Toward Disturbance Resistant Text Style Transfer

Jingyuan Li, Jiecheng Xu, Chuhan Feng

December 11, 2020

1 Introduction

Natural Language Processing (NLP) methods have achieved great success in recent days Devlin et al. (2018); Liu et al. (2019). In this project, we focus on the unsupervised sequence-to-sequence generation task Lewis et al. (2020); Li et al. (2020), named **Text Style Transfer** Shen et al. (2017), which aims to change the style related attributes of sentences without changing the objective contents. For example, when describing the same restaurant, different people might have different opinions about the foods provided, and with high probability, they express their opinions in different ways. Then the text style transfer methods allow us to explore the differences between the way they express their opinions. The advancements in text style transfer will consequently lead to advancements in the fields of machine translation, text editing, content correction, decipherment, as well as text style control.

Supported by the latest deep learning techniques, models for text style transfer can already provide promising results in in-lab experiments. For example, Yang et al. (2018) successfully reached 90% accuracy in sentiment modification tasks, indicating the mostly 'perfect' performance of the models under specific setting. However, when coming into the practical applications, the outputs of the models are still far from satisfying, regardless of the seemingly fruitful results demonstrated in academic papers. Motivated by this observation, in this project, we would like to investigate the sources of the vulnerability of the text style transfer models in actual application, and develop potential solutions to enhance the robustness of the models.

The performance differences between models from the laboratory and those deployed in practical application are caused by two reasons, including 1) the internal assumptions made when people are building the models and 2) the external causes from discrepancies between the user's behaviors and the system's expectations. In this project, we focus on the external vulnerability of the models, which can not be solved by simply patching the models. To be more specific, the entrance and exit of the models have predefined specifications that must be satisfied by the inputs, e.g. what kind of vocabulary should not exist. In a word, these models require high stability inputs. Once these specifications are violated, the black-box system could go wrong in unpredictable manners, and then lead to serious consequence that might harm the whole application. To relief the sensitivity, one straightforward way is to apply data pre-processing, like tokenizing, converting into lower cases, etc. Unfortunately, the overly random user inputs can be a lot more difficult than the cases that can be handled by the pre-processing methods, which makes it hard to defense predictively. As a result, the robustness of the text style transfer methods is still to be improved.

In this project, we make a step forward in the robustness of the text style transfer system. In specific, we propose an input-output normalization technique to make the user inputs deterministically satisfy the specification requirements, regardless of the randomness of the user inputs. To end this, we make use of the state-of-the-art language correction methods Omelianchuk et al. (2020) to perform the role of language normalizer, which ensures that the user inputs are converted into the form that is recognizable by the transfer model. We take advantage of the language correction methods' feature that they are not sensitive to the input of users and that their outputs are also following specific specifications. This makes these models perfect candidates to solve the instability of inputs. As a result, the robustness of the system is enhanced considerably by adding the normalization technique. Besides, because of the nature of text style transfer, the output results might appear to be counter-intuitive and hard to understand from the view of users. As a result, to further improve the quality of text style transfer methods, we also perform the normalization in the outputs of the text style transfer methods. Compared to state-of-the-art methods, a clear improvement in terms of output quality and readiness is achieved by our methods. Besides, our method is robust to the disturbance in the user inputs, where the models can successfully process sentences with unpredictable errors.

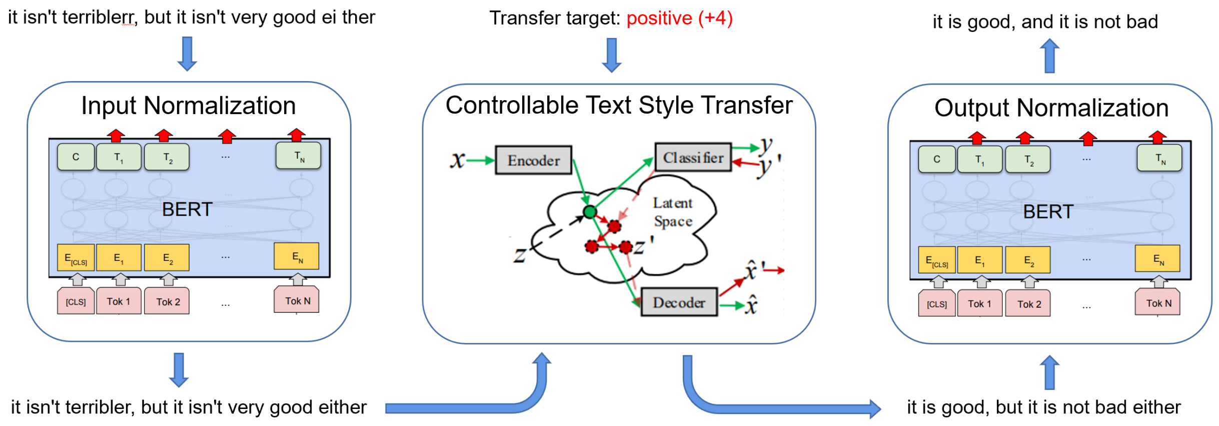

it isn't terriblerr, but it isn't very good ei ther
Transfer target: positive (+4)
it is good, and it is not bad
Input Normalization
C
T1
T2
...
TN
BERT
E[CLS]
E1
E2
EN
[CLS]
Tok 1
Tok 2
Tok N
Controllable Text Style Transfer
x
Encoder
Classifier
y
y'
Latent Space
z
z'
Decoder
x̂'
x̂
Output Normalization
it isn't terribler, but it isn't very good either
it is good, but it is not bad either

Figure 1: Model Pipeline

2 Related Works

Deep Learning based Text Style Transfer

Our methods are related to sequence-to-sequence generation with unpaired data, i.e. we don't have a corresponding 'label' sentence during training. Shen et al. (2017) developed an encoder-decoder pair to perform text style transfer, assisted by the reconstruction loss and KL loss. Xu et al. (2018); Gong et al. (2019); Luo et al. (2019b,a); Wu et al. (2019) make use of cycled reinforcement learning to enhance the sentiment-to-sentiment transfer models. Li et al. (2018) developed a deep learning model to solve the problem in a delete-and-refill fashion. Yang et al. (2018) use the language models as discriminators to build a GAN for sequence generation. John et al. (2019) propose to explicitly model the style and content space of sentence for style transfer, allowing the control of style. Dai et al. (2019) propose to control the style of text in latent space without separating the style and contents. Wang et al. (2019) further allow the detailed editing of style in the latent space. While the methods have made tremendous improvement on quality of the text style transfer results, they suffer from the external vulnerabilities of users. Besides, their outputs are usually optimized for the training targets, which might not be optimal for readability and correctness. In comparison, our model solves both of the issues, making our methods more usable in practice.

Language Correction Models

Recently, automatic language correction has been explored much. However, these language correction models are generally used only for correction during the text editing process Omelianchuk et al. (2020). We noticed that for these models, they naturally have the feature that they can accept inputs of many different forms, and they are robust to disturbance. In our work, we make special use of these models and design them as the language normalizer, which allows the input to the transfer model to be very stable. Besides, the model also works on the output of the transfer results, making them more readable.

3 Method

Our main contribution for this project is to improve text style transfer by input and output normalization. Our results show that the transformed sentences can be substantially better in both correctness and readability comparing to vanilla text style transfer model.

Pipeline

As shown in Figure 1, our entire pipeline consists of three parts, Input Normalization, Controllable Text Style Transfer (CTST), and Output Normalization. The input sentence is first corrected by Input Normalization module in both spelling and grammar. Then, the normalized input is sent to the CTST module. By providing a sentiment target to the CTST module, it performs sentiment transfer on the normalized input sentence. In the end, the generated output will be sent to the Output Normalization module to perform the final refinement.

Controllable Text Style Transfer

The backbone of our pipeline is the CTST model proposed by Wang et al. (2019), which is essentially a Transformer-NMT model with a latent space classifier as shown in Figure 2. The model is trained on yelp

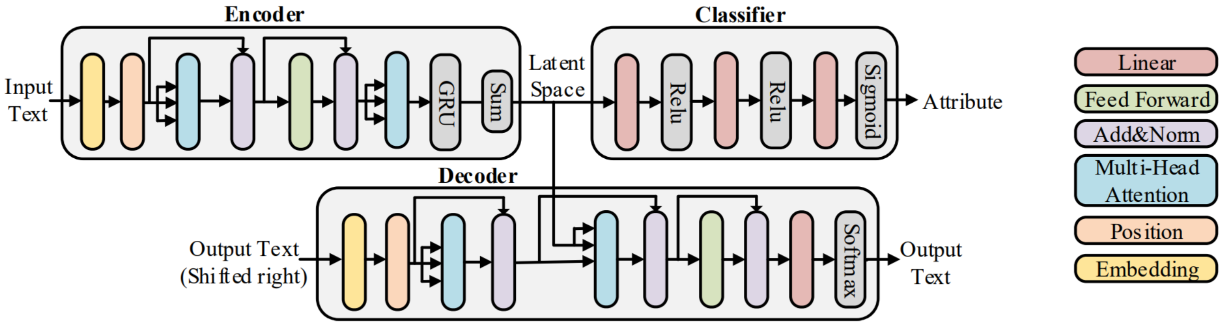


Figure 2: CTST Architecture

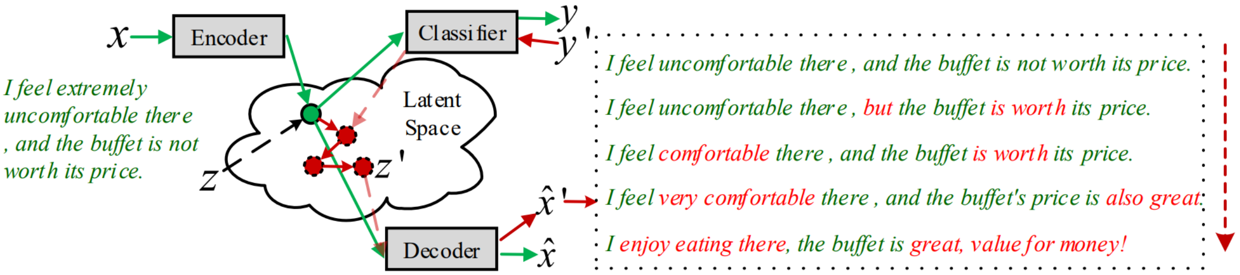

x
Encoder
Classifier
y
y'
I feel extremely uncomfortable there , and the buffet is not worth its price.
Latent Space
z
z'
Decoder
$\hat{x}'$
$\hat{x}$
I feel uncomfortable there , and the buffet is not worth its price.
I feel uncomfortable there , but the buffet is worth its price.
I feel comfortable there , and the buffet is worth its price.
I feel very comfortable there , and the buffet's price is also great.
I enjoy eating there, the buffet is great, value for money!

Figure 3: Controllable Text Style Transfer

review dataset. The CTST model takes as input an English sentence, which is then encoded into latent space representation by the transformer encoder. The latent space classifier then classifies the representation into a sentiment score. To modify the sentiment of the sentence, the representation is optimized to minimize the difference between the original sentiment score and target sentiment score. The updated representation is then decoded by the transformer decoder to generate the target sentence with the expected sentiment.

Input and Output Normalization

In order to ensure the input language distribution lies under the distribution of the training dataset, we employ a spelling and grammar correction model GECToR, proposed by Omelianchuk et al. (2020), to normalize the input sentence. GECToR is a BERT-like model trained on multiple grammar correction dataset. Unlike previously existing grammar correction methods based on seq2seq models, GECToR doesn't directly generate the target sentence. Instead, it predicts correction tags for each tokens in the input sentence. In this way, GECToR simplifies grammar correction into a sequence encoding problem and achieved state-of-the-art performance on grammar correction tasks.

In addition to that, we also applied GECToR to normalize the output from the CTST to refine the quality of the generated sentences.

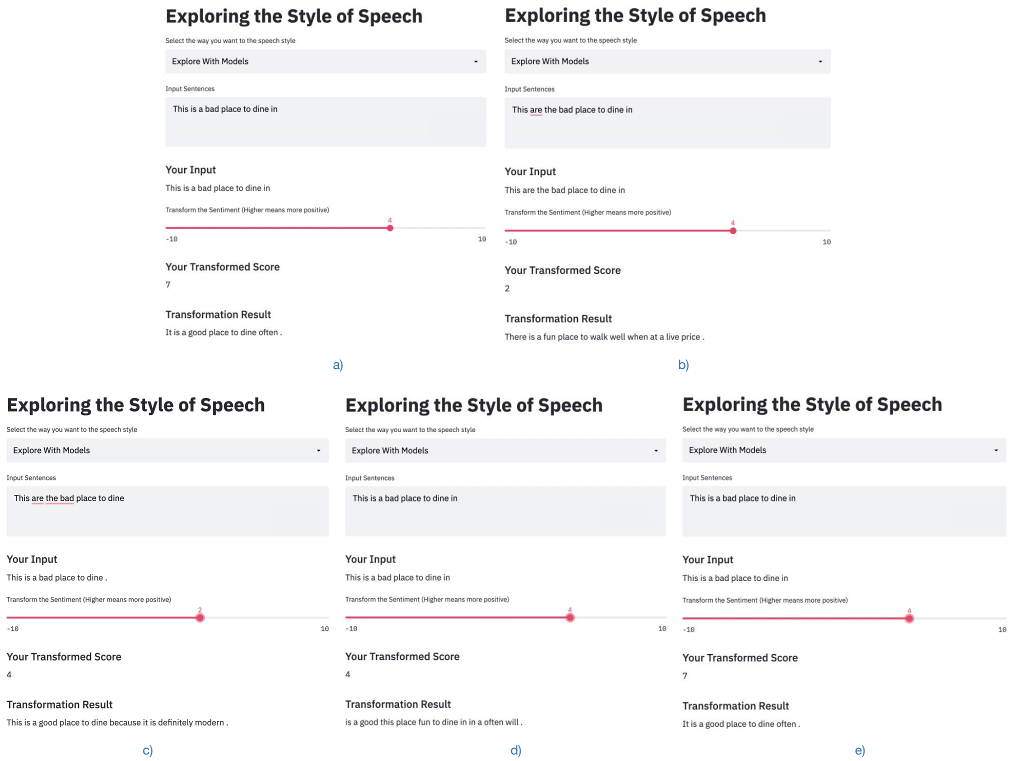
4 Results and Discussion

Datasets

In this project, the models are trained on several datasets. The text style transfer model is trained on the Yelp review dataset containing reviews for different restaurants. Although this dataset is relatively simple, it's a good for tesing the performance of text style transfer as the user ratings themselves are well aligned with the definition of 'style' in the tasks. For the input and output normalization model, a large scale dataset PIE-synthetic Awasthi et al. (2019) is used for training, and the model is fine-tuned on several smaller datasets Dahlmeier et al. (2013); Tajiri et al. (2012); Yannakoudakis et al. (2011); Bryant et al. (2019). Hopefully, this dataset combination would effectively enhance the stability of the model.

Results

Before the GECToR model was introduced to our application, our result was solely based on the performance of the sentiment transfer model, which is highly affected by the input quality. Such a bare model



Exploring the Style of Speech

Select the way you want to the speech style: Explore With Models

Input Sentences: This is a bad place to dine in

Your Input

This is a bad place to dine in

Transform the Sentiment (Higher means more positive): 4 (-10 – 10)

Your Transformed Score

7

Transformation Result

It is a good place to dine often .

a)

Exploring the Style of Speech

Select the way you want to the speech style: Explore With Models

Input Sentences: This are the bad place to dine in

Your Input

This are the bad place to dine in

Transform the Sentiment (Higher means more positive): 4 (-10 – 10)

Your Transformed Score

2

Transformation Result

There is a fun place to walk well when at a live price .

b)

Exploring the Style of Speech

Select the way you want to the speech style: Explore With Models

Input Sentences: This are the bad place to dine

Your Input

This is a bad place to dine .

Transform the Sentiment (Higher means more positive): 2 (-10 – 10)

Your Transformed Score

4

Transformation Result

This is a good place to dine because it is definitely modern .

c)

Exploring the Style of Speech

Select the way you want to the speech style: Explore With Models

Input Sentences: This is a bad place to dine in

Your Input

This is a bad place to dine in

Transform the Sentiment (Higher means more positive): 4 (-10 – 10)

Your Transformed Score

4

Transformation Result

is a good this place fun to dine in in a often will .

d)

Exploring the Style of Speech

Select the way you want to the speech style: Explore With Models

Input Sentences: This is a bad place to dine in

Your Input

This is a bad place to dine in

Transform the Sentiment (Higher means more positive): 4 (-10 – 10)

Your Transformed Score

7

Transformation Result

It is a good place to dine often .

e)

Figure 4: The output results of our methods are the compared method.

will be vulnerable to any changes that make the content not recognizable by the model. For instance, the sentiment transfer model works well on a well-structured, grammatical correct input sentence (Figure 4a), but got confused if user inputs contain grammatical mistakes (Figure 4b). From the view of the model, these mistakes have yielded 'new words' that haven't shown up in the training dataset. As a consequence, their performance becomes unpredictable.

To overcome this issue, we decided to introduce GECToR to perform language normalization on both our input and output. This way, we can ensure that the inputs to the model are correct and safe. As demonstrated by the example in (Figure 4c), by stabilizing our inputs, we were able to make a significant improvement regarding the robustness of the model. Even if the input to the model is completely messed up, the model produces very encouraging results.

We have also discovered that, sometimes the output of the style transfer model doesn't make sense and is not grammatically correct, which harms of overall user experience (Figure 4d). It's due to that when training the text style transfer model, the optimization target is usually inconsistent with the human experience. Thus the results can be difficult for people to understand. To address this issue, we also perform normalization on the outputs of the model to further stabilize the results and gain extra readability. (Figure 4f)

Further Improvements

Currently the sentiment transfer model output is not very stable since the model was trained only on the yelp review datasets. In terms of future improvement, we are planning to expand the training dataset beyond yelp reviews so that the sentiment transfer model would better generalize to a different context. Also, we are planning on expanding the vocabulary set used by the GECToR model so it could perform a better job with grammatical correction.

5 Conclusion

In this project, we explore the task text style transfer. We noticed the vulnerability of the text style transfer methods and showed that these vulnerabilities might make the model unusable in practice. To relieve the issue, we propose an input-output normalization scheme that ensures the specification inputs and outputs of the model. With our proposed technique, improvements in terms of readability and fluency as well as robustness are achieved.

References

Awasthi, A., Sarawagi, S., Goyal, R., Ghosh, S., and Piratla, V. (2019). Parallel iterative edit models for local sequence transduction. In *Proc. EMNLP*, pages 4259–4269.

Bryant, C., Felice, M., Andersen, Ø. E., and Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. In *Proc. ACL*, pages 52–75.

Dahlmeier, D., Ng, H. T., and Wu, S. M. (2013). Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proc. ACL Workshop*, pages 22–31.

Dai, N., Liang, J., Qiu, X., and Huang, X. (2019). Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proc. ACL*, pages 5997–6007.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gong, H., Bhat, S., Wu, L., Xiong, J., and Hwu, W.-m. (2019). Reinforcement learning based text style transfer without parallel training corpus. In *Proc. NAACL*, pages 3168–3180.

John, V., Mou, L., Bahuleyan, H., and Vechtomova, O. (2019). Disentangled representation learning for non-parallel text style transfer. In *Proc. ACL*, pages 424–434.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. ACL*, pages 7871–7880.

Li, J., Jia, R., He, H., and Liang, P. (2018). Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proc. ACL*.

Li, K., Chen, C., Quan, X., Ling, Q., and Song, Y. (2020). Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proc. ACL*, pages 7056–7066.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luo, F., Li, P., Yang, P., Zhou, J., Tan, Y., Chang, B., Sui, Z., and Sun, X. (2019a). Towards fine-grained text sentiment transfer. In *Proc. ACL*, pages 2013–2022.

Luo, F., Li, P., Zhou, J., Yang, P., Chang, B., Sun, X., and Sui, Z. (2019b). A dual reinforcement learning framework for unsupervised text style transfer. In *Proc. IJCAI*, pages 5116–5122.

Omelianchuk, K., Atrasevych, V., Chernodub, A., and Skurzhanskyi, O. (2020). GECToR – grammatical error correction: Tag, not rewrite. In *Proc. ACL Workshop*, pages 163–170.

Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Proc. NIPS*, pages 6830–6841.

Tajiri, T., Komachi, M., and Matsumoto, Y. (2012). Tense and aspect error correction for ESL learners using global context. In *Proc. ACL*, pages 198–202.

Wang, K., Hua, H., and Wan, X. (2019). Controllable unsupervised text attribute transfer via editing entangled latent representation. In *Proc. NeurIPS*, pages 11036–11046.

Wu, C., Ren, X., Luo, F., and Sun, X. (2019). A hierarchical reinforced sequence operation method for unsupervised text style transfer. In *Proc. ACL*.

Xu, J., Sun, X., Zeng, Q., Zhang, X., Ren, X., Wang, H., and Li, W. (2018). Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proc. ACL*.

Yang, Z., Hu, Z., Dyer, C., Xing, E. P., and Berg-Kirkpatrick, T. (2018). Unsupervised text style transfer using language models as discriminators. In *Proc. NeurIPS*, pages 7287–7298.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *Proc. ACL*, pages 180–189.