Enhancing Content Preservation in Text Style Transfer Using Reverse Attention and Conditional Layer Normalization

Dongkyu Lee **Zhiliang Tian** **Lanqing Xue** **Nevin L. Zhang**
Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology
{dleear, ztianac, lxueaa, lzhang}@cse.ust.hk

Abstract

Text style transfer aims to alter the style (e.g., sentiment) of a sentence while preserving its content. A common approach is to map a given sentence to content representation that is free of style, and the content representation is fed to a decoder with a target style. Previous methods in filtering style completely remove tokens with style at the token level, which incurs the loss of content information. In this paper, we propose to enhance content preservation by implicitly removing the style information of each token with reverse attention, and thereby retain the content. Furthermore, we fuse content information when building the target style representation, making it dynamic with respect to the content. Our method creates not only style-independent content representation, but also content-dependent style representation in transferring style. Empirical results show that our method outperforms the state-of-the-art baselines by a large margin in terms of content preservation. In addition, it is also competitive in terms of style transfer accuracy and fluency.

1 Introduction

Style transfer is a popular task in computer vision and natural language processing. It aims to convert an input with a certain style (e.g., sentiment, formality) into a different style while preserving the original content.

One mainstream approach is to separate style from content, and to generate a transferred sentence conditioned on the content information and a target style. Recently, several models (Li et al., 2018; Xu et al., 2018; Wu et al., 2019) have proposed removing style information at the token level by filtering out tokens with style information, which are identified using either attention-based methods (Bahdanau et al., 2015) or frequency-ratio based methods (Wu et al., 2019). This line of work is built upon the assumption that style is *localized* to certain tokens in a sentence, and a token has *either* content or style information, but *not both*. Thus by utilizing a style marking module, the models filter out the style tokens entirely when constructing a style-independent content representation of the input sentence. The drawback with the filtering method is that one needs to manually set a threshold to decide whether a token is stylistic or content-related. Previous studies address this issue by using the average attention score as a threshold (Li et al., 2018; Xu et al., 2018; Wu et al., 2019). A major shortcoming of this approach is the incapability of handling flat attention distribution. When the distribution is flat, in which similar attention scores are assigned to tokens, the style marking module would remove/mask out more tokens than necessary. This incurs information loss in content as depicted in Figure 1.

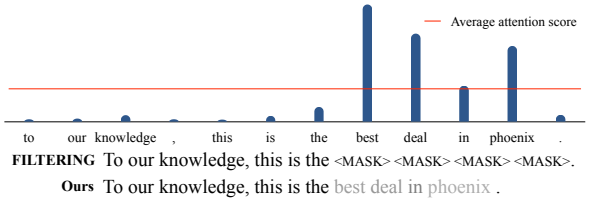

Figure 1: Illustration of difference between our method and filtering method in handling flat attention distribution. Each bar indicates attention score of the corresponding word.

In this paper, we propose a novel method for text style transfer. A key idea is to exploit the fact that a token often posses both style and content information. For example, the word "delicious" is a token with strong style information, but it also implies the subject is food. Such words play a pivotal role in representing style (e.g., positive sentiment) as well as presenting a hint at the subject matter/content (e.g., food). The complete removal of such tokens leads to the loss of content information.

For the sake of enhancing content preservation,

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 93–102
August 1–6, 2021.