Rethinking Sentiment Style Transfer

Ping Yu[1], **Yang Zhao**[1], **Chunyuan Li**[2], **Changyou Chen**[1]
[1] State University of New York at Buffalo, Buffalo, NY 14228
[2] Microsoft Research, Redmond, WA 98052
{pingyu,yzhao63,changyou}@buffalo.edu
{chunyl}@@microsoft.com

Abstract

Though remarkable efforts have been made in non-parallel text style transfer, the evaluation system is unsatisfactory. It always evaluates over samples from only one checkpoint of the model and compares three metrics, *i.e.,* transfer accuracy, BLEU score, and PPL score. In this paper, we argue the inappropriateness of both existing evaluation metrics and evaluation methods. Specifically, for evaluation metrics, we make a detailed analysis and comparison from three aspects: style transfer, content preservation, and naturalness; for the evaluation method, we reiterate the fallacy of picking only one checkpoint for model comparison. As a result, we establish a robust evaluation method by examining the trade-off between style transfer and naturalness, and between content preservation and naturalness. Notably, we elaborate the human evaluation and identify the inaccurate measurement of content preservation automatically computed by the BLEU score. To overcome this issue, we propose a graph-based method to extract attribute content and attribute-independent content from input sentences in the YELP dataset and IMDB dataset. With the modified datasets, we design a new evaluation metric called *"attribute hit"* and propose an efficient regularization to leverage the attribute-dependent content and attribute-independent content as guiding signals. Experimental results have demonstrated the effectiveness of the proposed strategy.

1 Introduction

Text style transfer aims to modify the input attribute while retaining the attribute-independent content and contextual relations. For instance, given the input "the food in this restaurant is really delicious," an expected sentiment transfer result from positive to negative could be "the food in this restaurant is really disgusting." In this process, we expect to flip the sentiment while preserving essential contents such as "food" and "restaurant." This paper focuses on the non-parallel sentiment style transfer, where the sentences before and after transfer are not paired in the training data. Most existing works follow this setting, which is more common in real applications due to the scarcity of parallel datasets.

Most recent research efforts of text style transfer have been put on the model architecture design (Hu et al., 2017; Shen et al., 2017; Fu et al., 2018; Xu et al., 2018; Zhao et al., 2018; Luo et al., 2019; Huang et al., 2020; Li et al., 2020b; Kim and Sohn, 2020; Li et al., 2020a; Shi et al., 2021) and methodological innovations (Zhang et al., 2018; Jin et al., 2019; Liu et al., 2020b; Krishna et al., 2020; Malmi et al., 2020; Yi et al., 2020). Though achieving much progress, we identify that the evaluation system is broadly unsatisfactory. Existing evaluation systems mainly carry out *automatic evaluation* and *human evaluation*:

(i) *Automatic evaluation*: Current works mainly adopt classification accuracy, BLEU score, and PPL score for automatic evaluation. We argue that these metrics are not effective for evaluating text style transfer due to inconsistent and unfair comparisons across different works. For example, PPL is reported based on different pre-trained language models. In addition, they always pick one checkpoint for model comparison from which we usually can't reach a consensus on a proposed model's actual performance.

(ii) *Human evaluation*: A typical way is to show workers the generated sentences along with original sentences and ask them for scoring. However, we believe the task is too complicated for random workers to evaluate, and the results are too noisy to be trusted.

To alleviate these issues, we propose targeted approaches: (i) for *automatic evaluation*, we conduct a detailed analysis and comparison for the current metrics from three aspects: style transfer, content preservation, and naturalness. We re-run the current state-of-the-art models and make a fair

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1569–1582
November 7–11, 2021.

comparison under the same setting. In addition, we propose robust style transfer evaluation by drawing curves reflecting the style transfer versus naturalness trade-off, and content preservation versus naturalness trade-off. With these trade-off curves, one could have an overall comparison. For example, one can find out whether one model is consistently better than another, or whether it is better only in some aspects (**Section 3**). (ii) For *human evaluation*, in order to eliminate the bias, we randomly mixed some manually labeled sentences to test the workers. Besides, we delicately design some rules to make human evaluation more reasonable and reliable (**Section 4**).

Through human evaluation and analysis, we found that the current automatic evaluation metrics retain the problem of detecting content preservation. To detect content preservation, the ideal automatic evaluation metric needs to be able to identify style-independent contents from an input sentence. However, the BLEU score simply calculates the continuous overlap without excluding style-related words. The Earth Mover Distance (EMD) from (Mir et al., 2019) alleviates the problem through masking style-related words and then calculating the earth mover distance. But style-related words are detected through checking human-labeled lexicon as a reference, making this method hard to be extended to other datasets. Therefore, how to effectively detect style-related words is the key challenge.

Thanks to the dependency parser, we can analyze the meaning, structure, and syntactical relationships in sentences and then formulate the general grammar rules to identify style-related contents. By leveraging this method, we pre-process the YELP and IMDB dataset. Furthermore, we introduce a regularization term that encourages the matching of attribute-independent tokens while discouraging others. We demonstrate improved model performance of our method (**Section 5**). The modified datasets will be released for future research.

2 Related Works

2.1 Style Transfer

Since our goal is to systematically evaluate text style transfer in a fair way, we carefully choose three recently proposed representative approaches that are open-source as baselines: the Style Transformer (ST) (Dai et al., 2019), Deep Latent Sequence Model (DLS) (He et al., 2020) and Fine Grained Style Transfer (FGST) (Liu et al., 2020a). Many other works either do not release the source code or the published results failed to be reproduced with the provided source code. Thus they are not considered in the comparison. Specifically, Dai et al. (2019) presents a Style Transformer that combines the Transformer (Vaswani et al., 2017) with adversarial learning to realize content preservation and text style transfer. He et al. (2020) proposes a probabilistic generative formulation that unites past work on unsupervised text style transfer. Liu et al. (2020a) proposes a new framework that treats the text style transfer as the continuous latent code movement with the guidance of the classification error's gradient.

2.2 Automatic Evaluation

To our best knowledge, Mir et al. (2019) is the only evaluation paper that analyzes style transfer evaluation systems. Still, this work only considers three old models: the cross-aligned autoencoder (Shen et al., 2017), adversarially regularized autoencoder (Zhao et al., 2018), and delete-and-retrieve models (Li et al., 2018). Two metrics were proposed in this paper: the EMD score for measuring the content preservation and a naturalness classifier for measuring the naturalness.
(i) To calculate the EMD score, a style lexicon form is first manually annotated for the YELP dataset. Then, the sentences are masked with style lexicon. Finally. the EMD score between the masked generated sentences and the masked original sentences is calculated. The work heavily depends on human labeling and is not easy to extend to other datasets. In contrast, our approach approaches the problem in a much more automatic and robust way.
(ii) To calculate the naturalness, a unigram regression classifier on original sentences and transferred sentences for each transfer model is trained. Via adversarial evaluation, this naturalness classifier is expected to distinguish human-generated inputs from machine-generated outputs.

2.3 Graph-based Methods

Sentence parsing can be helpful in understanding the meaning, structure, and syntactical relationships in sentences, which is suitable for style transfer. Shi et al. (2021) performs feature extractions and style transfer at linguistic graph level by leveraging graph neural networks. However, this style transfer task is different from analysis and reasoning tasks, which does not require a complete log-

Input	The store is dump looking and management needs to change.
Ground truth	Management is top notch, the place looks great.
Sample 1	The store is good looking and management does not need to change.
Sample 2	The store looks nice and I really like the management.
Sample 3	Friendly staff, reasonably organized and knowledge employees.
Sample 4	The store is dump.
Sample 5	The store dump dump.

Samples	Style Transfer		Content Preservation					Naturalness		
	Accuracy↑	Human↑	*self*-BLEU ↑	*ref*-BLEU ↑	EMD ↓	Attribute Hit↑	Human↑	PPL↓	Classifier↑	Human↑
Sample 1	100	100	0.00	0.00	0.41	100	5	78.03	0.92	4.7
Sample 2	100	100	0.00	0.00	0.77	60	3.3	56.51	0.97	5.0
Sample 3	100	100	0.00	0.00	1.03	0	0	112.18	0.99	5.0
Sample 4	0	0	0.21	0.00	0.75	20	1.3	114.25	0.24	3.3
Sample 5	0	0	0.00	0.00	0.76	20	1.3	812.76	0.06	1.3

Table 1: Evaluation of generated samples on YELP. (*Top*) Input is a negative sentence and the task is trying to generate a positive sentence based on this input sentence. Ground truth is the positive sentence. (*Bottom*) Evaluation results of the five samples based on current evaluation metrics and our proposed *Attribute hit*.

ical structure of a sentence. Moreover, it is also time-consuming for training with the whole graphs. Instead of leveraging the complete graph by graph neural networks, we leverage the dependency parsing tree to detect attribute-dependent and attribute-independent words in the data pre-processing step. With the help of our pre-processed datasets, linguistic knowledge is no longer needed in the modeling process.

3 Revisiting Automatic Evaluation

In this section, we will examine the current automatic evaluation metrics and automatic evaluation method from the following three aspects.

1. Style transfer accuracy: What's the success rate to transform from one style to another? For example, given an input sentence with negative sentiment, how successfully can the model transfer it to positive sentiment?

2. Content preservation: Whether the generated sentences maintain the same content as the input sentences. More specifically, we need to exam whether the generated sentences preserve the attribute-independent context from original sentences.

3. Naturalness: Are the generated sentences fluent and natural? Are there any grammatical errors?

3.1 Automated Evaluation Metrics

We will analyze current automatic evaluation metrics with some generated sentences. As an example, in Table 1, the 1st and 2nd generated samples are the desired style transfer results. Although the 3rd sample is fluent and stylized by the correct sentiment, the content appears to be unrelated. Both the 4th and 5th samples fail to transfer sentiment. The 5th sentence contains grammatical errors.

Style Transfer A pre-trained style classifier is used to detect the classification accuracy of style transfer, *e.g.*, the first three generated samples in Table 1 will be classified correctly.

Content Preservation Commonly used metrics are *self*-BLEU [1] and *ref*-BLEU scores [2]. In addition, (Mir et al., 2019) proposes to calculate the EMD score between the masked generated sentences and masked input sentences. In this paper, we propose an additional metric in Section 5, *Attribute Hit*, for the same purpose.

For example, in Table 1, both the 1st and 2nd samples preserve content from the input sentence. However, compared with the 1st sample, the 2nd sample is more flexible. The content from 3rd sample is totally unrelated. And both 4th and 5th cover partial contents (only talk about "store" without mentioning "management"). Since the 3rd sample contains correct emotion and is fluent, this sample will obtain a high score in both style transfer and naturalness detection. Our content preservation detection aims to detect this unrelated generation.

In Table 1, both *self*-BLEU score and *ref*-BLEU score are zero because there are no 4-gram over-

[1]Calculated between generated sentence and input sentence.

[2]Calculated between generated sentence and ground truth sentence. Note that, only YELP dataset contains ground truth sentences as the reference.

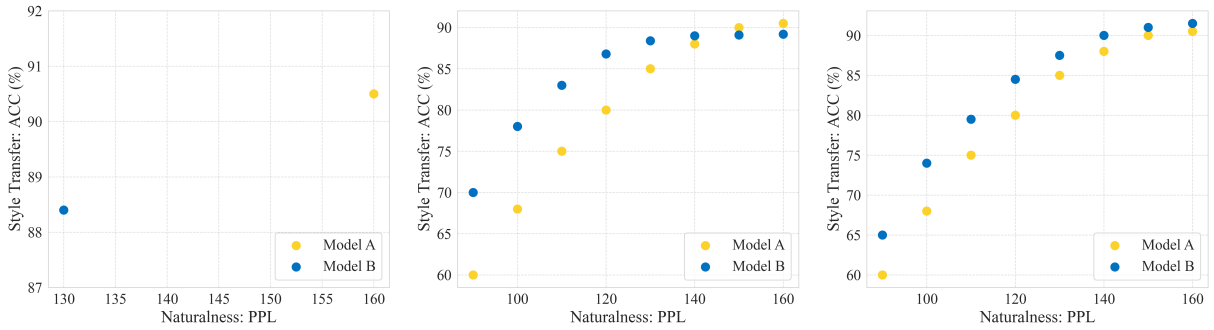


Figure 1: The importance of robust style transfer evaluation. **Left:** Current way of comparing style transfer models. The model in yellow achieves better accuracy score but worse naturalness score. In this case, it is impossible to come to any meaningful conclusion about which model (yellow or blue) dominates the other. **Middle:** The first simulated scenario (consist with the Left Figure). With this figure, the blue model should be used for better naturalness samples and yellow model should be used for high accuracy samples; **Right:** A second simulated scenario (also consist with the Left Figure) where the naturalness sweep reveals that the blue model dominates the yellow. That is, for any desired naturalness level, the blue model achieve better accuracy performance.

laps [3]. The BLEU score fails to detect unrelated generated sentences. It is not easy to have continuous overlaps between input sentences and transfer results since it needs to alter style-related words. Although YELP provides ground truths, this style transfer task is quite flexible, making it harder to calculate *ref*-BLEU. To avoid the above problem, EMD masks style-related words by checking the human-labeled lexicon and then calculates the earth mover distance between masked sentences. Our proposed Attribute Hit, by contrast, finds style-independent words by a graph-based method and then calculates whether generated sentences could hit these contents. Both EMD and Attribute Hit remove style-related words and successfully differentiate unrelated sentences (giving the lowest score to the 3rd sample in Table 1).

Naturalness PPL score from a pre-trained language model could indicate the fluency of generated sentences. (Mir et al., 2019) trains a neural logistic regression classifier to measure the naturalness. In addition, we can borrow the Grammarly software[4] for automatically scoring naturalness. Since Grammarly needs documents of at least 30 words to calculate the scores, we thus did not show the Grammarly score in Table 1. However, we will use it to calculate the generated samples in a batch[5] in the next section for measurement. In Table 1, the 5th sample contains grammar error. Both PPL and classification accuracy could give a reasonable score for measuring the naturalness in this example.

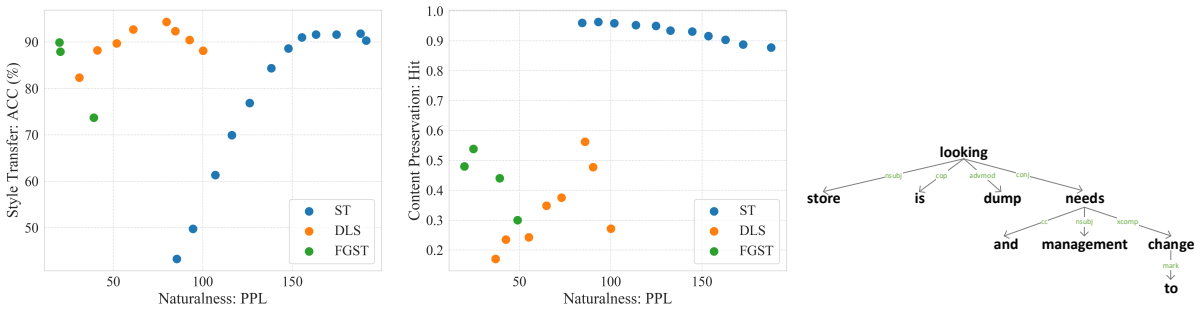
3.2 Robust Style Transfer Evaluation

The current evaluation protocol for style transfer is to pick one checkpoint for model comparison (Left figure in 1). Usually, this results in a situation where it is impossible to tell which model is superior since the actual scenario could be the Middle or Right figure. If the actual scenario is as the same trend as the Middle figure, the conclusion would be the model B should be used for generating better naturalness samples, and the model A should be used for generating high accuracy samples; However, if the actual scenario is in the similar trend as the Right figure, the conclusion would be the model B is superior to the model A.

We propose to build a robust style transfer evaluation by drawing curves of *Naturalness* versus *Style transfer* and *Naturalness* versus *Content preservation*, as demonstrated in Figure 1. During the training process, we track the naturalness value and divide naturalness into several intervals (*e.g.*, fit PPL value into 110-120, 120-130, 130-140, 140-150). In each interval, we record the best style transfer value and content preservation value. We run each method three times and report the average performance.

This new way of evaluating style transfer models allows practitioners to answer questions like: Does the new model improve others in general, or does it just improve the accuracy (successfully transfer style) at the expense of losing fluency of the generated sentences? Also, if one wants more fluent and smooth sentences rather than completing the style

[3] 4-gram BLEU scores are calculated in research papers
[4] https://app.grammarly.com/
[5] calculate 100 generated samples at once



(a) (b)

Figure 2: **(a)**: Robust style transfer evaluation among three models: the Style Transformer (ST) (Dai et al., 2019), Deep Latent Sequence Model (DLS) (He et al., 2020) and Fine Grained Style Transfer (FGST) (Liu et al., 2020a). **(b)**: Dependency parser tree generated by UDify (Kondratyuk and Straka, 2019). The original sentence is "the store is dump looking and management needs to change".

conversion, which model should be chosen?

Figure 2a shows the robust style transfer evaluation for three baseline models. In the *Naturalness-Style transfer* space, we can see that DLS could achieve a similar style transfer accuracy with higher naturalness when compared with the ST model. In the *Naturalness-Content preservation* space, the ST model achieves the highest content presentation results although sacrificing part of the naturalness. Through our robust style transfer evaluation, we could conclude that ST performs the best but scarifying part of its naturalness for the entire style transfer task. If we pay more attention to the naturalness of the generated sentences, DLS is also a good candidate.

4 Revisiting Human Evaluation

Human evaluation usually has been regarded as the ground truth of automated evaluation. However, the accuracy is affected by several factors: (1) large variance of human judgement – in the Mturk, one task will be distributed to many people, who have different scoring standards; (2) some tasks are too hard for workers to be understandable, even with examples; (3) some workers are of low quality.

We implement the following improvements. (1) To avoid the bias between different models, we associate each assignment for each worker 30 sentences, and 10 sentences per model. (2) To make rating the content preservation task more effective, we further provide some accepted good examples and rejected bad examples. We observe that this additional information brings in evident quality improvement on human evaluation. (3) To avoid the bias between different people and ensure workers to complete the work with high quality, we manually label 5 sentences and randomly mix them with 30 other sentences. Thus, each assignment contains 35 sentences, 5 of which are used to verify worker's quality. We will reject the whole assignment if the score for one of the 5 test cases different from our labels. (4) We reject all the assignments that do not match our requirements and block the workers from consistently providing low-quality submissions. The rejected assignments are re-collected until all assignments strictly match our manually labeled results. In this way, we can ensure human evaluation to be accurate. The AMT interface can be found in the supplementary material, along with more details.

Style Transfer In order to measure the success of style transfer, we instruct the workers from Mturk to rate the generated sentences with three levels: 0 (negative), 0.5 (neutral), and 1 (positive).

Content Preservation As not all raters may identity the same words as stylistic, it is impractical to ask them to ignore style-related words and rate the content preservation. To overcome this difficulty, (Mir et al., 2019) masked the style words using their style lexicon. However, their algorithm can add bias to human evaluation. Ideally, we do not wish an algorithm to affect human evaluation results. To this end, we provide raters with examples under each score (0 to 5; 0 for no relationship, and 5 for storing relationship) to educate the raters. Then we randomly set 5 test cases in each assignment to check whether the workers understand and complete the task with high quality.

Naturalness We ask whether the generated sentences are like what people say everyday, and score it from 1 to 5.

Models	Style Transfer		Content Preservation					Naturalness			
	Acc↑	Human↑	*self*-BLEU ↑	*ref*-BLEU ↑	EMD ↓	Hit↑	Human↑	PPL↓	Classifier↑	Grammarly↑	Human↑
ST	0.8760	0.7870	55.2000	20.3087	0.2113	0.8925	4.5929	112.1019	0.9102	49	2.9466
DLS	0.8830	0.8763	31.7160	12.2821	0.4688	0.6049	3.3232	42.2874	0.8141	66	3.5683
FGST	0.8699	0.8239	11.3361	5.0401	0.5364	0.5006	2.6357	21.2031	0.7655	70	3.8716

Models	Style Transfer	Content Preservation				Naturalness		
	ΔAcc	Δ*self*-BLEU	Δ*ref*-BLEU	ΔEMD	ΔHit	ΔPPL	ΔClassifier	ΔGrammarly
ST / FGST	**0.0508**	3.1268	2.6868	0.1365	**0.0402**	4.0482	0.4279	**0.0611**
DLS / FGST	**0.0485**	1.5370	1.1760	0.1348	**0.0525**	0.9161	0.1418	**0.0213**

Table 2: Evaluation of recent three models on YELP dataset. The above table is the absolute value measured by each metric; The bottom table is the amount of change compared to human evaluation of each metric, which is the smaller the better.

Models	Style Transfer		Content Preservation			Naturalness		
	Acc↑	Human↑	*self*-BLEU ↑	Hit↑	Human↑	PPL↓	Grammarly↑	Human↑
ST	0.8580	0.7979	66.1308	0.8569	4.5903	39.5525	49	2.8187
DLS	0.6679	0.7349	16.4723	0.3503	2.0471	265.66	53	2.6224

Models	Style Transfer	Content Preservation		Naturalness	
	ΔAcc	Δ*self*-BLEU	ΔHit	ΔPPL	ΔGrammarly
ST / DLS	**0.1989**	1.7722	**0.2039**	0.7763	**0.1503**

Table 3: Evaluation of recent three models on IMDB dataset.

4.1 Human Evaluation on YELP

We pick one checkpoint from each converged model for evaluation. Table 2 shows the results in terms of both automated metric and human evaluation. We also calculate the relative changed scores relative to the FGST model for a more clear comparison, which is defined as

$$\Delta \text{Acc}_{\text{ST/FGST}} = \left| \frac{\text{Acc}_{\text{ST}} - \text{Acc}_{\text{FGST}}}{\text{Acc}_{\text{FGST}}} - \frac{\text{Human}_{\text{ST}} - \text{Human}_{\text{FGST}}}{\text{Human}_{\text{FGST}}} \right|$$

We can conclude from the results that (1) for style transfer, accuracy is close to human evaluation scores. (2) for content preservation, both *self*-BLEU and *ref*-BLEU are significantly deviated from human evaluation. EMD is closer to human score, but it needs human labeled style lexicon for each dataset, which only exists for YELP dataset. Our proposed Attribute Hit is the closest to human evaluation results, and it could be easily extended to other datasets. (3) for naturalness, the pre-trained classifier is more accurate than PPL. Although Grammarly is the closest to the artificial result, it is much less flexible than the pretrained classifier as the generated sentences need to be manually copied into the software.

4.2 Human Evaluation on IMDB

Table 3 shows the results on the IMDB dataset. Because (Mir et al., 2019) only conducted experiment on the YELP dataset, implementing EMD for detecting content preservation and classifier for naturalness detection is unavailable. In addition, since the IMDB dataset does not provide the ground truth sentences, it is unable to calculate the *ref*-BLEU score. Thus, these metrics are ignored. The results on this dataset are similar to that of YELP. We observe that the classifier is great for detecting style transfer; Attribute Hit is great for content preservation; and Grammarly performs the best for measuring naturalness.

5 *Attribute Hit*

The key challenge in the task is to automatically identify style related and unrelated words. Since sentence parsing can be helpful in understanding the meaning, structure, and syntactical relationships in a sentences, we adapt it to analyze the sentence structure and detect attribute independent and dependent content.

5.1 Parsing Tree based Attribute Detection

5.1.1 Attribute-independent Content Detection

Our method is built on UDify (Kondratyuk and Straka, 2019), a single model that jointly parses Universal Dependencies (UPOS, UFeats, Lemmas, Deps). It accepts any of 75 supported languages as input (trained on UD v2.3 with 124 treebanks). Figure 2b shows an example parser tree built by UDify,

Modified Inputs	
Input	combine the bad writing and bad acting this movie just totally fail .
Attribute-independent	combine, writing, acting, movie
Attribute-dependent	bad, bad, fail
Input	this cinematic failure is littered with cheesy , cliche dialogue that 's worse than angsty teen poetry .
Attribute-independent	littered, dialogue, teen, poetry
Attribute-dependent	failure, worse
Input	after about 30 minutes i stopped the movie , went on-line to see how many minutes this disaster was .
Attribute-independent	minutes, i, movie, went, online, see, minutes, was
Attribute-dependent	stopped, disaster
Input	if you wish to have a truly traumatic experience , than this awful motion picture is for you .
Attribute-independent	you, wish, have, experience, motion, picture, you
Attribute-dependent	traumatic, awful
Input	my final comment is : do not waste your time and money to watch this uninspired and boring film .
Attribute-independent	comment, is, time, money, watch, film
Attribute-dependent	waste, boring

Table 4: Examples of our modified dataset on IMDB dataset. More examples for YELP dataset and IMDB dataset in Appendix.

which could clearly reveal structure information of each sentence.

Our method of extracting attribute-independent content is based on the intuition that attribute-independent content is usually described by nominal words or verbal words. We thus take the following steps to process the dataset:

- **Step 1:** Detect whether the *POS* [6] of each word belongs to a noun or a verb.
- **Step 2:** Use a rule based emotional classifier (Hutto and Gilbert, 2014) to detect the emotion of each verb and noun, and only keep the noun and verb with neural emotion.
- **Step 3:** Verbs can have various tenses, nouns can be in singular or plural forms, and the vocabulary of a generated sentence could be different from the original sentence (e.g., "needs" in the input sentence and "need" in the generated sentence 1 in Table 1). We thus leverage NLTK PorterStemmer class to perform stemming.
- **Step 4:** The results might end up with different pronouns, *e.g.*, the personal pronoun ("i", "you", "he", "she", *etc*), the interrogative pronoun ("which", "what", *etc*). We only consider the personal pronoun (except "it"), the possessive pronoun, and reflexive pronoun as they seem to have more impacts on the meaning of a sentence.

[6] part-of-speech tagging also called grammatical tagging

With the four steps, we can obtain the attribute-independent content for each sentence. For example, with the input "The store is dump looking and management needs to change", our attribute-independent list would be ["store", "look", "management", "need", "change"]. We use the list to calculate the *Attribute Hit* score defined as:

$$\text{Hit} = \text{Hit number}/\text{Total number of words} ,$$

where Hit number means how many words in the generated sentences are included in the attribute-independent list; the total number of words means the length of the attribute-independent list. For example, in Table 1, the 1st generated sentence contains all the words in the attribute-independent list, thus the *Attribute Hit* is 100%. The second sentence only contains "store", "look", and "management", thus the Hit is 3/5 = 60%.

This metric can also be adjusted according to our needs. For example, if we want our generated sentence more flexible, we could only use nouns. In our example, our attribute-independent list could only have ["store","look","management"]. In this case, the 2nd generated sentence in Table 1 will be selected.

5.1.2 Attribute-dependent Content Detection

We also need to extract a list of words related to the sentiment, called the attribute-dependent content. This will be used as the guide signal as the regularization in next section. We achieve this by the following steps:

- **Step 1:** Add the nouns and verbs in the sentence which has the emotional bias.
- **Step 2:** Find the modifiers (child node of nouns and verbs from Step 1).
- **Step 3:** Check whether these modifiers contain emotional bias. If yes, add them to the attribute-dependent list).

Table 4 shows some samples of our modified dataset on IMDB dataset. More examples from YELP and IMDB dataset listed in Appendix.

5.2 Regularization Term

With the attribute-independent and attribute-dependent lists for each sentence, we will leverage them to boost our training process. For each sentence, the desired transferred sentence should contain words from the attribute-independent list and avoid words from the attribute-dependent list. In other words, we want the generated sentence close to words in the attribute-independent list and far away from the words in the attribute-dependent list. To this end, we define an attribute loss:

$$\text{Loss} = \text{SIM}(E(y), E(i)) - \text{SIM}(E(y), E(d)) \ ,$$

where SIM means cosine similarity, E denotes a feature extractor, y is the generated sentence, d and i means attribute-dependent and attribute-independent words obtained from our modified datasets, respectively.

We add this attribute loss term as an extra loss term on the two best models evaluated in the previous sections: the ST and DLS model. The experiment results are shown in Figure 3. Compared with the ST model, the performance improvement is more significant for the DLS model. We argue that adding these style-related words and style-unrelated words can provide guidelines to make the model perform better.

6 Conclusion

We analyzed automatic evaluation metrics and introduced a robust style transfer evaluation method. By designing a more reliable human evaluation method, we further examined three state-of-the-art models and current evaluation metrics. As confirmed in our experiments, leveraging a classifier to evaluate style transformation is close to human

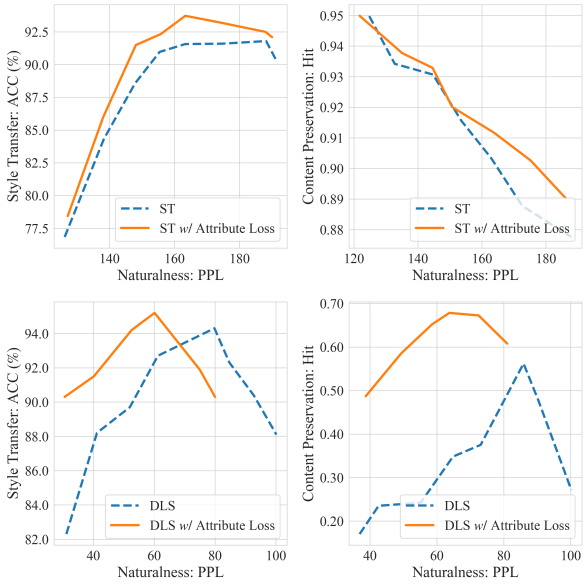


Figure 3: Add regularization term on two models: ST (above) and DLS (bottom)

evaluation. However, the current standard evaluation metric: BLEU scores are not accurate when measuring content preservation in style transfer. Similarly, PPL score also not ideal in measuring naturalness.

To overcome this issue, we propose a graph-based method to extract attribute-dependent content and attribute-independent content from input sentences in the YELP and IMDB dataset. With the modified datasets, we design a new evaluation metric called "*attribute hit*," which is a general method and could better measure content preservation. In addition, we tried to use software – Grammarly to measure the naturalness. However, borrowing the Grammarly software is not convenient since it needs manually copy the generated sentences. In addition, there are also many limitations in the software, such as not too many or too few characters in a calculation. Designing better and more general metrics that can estimate sentence fluency is also a challenge for the whole NLP community. By leveraging our modified datasets, we add the cosine similarity regularization as the guiding signal, which could further boost style transfer performance. By leveraging our published graph-based attribute extraction code, people could modify any other sentiment style transfer datasets. Also, this could help follow-up research to improve the style transfer method by leveraging style-dependent content and style-independent content.

7 Ethical Considerations

We described details about our human evaluation for a reader to understand our endeavor of providing unbiased and reliable experiments. We carried out our human evaluation on Mturk. They all voluntarily participated in our human evaluation and have been compensated fairly.

This style transfer task belongs to text generation, which could have a potential issue of generating unsafe sequences. We assessed whether those generations were safe or not using an unsafe word list and filtered out unsafe words.

References

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. *arXiv preprint arXiv:1905.05621*.

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

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. *arXiv preprint arXiv:2002.03912*.

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.

Yufang Huang, Wentao Zhu, Deyi Xiong, Yiye Zhang, Changjian Hu, and Feiyu Xu. 2020. Cycle-consistent adversarial autoencoders for unsupervised text style transfer. *arXiv preprint arXiv:2010.00735*.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. Imat: Unsupervised text attribute transfer via iterative matching and translation. *arXiv preprint arXiv:1901.11333*.

Heejin Kim and Kyung-Ah Sohn. 2020. How positive are you: Text style transfer using adaptive style embedding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2115–2125.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099*.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. *arXiv preprint arXiv:2010.05700*.

Jingjing Li, Zichao Li, Lili Mou, Xin Jiang, Michael R Lyu, and Irwin King. 2020a. Unsupervised text generation by learning from search. *arXiv preprint arXiv:2007.08557*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Xiao Li, Guanyi Chen, Chenghua Lin, and Ruizhe Li. 2020b. Dgst: a dual-generator network for text style transfer. *arXiv preprint arXiv:2010.14557*.

Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. 2020a. Revision in continuous space: Unsupervised text style transfer without adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8376–8383.

Yixin Liu, Graham Neubig, and John Wieting. 2020b. On learning text style transfer with direct rewards. *arXiv preprint arXiv:2010.12771*.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. *arXiv preprint arXiv:1905.10060*.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. *arXiv preprint arXiv:2010.01054*.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. *arXiv preprint arXiv:1904.02295*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *arXiv preprint arXiv:1705.09655*.

Yukai Shi, Sen Zhang, Chenxing Zhou, Xiaodan Liang, Xiaojun Yang, and Liang Lin. 2021. Gtae: Graph-transformer based auto-encoders for linguistic-constrained text style transfer. *arXiv preprint arXiv:2102.00769*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Jingjing Xu, Xu Sun, Qi Zeng, Xuancheng Ren, Xiaodong Zhang, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. *arXiv preprint arXiv:1805.05181*.

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2020. Text style transfer via learning style instance supported latent space. IJCAI.

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*.

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *International conference on machine learning*, pages 5902–5911. PMLR.

A Appendices

Modified Negative Inputs	
Input	definitely disappointed that i could not use my birthday gift !
Attribute-independent	i, use, birthday, gift
Attribute-dependent	disappoint
Input	don't waste your time or money at this denny's .
Attribute-independent	time, money, denny
Attribute-dependent	waste
Input	my biggest complaint, however, is what happened with our meals .
Attribute-independent	happened, meals
Attribute-dependent	complaint
Input	unfortunately my family decided to go here again tonight .
Attribute-independent	family, decided, go, tonight
Attribute-dependent	unfortunately
Input	bad food , slow service and rude managers .
Attribute-independent	food, service, manager
Attribute-dependent	bad, rude

Modified Positive Inputs	
Input	they also have daily specials and ice cream which is really good .
Attribute-independent	they, have, daily, specials, ice, cream
Attribute-dependent	good
Input	the best fish and chips you 'll ever enjoy and equally superb fried shrimp .
Attribute-independent	fish, chips, you, shrimp
Attribute-dependent	best, enjoy, superb
Input	excellent fish sandwich , wonderful reuben sandwich , even the stuffed cabbage tastes homemade .
Attribute-independent	fish, sandwich, reuben, sandwich, cabbage, tastes
Attribute-dependent	excellent, wonderful
Input	fantastic wings that are crispy and delicious , wing night on tuesday and thursday !
Attribute-independent	wings, wing, night, tuesday, thursday
Attribute-dependent	fantastic, delicious
Input	friendly staff , good food , great beer selection , and relaxing atmosphere .
Attribute-independent	staff, food, beer, selection, atmosphere
Attribute-dependent	friendly, good, great, relaxing

Table 5: Examples of our modified dataset on YELP.

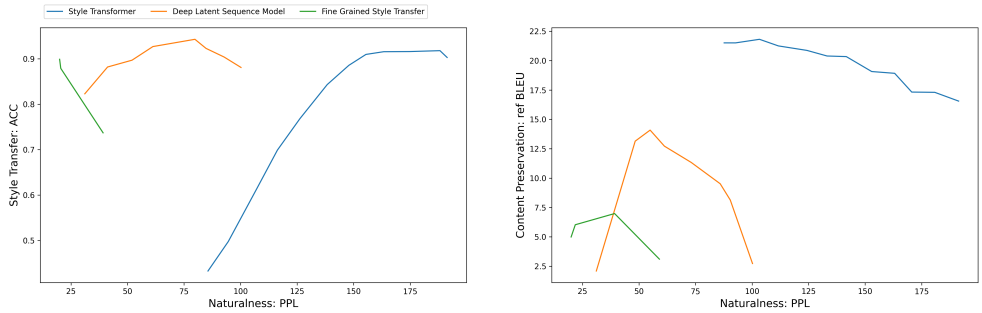


Figure 4: Model Comparison

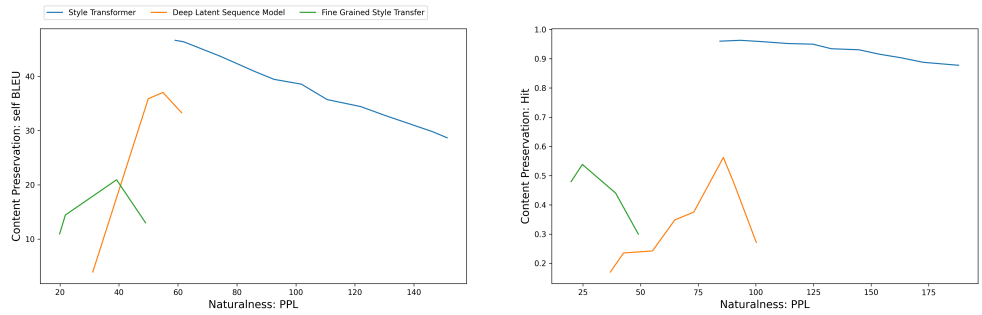


Figure 5: Model Comparison

Modified Negative Inputs	
Input	combine the bad writing and bad acting this movie just totally fail .
Attribute-independent	combine, writing, acting, movie
Attribute-dependent	bad, bad, fail
Input	this cinematic failure is littered with cheesy , cliche dialogue that 's worse than angsty teen poetry .
Attribute-independent	littered, dialogue, teen, poetry
Attribute-dependent	failure, worse
Input	after about 30 minutes i stopped the movie , went on-line to see how many minutes this disaster was .
Attribute-independent	minutes, i, movie, went, online, see, minutes, was
Attribute-dependent	stopped, disaster
Input	if you wish to have a truly traumatic experience , than this awful motion picture is for you .
Attribute-independent	you, wish, have, experience, motion, picture, you
Attribute-dependent	traumatic, awful
Input	my final comment is : do not waste your time and money to watch this uninspired and boring film .
Attribute-independent	comment, is, time, money, watch, film
Attribute-dependent	waste, boring
Modified Positive Inputs	
Input	i am a great fan of this movie and would , and have , recommended it to all .
Attribute-independent	i, movie, have
Attribute-dependent	great, fan, recommended
Input	fantastic chaplin movie with many memorable moments as charlie joins the army to fight in ww 1 .
Attribute-independent	chaplin, movie, moments, charlie, joins, army, fight, ww
Attribute-dependent	fantastic
Input	it 's one of the all-around funniest movies i 've ever seen .
Attribute-independent	movies, i, seen
Attribute-dependent	funniest
Input	powerful , to the point , beautifully acted , mysterious in it 's ending , and just downright superb .
Attribute-independent	point, acted, ending
Attribute-dependent	powerful, beautifully, superb
Input	his happy-go-lucky exterior is there , but he reveals his soul to show us the underlying loneliness .
Attribute-independent	go, exterior, he, reveals, soul, show, loneliness
Attribute-dependent	happy, lucky

Table 6: Examples of our modified dataset on IMDB.

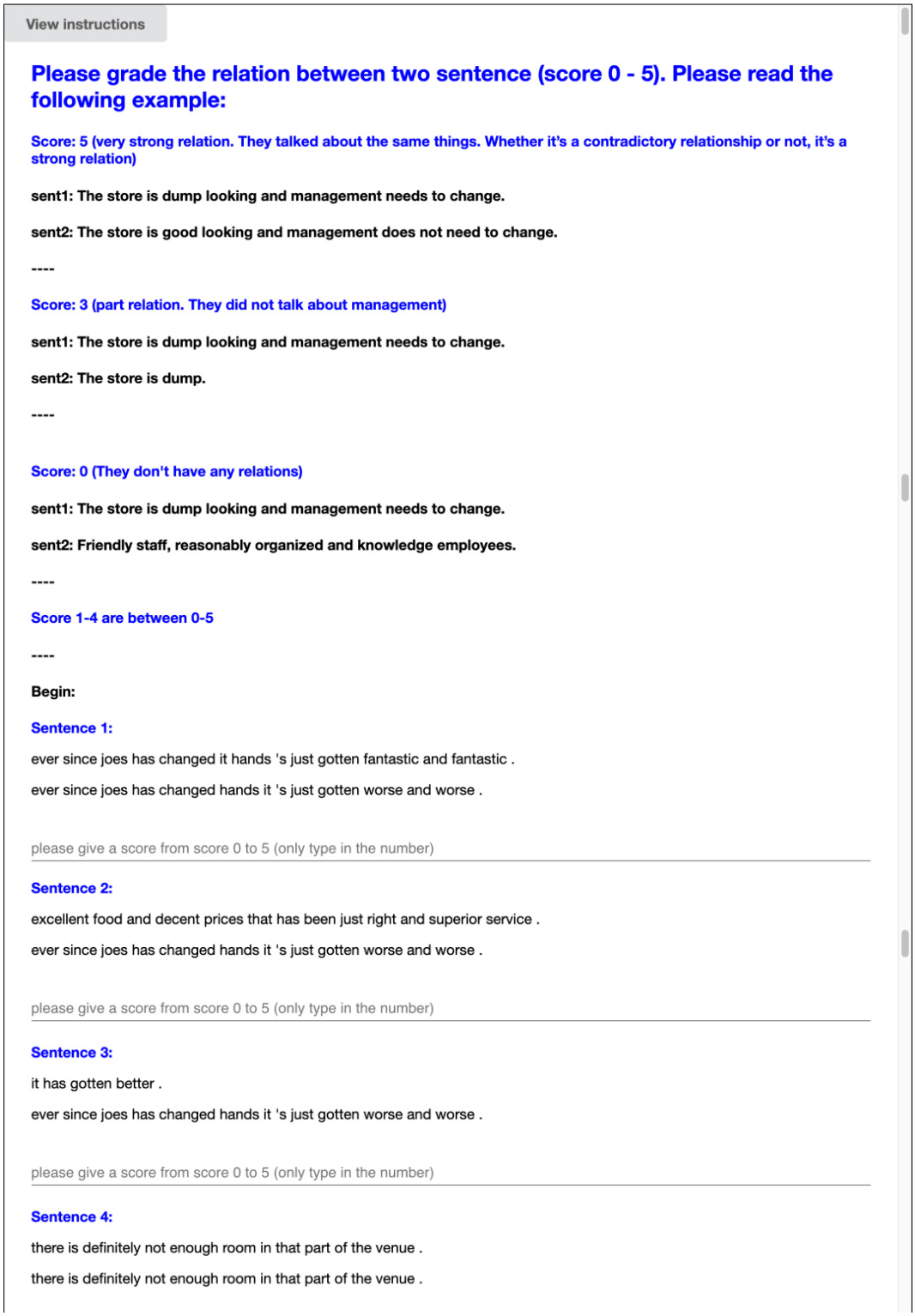


Figure 6: Mturk Evaluation