entropy

MDPI

Article

Quick Estimate of Information Decomposition for Text Style Transfer

Viacheslav Shibaev [1], Eckehard Olbrich [2], Jürgen Jost [2] and Ivan P. Yamshchikov [2,3,*]

[1] Department of Intelligent Information Technologies, Ural Federal University, 620075 Ekaterinburg, Russia
[2] Max Planck Institute for Mathematics in the Sciences Leipzig, 04103 Leipzig, Germany
[3] CEMAPRE, University of Lisbon, 1649-004 Lisboa, Portugal
* Correspondence: ivan@yamshchikov.info

Abstract: A growing number of papers on style transfer for texts rely on information decomposition. The performance of the resulting systems is usually assessed empirically in terms of the output quality or requires laborious experiments. This paper suggests a straightforward information theoretical framework to assess the quality of information decomposition for latent representations in the context of style transfer. Experimenting with several state-of-the-art models, we demonstrate that such estimates could be used as a fast and straightforward health check for the models instead of more laborious empirical experiments.

Keywords: text style transfer; natural language processing; information decomposition

Citation: Shibaev, V.; Olbrich, E.; Jost, J.; Yamshchikov, I.P. Quick Estimate of Information Decomposition for Text Style Transfer. *Entropy* **2023**, *25*, 322. https://doi.org/10.3390/e25020322

Academic Editors: Adam Lipowski, Jaroslaw Krzywanski, Yunfei Gao, Marcin Sosnowski, Karolina Grabowska, Dorian Skrobek, Ghulam Moeen Uddin, Anna Kulakowska, Anna Zylka and Bachil El Fil

Received: 20 November 2022
Revised: 21 January 2023
Accepted: 1 February 2023
Published: 10 February 2023

1. Introduction

Natural language generation (NLG) is a challenging task. The discrete nature of textual information [1] leads to non-smooth disentangled representations and the absence of local information continuity [2] that make natural language generation even more complicated. One of the NLG tasks is style transfer for texts. This task is often addressed in the context of disentangled latent representations [1,3–9]. These works use an encoder–decoder architecture with one or multiple style discriminators to improve latent representations. An encoder takes a given sentence as an input and generates a style-independent content representation. The decoder then uses this content representation and a target style representation to generate a new sentence in the needed style (for a detailed review of modern text style transfer we address the reader to [10]).

There is a variety of benchmarks and methods used to compare the relative performance of the proposed architectures, see [11–13]. Yet, there is little rigorous work on how one could assess the quality of the resulting representations. For example, ref. [14] demonstrate that the quality of style and content decomposition depends on the particular architecture and models, with better information decomposition quality outperforming the state-of-the-art models in terms of BLEU (bilingual evaluation understudy) [15] between output and human-written reformulations. (BLEU's output is always a number between 0 and 1. This value indicates how similar the candidate text is to the reference texts, with values closer to 1 representing more similar texts.) However, the experiments that illustrate this are very computationally intensive. This paper suggests assessing the quality of the obtained representations in a more effective, straightforward way and shows that the proposed theoretical estimates correspond to the empirical results. Since information decomposition within a latent representation might play a key role in minute text manipulation, ref. [1] hope that a computationally light model that assesses the quality of information decomposition in a given architecture could be instrumental for further research in natural language generation.

2. Related Work

The problem of text style transfer (TST) needs a more rigorous definition [16]. However, there are some attempts to quantify literary style, see [17]. Ref. [18] states that stylized texts could be generated if a system is trained on a dataset of stylistically similar texts. Ref. [19] show that the literary styles of the authors could be learned end-to-end.

Most recent contributions in the field address specific narrow aspects of style that could be empirically measured. Such stylistic attributes of text range from politeness [20], the *'style of the time'* [17] and formality of speech [11] to author-specific attributes (see [21] or [22] on 'shakespearization'), gender or political slant [23]. These attributes themselves are defined with varying degrees of rigor. Refs. [4,24] define a style as a set of arbitrary quantitatively measurable categorical or continuous parameters that could be automatically estimated with an external classifier. Further in this paper, we work with this empirical paradigm of literary style. It is widely used in modern text style transfer research since it allows natural extensions due to the compositionality of stylistic features. For example, ref. [12] provide a dataset for fine-grained stylistic changes as building blocks for more complex, high-level transfers, or [25] suggest treating style transfer as one-to-many mapping instead of one-to-one correspondence.

Many TST contributions either use an idea of an adversarial component to ensure that semantic representations contain no stylistic information [4] or combine it with some additional constrictions. For example, ref. [3] apply a GAN to align hidden representations of sentences from two corpora and use an adversarial loss to decompose information about the form of a sentence. Ref. [6] introduce adversarial–motivational training that includes a special motivational loss to encourage a better decomposition. Ref. [5] develop a structured content-preserving model that leverages linguistic information in the structured fine-grained supervision to preserve the style-independent content better. Ref. [7] show that the decomposition of style and content could be improved with an auxiliary multi-task for label prediction and adversarial objective for a bag-of-words prediction.

Recently, ref. [26] propose a new information-theory-motivated architecture for style transfer. They develop a method that leverages mutual information upper bound to measure dependence between style and content. Ref. [27] propose a method that can decompose speech into four components by introducing three information bottlenecks. The majority of the approaches mentioned above use some form of disentangled representation learning (DRL) [7], yet there are only a handful of methods to assess the relative quality of the obtained representations provided by different architectures. This paper addresses latent representation quality assessment in an information-theoretic framework called partial information decomposition. We propose a straightforward information-theory-based approach and demonstrate that the proposed estimates correspond to the empirical results but are significantly less computationally demanding.

In this paper, we experiment with a subtask of sentiment transfer. There is a discussion if the sentiment of a text could be regarded as its stylistic attribute, see [28]. However, numerous style transfer papers regard sentiment transfer as a viable task for the style transfer system. For example, refs. [29–31] estimate the quality of the style transfer with pre-trained binary sentiment classifiers.

3. Style Transfer

Consider the text style transfer that comprises an encoder–generator pair $\mathcal{M} = \{f_{\theta_{\text{enc}}}, f_{\theta_{\text{gen}}}\}$ parameterized by neural networks. The input to the model is a sentence $\mathbf{x} = (\mathbf{w}_1, \ldots, \mathbf{w}_T)$ where $\mathbf{w}_i \in \mathbb{R}^{d_w}$ and its style variable $y \in \{0, 1\}$. The encoder maps $\mathbf{x}$ to a latent representation $\mathbf{z} \in \mathbb{R}^{d_z}$. The generator then takes $\mathbf{z}$ and y as inputs to generate a new sentence $\hat{\mathbf{x}}$. Ideally, changing the value of y should generate $\hat{\mathbf{x}}$ in a different style.

There are various methods for the evaluation of text style transfer models. For a detailed overview of modern semantic similarity measures and their applicability to the problem of style transfer, we address the reader to [32]. Instead of assessing the system's overall performance, this paper focuses on the latent space that the style transfer model

uses. Several TST papers state that precise text manipulation is enabled through effective information representation. However, there is no method that could compare the quality of latent spaces obtained by two different architectures. In this paper, we demonstrate that one can compare the rate of information decomposition achieved by a given model using measures from information theory.

4. Qualifying Latent Representations with Coinformation

Mutual information (MI) was originally proposed in Claude Shannon's article "A Mathematical Theory of Communication" [33]. Given three jointly distributed random variables $(X, Y, Z) \sim P$, the mutual information between X and Y and Z can be decomposed into information that Y has about X that is *unknown* to Z (we call this the *unique* information of Y w.r.t. Z) and information that Y has about X that is *known* to Z (we call this the *shared* or *redundant* information. Using the chain rule, the mutual information between X and (Y, Z) can be decomposed into four terms:

$$\begin{aligned} & I(X;Y,Z) \\ = & \underbrace{UI(X;Y\backslash Z)}_{\text{unique information of } Y \text{ w.r.t. } Z} + \underbrace{SI(X;Y,Z)}_{\text{shared information}} \\ + & \underbrace{UI(X;Z\backslash Y)}_{\text{unique information of } Z \text{ w.r.t. } Y} + \underbrace{CI(X;Y,Z)}_{\text{complementary information}}. \end{aligned} \tag{1}$$

This decomposition is part of a framework called *partial information decomposition (PID)* and was originally proposed by [34]. Ref. [35] made a now widely used proposal for concrete measures for the terms in Equation (1). The difference of the shared and synergistic information is equal to the *coinformation* [36,37], a symmetric measure of the correlation between three random variables:

$$\begin{aligned} CoI(X;Y;Z) &= SI(X;Y,Z) - CI(X;Y,Z) \\ &= I(X;Y) - I(X;Y|Z) \\ &= I(Y;Z) - I(Y;Z|X) \\ &= I(X;Z) - I(X;Z|Y) \end{aligned} \tag{2}$$

Coinformation is also widely used in neurosciences, with negative values interpreted as synergy and positive values as redundancy.

In the text style transfer (TST) setting, X represents the input text, Y its stylistic content and Z is the latent representation of the input that ideally should only capture the semantic content of X. In terms of the information decomposition in Equation (1) this would mean that when we decompose X there is unique information while the shared and the complementary information vanish. Since style Y is independent of latent representation Z given original input text X, the following statement holds

$$\begin{aligned} CoI(X;Y;Z) &= I(Y;Z) - \underbrace{I(Y;Z|X)}_{=0} \\ &= I(Y;Z). \end{aligned} \tag{3}$$

We propose to use $I(Y;Z)$ as a proxy to measure the quality of the latent representation Z. Because Equation (3) implies that the shared information $SI(X;Y,Z)$ is always greater than or equal to the complementary information $CI(X;Y,Z)$, we can say that a successful text style transfer should transfer only the semantic aspect of X into Z and therefore has low $I(Y;Z)$. In other words, these models should learn the latent representation Z in such a way that it keeps the redundant information $SI(X;Y,Z)$ as low as possible.

5. Experiments

This section calculates the proposed $I(Y;Z)$ for eight different style transfer architectures and shows how such a quantity can characterize the quality of the information decomposition in two given latent spaces.

5.1. Calculating Mutual Information

A framework for MI estimation between two continuous distributions was proposed [38]. This method is based on the estimator for differential entropy. In particular, one could apply the framework for the case where one of the distributions is continuous and another one is discrete [39] as follows

$$I(X,Y) = \psi(N) - \langle \psi(N_x) \rangle + \psi(k) - \langle \psi(m) \rangle,$$

where $\psi(\cdot)$ is the digamma function, $\langle \cdot \rangle$ is the overall averaging of points from the dataset, N is the total number of points in the dataset, N_x is the number of points for the given value x of discrete distribution, k is the number of nearest neighbors and m is the number of points which are closer than the k-th neighbor to the given point.

We conduct our experiments on [5] human-rewritten Yelp! Reviews: the dataset contains 998 original and 998 reformulated Yelp! Reviews that are rewritten into either positive or negative sentiment. As shown in Figure 1, we experiment with six different text style transfer models, namely [1]'s autoencoder with discriminators that we further denote as (**Baseline**);autoencoder model with discriminators (**ZDiscr**), shifted autoencoder (**SAE**) and shifted autoencoder with discriminators (**SAEZDiscr**) introduced in [9]; Ref. [3]'s autoencoder for mapping texts written in different styles in the same latent space (**Shen**); Ref. [5]'s autoencoder with discriminator which trained with additional language model and part of speech losses (**TianFull**); a version of TianFull trained without additional language model loss (**TianWithoutLM**); and a version of TianFull trained without additional part of speech loss (**TianWithoutPOS**). In every figure, solid lines represent inputs and the dashed blue lines connect inputs compared by discriminators. In contrast, the dashed red line stands for the soft output of the architecture passed to the encoder to explicitly minimize the distance between latent representations of input and output.

Out of the architectures in question, only the one proposed by [3] does not use additional tools to estimate the quality of the output $\hat{\mathbf{x}}$ to improve information decomposition in the encoder. Instead, it explicitly tries to minimize the distance between the aligned mapping of the sentences with different styles.

5.2. Exploring Latent Spaces

Let us obtain some intuition on the underlying geometry of the obtained latent spaces. k-means clustering could be a straightforward and intuitive method to obtain such intuition. K-means, by default, uses a within-cluster sum of squares criterion. Since the estimate in Equation (4) uses nearest neighbors, it could be prone to noise depending on the structure of the resulting latent space. Figure 2 shows how the within-cluster sum of squares criterion changes for all architectures and their modifications as we choose different numbers of clusters. Figure 2 highlights vital differences between architectures. The lower clustering coefficient implies local dense clusters in a latent space, while the higher clustering coefficient implies that clusters are not dense. We see that some architectures such as SAE provide latent spaces with a smaller number of dense clusters, while others, like the model introduced in [5], obtain latent representations with no distinct structure, rather, latent representations form a cloud of points in the latent space.

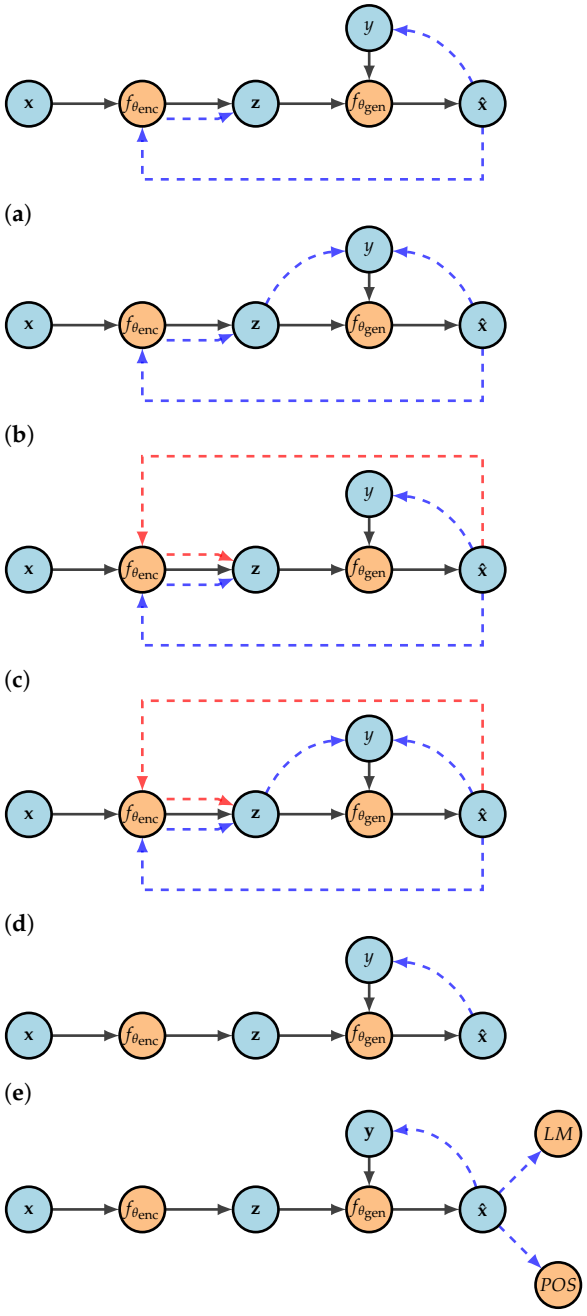

Figure 1. Text style transfer models evaluated in this work. Dashed lines indicate auxiliary training procedures that ensure **z** capturing the semantic content: the discriminators are denoted with blue color and the distance constraint is denoted with red. (**a**) **Baseline**: Variational autoencoder with discriminators [1]. (**b**) **ZDiscr**: Autoencoder with an additional discriminator [9]. (**c**) **SAE**: Shifted autoencoder [9]. (**d**) **SAEZDiscr**: Shifted autoencoder with an additional discriminator [9]. (**e**) **Shen et al. (2017)**: Aligned autoencoder [3]. (**f**) **Tian et al. (2018)**: Autoencoder with LM and POS losses [5].

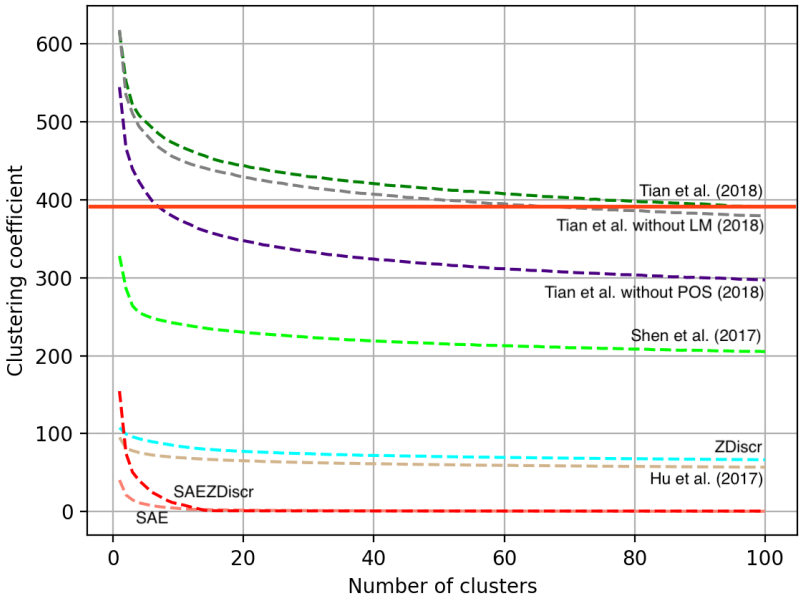

Figure 2. Clustering coefficient as a function of a number of clusters in a latent space obtained by architectures. Refs. [1,3,5].

Having this basic intuition, let us now calculate a mutual information estimate according to Equation (4) setting the number of nearest neighbors to three as recommended in [39]. Figure 3 shows the obtained estimates of MI alongside BLEU scores between model outputs and human rewrites. Higher BLEU corresponds to better performance of the model. One sees that there is a general correspondence between the proposed MI estimate and the performance of the model. The nature of the models is different, and the results are prone to noise, yet there is a general tendency that models that achieve lower mutual information between stylistic variable and semantic representation tend to perform better in terms of BLEU between model outputs and actual human rewrites.

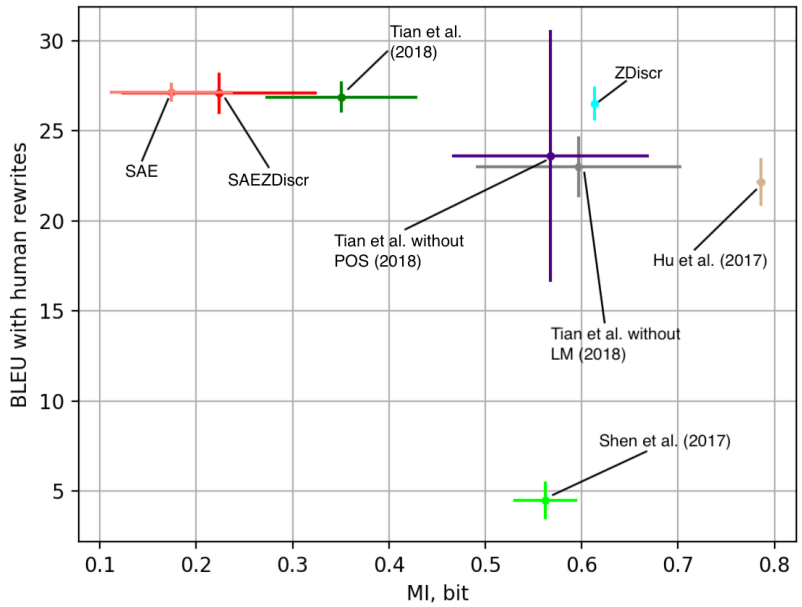

Figure 3. Estimated mutual information between target style and latent representation and BLEU between model output and human rewrites. Lower mutual information corresponds to higher BLEU and better performance. Refs. [1,3,5].

5.3. Correspondence with Empirical Results

In [14], the authors used a method proposed originally in [40] to see if better information decomposition corresponds to better style transfer performance. We reproduce and enhance these experiments here to see if our methodology aligns with these empirical results. The methodology originally proposed in [40] is as follows. To see if the encoder manages to decompose semantic and stylistic information, one can train a stand-alone artificial neural network that tries to predict the style of the input using its resulting latent representation. The lower accuracy of such classifiers corresponds to the higher quality of information decomposition. In [9], the authors show that the architectures used for style transfer are noisy. This means that to calculate the standard deviation for the results obtained with this method of information decomposition quality estimation, one has to retrain every architecture from scratch. Such experiments are computationally intensive and have to be tailored for every architecture.

Ref. [40] train a stand-alone artificial neural network that tries to predict the style of the input using its resulting latent representation. The authors suggest that the lower accuracy of such classifiers corresponds to a higher quality of information decomposition. Figure 4 shows the measurement of such external classifiers' accuracy along with the estimates of MI proposed earlier in this paper. We run eight experiments with every architecture. Figures 3 and 4 show standard deviations of the results along with the average numbers. One could see that the architectures with lower external classifier accuracy tend to deliver better performance in terms of BLEU with human-written rewrites. This means that measures for information decomposition quality could be useful for further advances in the research on minute language manipulation. However, this methodology demands the training of a stand-alone artificial neural network, and the results are prone to error due to inadequate choice of architecture. Naturally, if the architecture is not complex enough in comparison with the encoder, it could not provide adequate results, see [14]. However, comparing Figures 3 and 4, one can see that our information-theory-inspired methodology gives results that are in line with the method proposed by [40] yet uses a fraction of computational resources and is not prone to the errors due to inadequate architecture design. One could use the proposed methodology as a simple and straightforward health check for the models that rely on information decomposition.

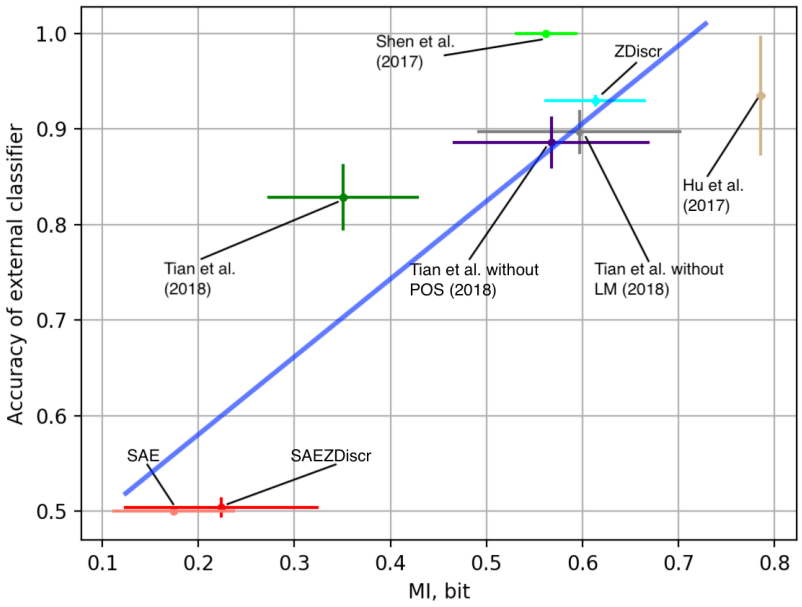

Figure 4. Accuracy of an external classifier and the proposed mutual information estimates, $R^2 = 0.78$. The proposed method for mutual information estimations obtains results similar to the previous empirical methods. Refs. [1,3,5].

6. Discussion

Out of the eight models that we experiment with, there is one that stands out, namely, the method proposed in [3]. Figure 3 shows that despite the relatively low value of mutual information estimation, the model performs rather poorly in terms of the BLEU between human rewrites and the model's output. Indeed, unlike other methods, [3] uses cross-alignment of the training data and does not provide any specific mechanism for estimation of the output quality as a part of the model. Instead, it tries to force distributional alignment over the latent space or sentence populations since the model is originally developed to work with nonparallel corpora rather than with style transfer on parallel texts. Thus, the model minimizes mutual information between Y and Z but does not explicitly maximize $I(X, Y, Z)$. Under these structural assumptions, better information decomposition obtained does not guarantee better performance of the model, which we see in the experiments. This highlights the clear limitation of the proposed method: if the architecture does not explicitly maximize $I(X, Y, Z)$, the decomposition quality assessment is not aligned with the performance of the model on the downstream task. Another explicit limitation of the proposed method corresponds to the applicability of the estimation methods proposed [38,39]: latent representation distribution has to be continuous while the distribution of the stylistic variable has to be discrete.

7. Conclusions

In this work, we have presented an alternative approach for evaluating the latent representation learned by TST models. If models learn the latent representations that capture only the semantic content of the inputs, the low value of mutual information between the latent representations and the target variable could measure the relative information decomposition quality obtained by various systems. Such methodology yields results in line with previous experiments yet is transparent and computationally efficient.

Author Contributions: Conceptualization, E.O., J.J. and I.P.Y.; Methodology, E.O., J.J. and I.P.Y.; Software, V.S.; Validation, V.S.; Investigation, V.S.; Writing–original draft, I.P.Y.; Writing–review & editing, E.O. and J.J.; Supervision, J.J. and I.P.Y.; Project administration, I.P.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; Xing, E.P. Toward Controlled Generation of Text. In Proceedings of the International Conference on Machine Learning, Ho Chi Minh City, Vietnam, 13–16 January 2017; pp. 1587–1596.
2. Bowman, S.R.; Angeli, G.; Potts, C.; Manning, C.D. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 632–642.
3. Shen, T.; Lei, T.; Barzilay, R.; Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6830–6841.
4. 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, New Orleans, LA, USA, 2–3 February 2018; Volume 32.
5. Tian, Y.; Hu, Z.; Yu, Z. Structured Content Preservation for Unsupervised Text Style Transfer. *arXiv* **2018**, arXiv:1810.06526
6. Romanov, A.; Rumshisky, A.; Rogers, A.; Donahue, D. Adversarial Decomposition of Text Representation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Minneapolis, MI, USA, 2019; pp. 815–825.
7. John, V.; Mou, L.; Bahuleyan, H.; Vechtomova, O. Disentangled Representation Learning for Non-Parallel Text Style Transfer. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 424–434.
8. Halevi, G.; Wadhawan, K. Text Style Transfer Using Partly-Shared Decoder. In Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 4–6 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1563–1567.

9. Tikhonov, A.; Shibaev, V.; Nagaev, A.; Nugmanova, A.; Yamshchikov, I.P. Style Transfer for Texts: Retrain, Report Errors, Compare with Rewrites. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3927–3936.
10. Hu, Z.; Lee, R.K.W.; Aggarwal, C.C.; Zhang, A. Text style transfer: A review and experimental evaluation. *ACM Sigkdd Explor. Newsl.* **2022**, *24*, 14–45. [CrossRef]
11. Rao, S.; Tetreault, J. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 129–140.
12. Lyu, Y.; Liang, P.P.; Pham, H.; Hovy, E.; Poczós, B.; Salakhutdinov, R.; Morency, L.P. StylePTB: A Compositional Benchmark for Fine-grained Controllable Text Style Transfer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 2116–2138.
13. Briakou, E.; Lu, D.; Zhang, K.; Tetreault, J. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 3199–3216.
14. Yamshchikov, I.P.; Shibaev, V.; Nagaev, A.; Jost, J.; Tikhonov, A. Decomposing Textual Information For Style Transfer. In Proceedings of the 3rd Workshop on Neural Generation and Translation, Hong Kong, China, 4 November 2019; pp. 128–137.
15. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. GBLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
16. Xu, W. From Shakespeare to Twitter: What are Language Styles all about? In Proceedings of the Workshop on Stylistic Variation, Copenhagen, Denmark, 8 September 2017; pp. 1–9.
17. Hughes, J.M.; Foti, N.J.; Krakauer, D.C.; Rockmore, D.N. Quantitative patterns of stylistic influence in the evolution of literature. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 7682–7686. [CrossRef] [PubMed]
18. Potash, P.; Romanov, A.; Rumshisky, A. GhostWriter: Using an LSTM for Automatic Rap Lyric Generation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, 17–21 September 2015; pp. 1919–1924.
19. Yamshchikov, I.P.; Tikhonov, A. Learning Literary Style End-to-end with Artificial Neural Networks. *Adv. Sci. Technol. Eng. Syst. J.* **2019**, *4*, 115–125. [CrossRef]
20. Sennrich, R.; Haddow, B.; Birch, A. Controlling politeness in neural machine translation via side constraints. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 35–40.
21. Xu, W.; Ritter, A.; Dolan, W.B.; Grishman, R.; Cherry, C. Paraphrasing for style. In Proceedings of the COLING, Mumbai, India, 8–15 December 2012; pp. 2899–2914.
22. Jhamtani, H.; Gangal, V.; Hovy, E.; Nyberg, E. Shakespearizing Modern Language Using Copy-Enriched Sequence-to-Sequence Models. In Proceedings of the Workshop on Stylistic Variation, Copenhagen, Denmark, 8 September 2017; pp. 10–19.
23. Prabhumoye, S.; Tsvetkov, Y.; Black, A.W.; Salakhutdinov, R. Style Transfer Through Back-Translation. *arXiv* **2018**, arXiv:1804.09000.
24. Ficler, J.; Goldberg, Y. Controlling linguistic style aspects in neural language generation. In Proceedings of the Workshop on Stylistic Variation, Copenhagen, Denmark, 8 September 2017; pp. 94–104.
25. Lin, K.; Liu, M.Y.; Sun, M.T.; Kautz, J. Learning to Generate Multiple Style Transfer Outputs for an Input Sentence. In Proceedings of the Fourth Workshop on Neural Generation and Translation, Online, 10 July 2020; pp. 10–23.
26. Cheng, P.; Min, M.R.; Shen, D.; Malon, C.; Zhang, Y.; Li, Y.; Carin, L. Improving Disentangled Text Representation Learning with Information-Theoretic Guidance. *arXiv* **2020**, arXiv:2006.00693.
27. Qian, K.; Zhang, Y.; Chang, S.; Hasegawa-Johnson, M.; Cox, D. Unsupervised speech decomposition via triple information bottleneck. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual, 13–18 July 2020; pp. 7836–7846.
28. Tikhonov, A.; Yamshchikov, I.P. What is wrong with style transfer for texts? *arXiv* **2018**, arXiv:1808.04365.
29. Kabbara, J.; Cheung, J.C.K. Stylistic transfer in natural language generation systems using recurrent neural networks. In Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods, Austin, TX, USA, 5 November 2016; pp. 43–47.
30. Li, J.; Jia, R.; He, H.; Liang, P. 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, New Orleans, LA, USA, 1–6 June 2018; Volume 1, pp. 865–1874.
31. 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* **2018**, arXiv:1805.05181.
32. Yamshchikov, I.P.; Shibaev, V.; Khlebnikov, N.; Tikhonov, A. Style-transfer and Paraphrase: Looking for a Sensible Semantic Similarity Metric. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 35, pp. 14213–14220.
33. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]

34. Williams, P.; Beer, R. Nonnegative Decomposition of Multivariate Information. *arXiv* **2010**, arXiv:1004.2515v1.
35. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183. [CrossRef]
36. McGill, W. Multivariate information transmission. *Trans. Ire Prof. Group Inf. Theory* **1954**, *4*, 93–111. [CrossRef]
37. Bell, A.J. The co-information lattice. In Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA, Granada, Spain, 22–24 September 2003; Volume 2003.
38. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef] [PubMed]
39. Ross, B.C. Mutual information between discrete and continuous data sets. *PLoS ONE* **2014**, *9*, e87357. [CrossRef] [PubMed]
40. Subramanian, S.; Lample, G.; Smith, E.M.; Denoyer, L.; Ranzato, M.A.; Boureau, Y.L. Multiple-Attribute Text Style Transfer. *arXiv* **2018**, arXiv:1811.00552.