represents with soft-sampling.

	ACC (%)	BLEU
Yang's results		
LM	85.4	13.4
LM+Classifier	90.0	**22.3**
w/frequency-ratio		
AC-MLM	58.0	18.7
AC-MLM-SS	93.7	17.5
w/attention-based		
AC-MLM	40.0	21.8
AC-MLM-SS	**98.5**	20.5
w/fusion-method		
AC-MLM	44.1	21.3
AC-MLM-SS	97.3	20.7

Table 3: Automatic evaluation on Yelp dataset performed by tools from [Yang *et al.*, 2018][6].

ily indicates that the system can correctly preserve content by retaining the same words from the source sentence as the reference.

Table 2 shows the performance obtained with [Li *et al.*, 2018]'s tools. Our base model AC-MLM achieves all the best BLEU, but performs poorly in accuracy. After applying the discriminator constraint to apply constraint (i.e. AC-MLM-SS), the accuracy improves significantly, though with the slight decline of BLEU. Previous models are hard to perform well on accuracy and BLEU simultaneously. Among our models, AC-MLM-SS using fusion-methods in *mask* step achieves the most satisfactory performance considering on both accuracy and BLEU. Table 3 shows the performance on [Yang *et al.*, 2018]'s tools. Compared to Yang's best model LM+Classifier, we perform better on accuracy, but slightly lower on BLEU.

Human Evaluation

We hired three annotators (not authors) to rate outputs for our models and [Li *et al.*, 2018]'s best model (DeleteAndRetrieval). We adopt three criteria range from 1 to 5 (1 is very bad and 5 is very good): grammaticality, similarity to the target attribute, and preservation of the source content. For each dataset, we randomly sample 200 examples for each target attribute. Table 5 shows the human evaluation results. Among them, our AC-MLM-SS with the fusion-method achieves best results. Pre-trained MLM infills the masked positions considering the context from both directions, being able to meet grammatical requirements better. Benefiting from explicitly separating content and style, our sentences' structures are kept and we perform better on the preservation of the source content. Then by introducing the pre-trained discriminator, MLM infills more accurate attribute words.

5 Analysis

Trade-off between Content and Attribute Our loss function in Equation 10 consists of two parts: reconstruction loss and discriminative loss. Reconstruction loss guides MLM to infill contextual-compatible words. Compared to weak label-constraint introduced by attribute embeddings, discriminative

From negative to positive (YELP)	
Source	it 's not much like an actual irish pub , which is depressing .
Human	It's like an actual irish pub .
CrossAligned	it 's not good for a clean and inviting , i textcolorbluelove food .
StyleEmbedding	it 's not much like an neat of vegetarian - but tiny crust .
MultiDecoder	it 's not much like an vegetarian bagel ... much is food .
CycledReinforce	it 's not much like an actual irish pub , excellent sweet sweet !
TemplateBased	it 's not much like an actual irish pub , which is most authentic .
RetrievalOnly	i like their food , i like their prices and i like their service .
DeleteOnly	it 's not like much an actual irish pub , which is very fun .
DeleteAndRetrieval	it 's not much like an actual irish pub , which is my favorite .
AC-MLM	it 's not much like an actual irish pub , which is quite nice .
AC-MLM-SS	it 's pretty much like an actual irish pub , which is very fantastic.

Table 4: Example outputs on YELP. "Human" line is a human annotated sentence. Original negative attribute markers are colored in red, transferred positive ones are colored in blue.

	YELP			AMAZON		
	Gra	Con	Att	Gra	Con	Att
DeleteAndRetrieval	3.4	3.5	3.6	3.5	3.2	3.3
w/frequency-ratio AC-MLM-SS	3.9	3.2	4.2	3.8	3.6	3.7
w/attention-based AC-MLM-SS	4.0	3.8	**4.4**	3.9	3.7	3.7
w/fusion-method AC-MLM-SS	**4.2**	**4.0**	**4.4**	**4.1**	**4.0**	**4.0**

Table 5: Human evaluation results on two datasets. We show average human ratings for grammaticality (Gra), content preservation (Con), target attribute match (Att).
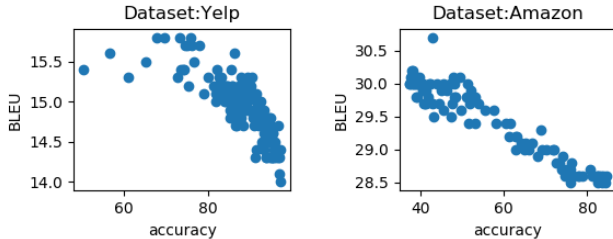

Figure 3: The trend of BLEU with the increase of accuracy.

loss from classifier improves accuracy largely by encouraging more attribute-compatible words, which decreases the diversity. The two losses are adversarial and cooperate to balance the accuracy and BLEU of transferred sentences. We plot the trade-off curve of the two indicators in Figure 3. We can modify the hyperparameter η in Equation 10 to control the trend.

Comparing Three Mask Methods In general, fusion-method performs best on automatic evaluation. On the high-quality Yelp dataset, fusion-base method gains slight improvement and performs closely to the attention-based method. But on the Amazon dataset, not as good as Yelp, the fusion-method gains large improvement by combining the merits of the attention-based and frequency-ratio methods.

Benefits of Masked Language Model Compared to the models using LSTM decoder in Table 2, our models highly improve the BLEU on both datasets. Compared to [Yang *et al.*, 2018] in table 3, which introduces the target domain language model as the discriminator, we also achieve comparable performance. But training a target domain language model from scratch is very time consuming and in need of large corpus. We directly fine-tune on the pre-trained BERT.

Connection to PG-Net Our models can be viewed as a collaboration between copying and generating, which is similar to the hybrid pointer-generator network [See *et al.*, 2017] for text summarization task. Copying aids accurate reproduction of content information, while retaining the ability to produce stylish words through the generator. Unlike PGNet which copies words from the source text via pointing, our approach copies words from the source via masking.

6 Conclusions and Future Work

In this paper, we focus on non-parallel sentiment transfer task and propose a two-stage "mask and infill" approach that enables training in the absence of parallel training data. Experimental results on two review datasets show that our approach outperforms the state-of-the-art systems, both in transfer accuracy and semantic preservation. For future work, we would like to explore a fine-grained version (more than two sentiments) of sentiment transfer, and explore how to apply the masked language model to other tasks of natural language generation beyond style transfer.

Acknowledgments

References

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 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*, 2015.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734, 2014.

[Dai and Le, 2015] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. pages 3079–3087, 2015.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Fu *et al.*, 2018] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670, 2018.

[He and McAuley, 2016] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. pages 507–517, 2016.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Howard and Ruder, 2018] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. 1:328–339, 2018.

[Hu *et al.*, 2017] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. pages 1587–1596, 2017.

[Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. pages 1501–1510, 2017.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751, 2014.

[Li *et al.*, 2018] Juncen Li, Robin Jia, He He, and Percy Liang. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1865–1874, 2018.

[Logeswaran *et al.*, 2018] Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. Content preserving text generation with attribute controls. pages 5108–5118, 2018.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. pages 311–318, 2002.

[Prabhumoye *et al.*, 2018] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 866–876, 2018.

[Radford *et al.*, 2018] Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI, 2018.

[See *et al.*, 2017] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083, 2017.

[Shen *et al.*, 2017] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. pages 6830–6841, 2017.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. pages 3104–3112, 2014.

[Taylor, 1953] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433, 1953.

[Wang *et al.*, 2018] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 353–355, 2018.

[Xu *et al.*, 2018] Jingjing Xu, Sun Xu, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. pages 979–988, 2018.

[Yang *et al.*, 2018] Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. pages 7298–7309, 2018.

[Zhu *et al.*, 2019] Wanrong Zhu, Zhiting Hu, and Eric Xing. Text infilling. *CoRR*, abs/1901.00158, 2019.