The Principles and Functions of Various Models for Text Style Transformation

Haolun He

School of computing, Bejing University of Technology, Bejing, China
hehaolun@emails.bjut.edu.cn

Abstract. Text style transfer is a significant research task in natural language processing, with its core objective being to transform the style of a text while maintaining the original semantic content. This paper reviews and summarizes existing research from the perspectives of supervised learning, semi-supervised learning, and unsupervised learning. Firstly, it introduces the supervised methods based on parallel corpora and their advantages and disadvantages. Secondly, it analyzes the semi-supervised methods that combine limited parallel data with large-scale non-parallel data. Finally, it focuses on the unsupervised methods that rely solely on non-parallel corpora and their latest progress. This paper also summarizes the commonly used datasets and evaluation metrics, points out the shortcomings of existing automated evaluations, and discusses the dual impact of text style transfer on information dissemination, human-computer interaction, and social applications. In conclusion, the development of text style transfer not only depends on algorithm innovation but also requires consideration of data construction, evaluation systems, and ethical norms, which will lay an important foundation for future research and applications.

Keywords: Text Style Transfer, Supervised Learning, Semi-supervised Learning, Unsupervised Learning.

1 Introduction

In the era of rapid development of artificial intelligence, text is not only a carrier of information and knowledge, but also a reflection of human emotions, personalities and cultures. In various application scenarios, people often hope to adjust the expression style of the text while keeping the meaning unchanged. For instance, in social media operation, the formality of the language is modified according to the audience; in cross-cultural communication, localized expressions are transformed into more universal ones; in education and accessible reading, complex academic texts are simplified to facilitate understanding by more groups. These demands have driven the development of text style transfer, making it an important direction in natural language processing.

The research methods for text style transfer can be broadly classified into three categories: supervised learning, semi-supervised learning, and unsupervised learning. The main differences among these three lie in the degree of reliance on language resources and the trade-offs made between the generalization ability and stability of the

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_39

models. With the development of deep learning technology and pre-trained models, these three methods have gradually evolved into a rich variety of variants and combinations.

This article mainly conducts a review from the following three perspectives. Firstly, it introduces the task definition and classification of text style transfer. Secondly, it discusses methods based on supervised, semi-supervised, and unsupervised learning respectively. Thirdly, it summarizes the commonly used datasets and evaluation metrics. Finally, it analyzes the limitations of existing methods and the future development directions. Through a systematic review, this article aims to provide references and inspiration for subsequent research.

2 Text Style Transfer Based on Supervised Learning

Supervised learning is the most direct and traditional method for text style transfer. Its core idea is to train the model using a large amount of parallel corpora, enabling the mapping relationship between the input text and the corresponding target style text to be automatically learned. In other words, supervised learning equates the style transfer problem to the machine translation task: although the source sentence and the target sentence have basically the same semantics, there are significant differences in expression (such as sentiment, tone, formality). The goal of model training is to predict the differences between the sentence and the real target sentence, thereby gradually mastering the balance between content preservation and style transformation. At the modeling level, the supervised learning method usually adopts the Seq2Seq framework, consisting of an encoder and a decoder. The encoder converts the source text into context semantic vectors, while the decoder gradually generates new sentences under the condition of the target style. To further enhance the model's capturing ability, researchers often combine the attention mechanism, enabling the decoder to dynamically focus on the key information in the input text during generation. With the development of deep learning, the Transformer architecture has gradually replaced the traditional RNN, achieving higher performance in text generation and style transformation. At the data level, supervised learning relies on manually constructed parallel style datasets. For example, the GYAFC dataset constructed by Rao and Tetreault contains a large number of sentence pairs with formal and informal expressions, allowing the model to learn the transformation rules between formal and everyday styles. Such datasets provide a unified evaluation benchmark for researchers and have also promoted the development of this field [1]. However, building large-scale and high-quality parallel data is often costly, which limits the application scope of supervised learning methods. In terms of model implementation, the "Delete-Retrieve-Generate" framework proposed by Li et al is a typical editing-based approach [2]. It assumes that the style is mainly determined by local words or phrases, so it first removes the style markers from the source text through the "deletion" operation, then retrieves suitable replacement segments from the target corpus, and finally generates complete sentences. This method strikes a balance between interpretability and flexibility, and performs particularly well in sentiment style transfer tasks. Another type

of method emphasizes the replication and alignment mechanisms. For example, the "Copy-enriched Seq2Seq model" proposed by Jhamtani et al. performs exceptionally well in the transfer from Shakespearean style to modern English [3]. During the decoding process, this model can directly replicate some segments of the input text, thereby ensuring the integrity of the content. At the same time, it achieves style transformation by learning to replace local words. Additionally, some studies have attempted to introduce reinforcement learning and multi-task learning ideas into supervised learning. [Zhang and Lapata] ([4]) introduced a reinforcement learning mechanism based on a reward function in sentence simplification and style-related tasks, enabling the model to optimize for fluency, simplicity, and semantic preservation [4]. [Niu and Bansal] proposed a multi-task learning framework, jointly training style transfer and dialogue generation tasks, thereby enhancing the model's generalization ability on small-scale datasets [5]. Overall, supervised learning methods can generate high-quality target style texts under conditions with sufficient parallel corpora, possessing advantages such as stable performance and strong style controllability. However, the bottleneck of this type of method lies in its strong dependence on parallel data, making it difficult to be extended to diverse style scenarios in the real world. Therefore, although supervised learning laid the foundation for text style transfer, the research focus is gradually shifting towards semi-supervised and unsupervised methods to alleviate the problem of data scarcity.

3 Text Style Transfer Based on Semi-supervised Learning

The core of semi-supervised learning methods lies in combining limited parallel corpora and large-scale non-parallel corpora to strike a balance between data utilization efficiency and model generalization ability. Compared with supervised learning, semi-supervised methods significantly reduce the reliance on manually annotated data; compared with unsupervised learning, semi-supervised methods can utilize a small amount of real parallel pairs to stabilize the training process. The training objective usually consists of three parts: supervised loss: using the existing parallel data to directly optimize the differences between the source sentence and the target sentence; unsupervised loss: generating pseudo parallel sentence pairs using non-parallel data, or improving stability through consistency constraints; discriminator loss: the idea of GAN, using the discriminator to constrain the style attributes of generated sentences. The Cross-Alignment framework proposed by Shen et al. is a representative of semi-supervised methods [6]. It aligns sentences of different styles in a shared semantic space and then uses the discriminator to constrain style consistency, thereby generating pseudo parallel data and gradually improving model performance. This framework has pioneered the idea of using semantic space alignment to solve style transfer. In terms of the introduction of reinforcement learning, Xu et al.proposed the Cycled Reinforcement Learning method, which uses "forward transfer - reverse transfer" cyclic training to gradually improve style consistency and semantic retention in the absence of parallel data [7]. This method uses policy gradient to optimize the reward signal, effectively alleviating the problem of content loss. Yang et al. further proposed the

DualRL framework, using bidirectional generators and discriminators, and optimizing the style accuracy through reinforcement learning signals, achieving significant results in sentiment transfer tasks [8]. Another approach is to introduce machine translation methods. The unsupervised machine translation framework proposed by Lample et al. uses recurrent consistency and adversarial training to enable the model to learn cross-language mapping without parallel data [9]. The idea of this framework has been widely borrowed in the field of style transfer to achieve a similar "translation" process between different style domains. Additionally, Wu et al. proposed the Mask-and-Fill method, generating pseudo labels by randomly masking and predicting missing fragments in sentences, and then iteratively optimizing the model using these data [10]. This method significantly improves the model's utilization of non-parallel corpora under the semi-supervised framework and reduces reliance on real parallel data. In summary, semi-supervised methods provide a compromise solution between performance and data availability. It can fully utilize large-scale internet text under limited annotation conditions and achieve results close to those of supervised methods. However, its challenge lies in the difficulty in ensuring the quality of pseudo parallel data. If generated unstably, it may introduce noise and amplify errors.

4 Text Style Transfer Based on Unsupervised Learning

Unsupervised learning methods rely entirely on non-parallel corpora and are the core direction of current text style transfer research. Their main goal is to decouple content and style while maintaining semantic consistency during generation. The difficulty of unsupervised methods lies in the lack of clear supervisory signals, so they usually combine mechanisms such as adversarial training, cycle consistency, edit-based generation, and controllable decoding. In terms of representation learning and adversarial decoupling, the Cross-Aligned Auto-Encoder proposed by Shen et al. attempts to separate content and style by sharing semantic representations and adversarial discriminators, enabling cross-style generation without parallel corpora [6]. The Multi-Decoder framework proposed by Fu et al. designs independent decoders for different styles, allowing the model to output target style text while maintaining content [11]. The Adversarially Regularized Autoencoder (ARAE) proposed by Zhao et al. uses adversarial regularization to constrain the generator and improves the stability of style decoupling [12]. In the context of cycle consistency and back-translation methods, the Back-Translation framework proposed by Prabhumoye guides the hidden states to move towards the target style through gradient guidance and introduces a cycle consistency loss to ensure semantic preservation. This mechanism significantly improves the comprehensibility and consistency of generated text in the absence of parallel data [13]. The edit-based methods, represented by Li et al.'s Delete-Retrieve-Generate, assume that style information is mainly determined by local words or phrases, and achieve transfer through deletion, retrieval, and generation [1]. These methods are not only highly interpretable but also show high controllability in emotion and tone transfer tasks. With the development of large-scale pre-trained language models, researchers have begun to explore controllable decoding. PPLM proposed by Dathathri

guides the hidden states to move towards the target style during decoding through gradients [14]; GeDi proposed by Krause reweights the word distribution during generation to achieve style control [15]; FUDGE proposed by Yang and Klein predicts "future" attributes to dynamically adjust the probability distribution, enhancing the flexibility of control [16]. Recently, Diffusion-LM proposed by Li et al. models text generation as a diffusion process and applies style and semantic constraints in the latent space, achieving more fine-grained control effects [17]. In summary, unsupervised methods have broad application prospects in real open environments and can cover a wider range of style types. However, these methods also have issues such as insufficient model stability, complex loss design, and an incomplete evaluation system. Therefore, unsupervised learning is not only a current research hotspot but also represents the future development direction.

5 Dataset

The dataset is the core foundation of text style transfer research. Based on whether it has parallel annotations, datasets can be divided into two types: Parallel datasets: Each source text corresponds to a target text with a different style but the same content, suitable for supervised learning. However, its construction cost is high and the coverage is limited. Non-parallel datasets: Only contain a single text and style labels, closer to the real Internet environment, covering diverse styles, and thus widely used in unsupervised and semi-supervised methods.

6 Evaluation Scheme

The manual evaluation dimensions include: accuracy of style transfer, content retention, and text fluency. Manual evaluation can comprehensively and flexibly grasp the quality, but its drawbacks are high cost and poor repeatability, which makes it unsuitable for large-scale experiments. Automatic evaluation: common automated indicators include style transfer accuracy, content retention rate, and fluency. These methods are efficient, low-cost, and suitable for large-scale experiments, but their accuracy and comprehensiveness are limited, especially their correlation with manual evaluation is low.

The existing automatic evaluation indicators are difficult to comprehensively measure the performance of text style transfer. There is often a conflict among style accuracy, content retention, and fluency, and there is a lack of a unified comprehensive indicator. At the same time, the correlation between automatic evaluation and manual evaluation is weak, and this gap becomes more significant as the model performance improves. Moreover, the application of TST has a dual nature: on the one hand, it can promote the development of information dissemination and human-computer interaction; on the other hand, it may also be used to generate false information or fraudulent texts, bringing potential risks. Therefore, research needs to balance technological advancement with ethical and security issues.

7 Conclusions

This article provides a systematic review of the development history of text style transfer, covering task definition, research methods, datasets, evaluation schemes, and their application impacts. Overall, TST, as an important branch of natural language processing, is constantly evolving. It has shifted from relying on parallel corpora for supervised learning to more flexible and realistic unsupervised and semi-supervised learning. Future research directions may focus on the following points: larger-scale and higher-quality dataset construction; multi-dimensional style modeling and transfer; comprehensive and interpretable automatic evaluation metrics design; and security and ethical norms in technology applications. In conclusion, text style transfer is not only a technical challenge but also an important topic combining artificial intelligence and social applications. Its development will profoundly affect the way human-computer communication and knowledge dissemination occur.

References

1. Rao, S., Tetreault, J.: Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. arXiv preprint arXiv:1803.06535 (2018)
2. Li, J., Jia, R., He, H., Liang, P.: Delete, retrieve, generate: A simple approach to sentiment and style transfer. arXiv preprint arXiv:1804.06437 (2018)
3. Jhamtani, H., Gangal, V., Hovy, E., Nyberg, E.: Shakespearizing modern language using copy-enriched sequence-to-sequence models. arXiv preprint arXiv:1707.01161 (2017)
4. Zhang, X., Lapata, M.: Sentence simplification with deep reinforcement learning. arXiv preprint arXiv:1703.10931 (2017)
5. Niu, T., Bansal, M.: Polite dialogue generation without parallel data. Transactions of the Association for Computational Linguistics 6, 373–389 (2018)
6. Shen, T., Lei, T., Barzilay, R., Jaakkola, T.: Style transfer from non-parallel text by cross-alignment. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
7. Xu, J., Sun, X., Zeng, Q., Ren, X., Zhang, X., Wang, H., Li, W.: Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. arXiv preprint arXiv:1805.05181 (2018)
8. Yang, Z., Hu, Z., Dyer, C., Xing, E.P., Berg-Kirkpatrick, T.: Unsupervised text style transfer using language models as discriminators. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
9. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.A.: Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043 (2017)
10. Wu, X., Zhang, T., Zang, L., Han, J., Hu, S.: Mask and infill: Applying masked language model to sentiment transfer. arXiv preprint arXiv:1908.08039 (2019)
11. Fu, Z., Tan, X., Peng, N., Zhao, D., Yan, R.: Style transfer in text: Exploration and evaluation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1 (2018)
12. Zhao, J., Kim, Y., Zhang, K., Rush, A., LeCun, Y.: Adversarially regularized autoencoders. In: International Conference on Machine Learning, pp. 5902–5911. PMLR (2018)
13. Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., Black, A.W.: Style transfer through back-translation. arXiv preprint arXiv:1804.09000 (2018)

14. Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Liu, R.: Plug and play language models: A simple approach to controlled text generation. arXiv preprint arXiv:1912.02164 (2019)
15. Krause, B., Gotmare, A.D., McCann, B., Keskar, N.S., Joty, S., Socher, R., Rajani, N.F.: Gedi: Generative discriminator guided sequence generation. arXiv preprint arXiv:2009.06367 (2020)
16. Yang, K., Klein, D.: FUDGE: Controlled text generation with future discriminators. arXiv preprint arXiv:2104.05218 (2021)
17. Li, X., Thickstun, J., Gulrajani, I., Liang, P.S., Hashimoto, T.B.: Diffusion-LM improves controllable text generation. In: Advances in Neural Information Processing Systems, vol. 35, pp. 4328–4343 (2022).