Style Transfer for Texts: Retrain, Report Errors, Compare with Rewrites

Alexey Tikhonov*
Yandex
Berlin, Germany
altsoph@gmail.com

Viacheslav Shibaev*
Ural Federal University
Ekaterinburg, Russia

Aleksander Nagaev
Ural Federal University/Sberbank
Ekaterinburg, Russia

Aigul Nugmanova
Speech Technology Center
St. Petersburg, Russia

Ivan P. Yamshchikov*
Max Planck Institute
for Mathematics in the Sciences
Leipzig, Germany
ivan@yamshchikov.info

Abstract

This paper shows that standard assessment methodology for style transfer has several significant problems. First, the standard metrics for style accuracy and semantics preservation vary significantly on different re-runs. Therefore one has to report error margins for the obtained results. Second, starting with certain values of bilingual evaluation understudy (BLEU) between input and output and accuracy of the sentiment transfer the optimization of these two standard metrics diverge from the intuitive goal of the style transfer task. Finally, due to the nature of the task itself, there is a specific dependence between these two metrics that could be easily manipulated. Under these circumstances, we suggest taking BLEU between input and human-written reformulations into consideration for benchmarks. We also propose three new architectures that outperform state of the art in terms of this metric.

1 Introduction

Deep generative models attract a lot of attention in recent years (Hu et al., 2017b). Such methods as variational autoencoders (Kingma and Welling, 2013) or generative adversarial networks (Goodfellow et al., 2014) are successfully applied to a variety of machine vision problems including image generation (Radford et al., 2017), learning interpretable image representations (Chen et al., 2016) and style transfer for images (Gatys et al., 2016). However, natural language generation is more challenging due to many reasons, such as the discrete nature of textual information (Hu et al., 2017a), the absence of local information continuity and non-smooth disentangled representations (Bowman et al., 2015). Due to these difficulties, text generation is mostly limited to specific narrow applications and is usually working in supervised settings.

Content and style are deeply fused in natural language, but style transfer for texts is often addressed in the context of disentangled latent representations (Hu et al., 2017a; Shen et al., 2017; Fu et al., 2018; John et al., 2018; Romanov et al., 2018; Tian et al., 2018). Intuitive understanding of this problem is apparent: if an input text has some attribute A, a system generates new text similar to the input on a given set of attributes with only one attribute A changed to the target attribute $\tilde{A}$. In the majority of previous works, style transfer is obtained through an encoder-decoder architecture with one or multiple style discriminators to learn disentangled representations. The encoder takes a sentence as an input and generates a style-independent content representation. The decoder then takes the content representation and the target style representation to generate the transformed sentence. In (Subramanian et al., 2018) authors question the quality and usability of the disentangled representations for texts and suggest an end-to-end approach to style transfer similar to an end-to-end machine translation.

Contribution of this paper is three-fold: 1) we show that different style transfer architectures have varying results on test and that reporting error margins for various training re-runs of the same model is especially important for adequate assessment of the models accuracy, see Figure 1; 2) we show that BLEU (Papineni et al., 2002) between input and output and accuracy of style transfer measured in terms of the accuracy of a pre-trained external style classifier can be manipulated and naturally diverge from the intuitive goal of the style transfer task starting from a certain threshold; 3) new architectures that perform style transfer using improved latent representations are shown to outperform state of the art in terms of BLEU between output and

* Equal contribution

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3934–3943,
Hong Kong, China, November 3–7, 2019.

human-written reformulations.

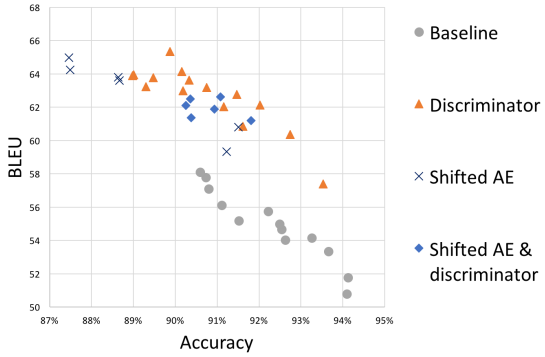

Figure 1: Test results of multiple runs for four different architectures retrained several times from scratch. In-depth description of the architectures can be found in Section 3.

2 Related Work

Style of a text is a very general notion that is hard to define in rigorous terms (Xu, 2017). However, the style of a text can be characterized quantitatively (Hughes et al., 2012); stylized texts could be generated if a system is trained on a dataset of stylistically similar texts (Potash et al., 2015); and author-style could be learned end-to-end (Tikhonov and Yamshchikov, 2018b,c; Vechtomova et al., 2018). A majority of recent works on style transfer focus on the sentiment of text and use it as a target attribute. For example, in (Li et al., 2018; Kabbara and Cheung, 2016; Xu et al., 2018) estimate the quality of the style transfer with binary sentiment classifier trained on the corpora further used for the training of the style-transfer system. (Ficler and Goldberg, 2017) and especially (Fu et al., 2018) generalize this ad-hoc approach defining a style as a set of arbitrary quantitively measurable categorial or continuous parameters. Such parameters could include the *'style of the time'* (Hughes et al., 2012), author-specific attributes (see (Xu et al., 2012) or (Jhamtani et al., 2017) on 'shakespearization'), politeness (Sennrich et al., 2016), formality of speech (Rao and Tetreault, 2018), and gender or even political slant (Prabhumoye et al., 2018).

A significant challenge associated with narrowly defined style-transfer problems is that finding a good solution for one aspect of a style does not guarantee that you can use the same solution for a different aspect of it. For example, Guu et al. (2018) build a generative model for sentiment transfer with a retrieve-edit approach. In (Li et al., 2018) a delete-retrieve model shows good results for sentiment transfer. However, it is hard to imagine that these retrieval approaches could be used, say, for the style of the time or formality, since in these cases the system is often expected to paraphrase a given sentence to achieve the target style.

In (Hu et al., 2017a) the authors propose a more general approach to the controlled text generation combining variational autoencoder (VAE) with an extended wake-sleep mechanism in which the sleep procedure updates both the generator and external classifier that assesses generated samples and feedbacks learning signals to the generator. Authors had concatenated labels for style with the text representation of the encoder and used this vector with "hard-coded" information about the sentiment of the output as the input of the decoder. This approach seems promising, and some other papers either extend it or use similar ideas. Shen et al. (2017) applied a GAN to align the hidden representations of sentences from two corpora using an adversarial loss to decompose information about the form. In (Zhao et al., 2017) model learns a smooth code space and can be used as a discrete GAN with the ability to generate coherent discrete outputs from continuous samples. Authors use two different generators for two different styles. In (Fu et al., 2018) an adversarial network is used to make sure that the output of the encoder does not have style representation. (Hu et al., 2017a) also uses an adversarial component that ensures there is no stylistic information within the representation. Fu et al. (2018) do not use a dedicated component that controls the semantic component of the latent representation. Such a component is proposed by John et al. (2018) who demonstrate that decomposition of style and content could be improved with an auxiliary multi-task for label prediction and adversarial objective for bag-of-words prediction. Romanov et al. (2018) also introduces a dedicated component to control semantic aspects of latent representations and an adversarial-motivational training that includes a special motivational loss to encourage a better decomposition. Speaking about preservation of semantics one also has to mention works on paraphrase systems, see, for example (Prakash et al., 2016; Gupta et al., 2018; Roy and Grangier, 2019). The methodology described in this paper could be extended to paraphrasing systems in terms of semantic preservation measurement, however,

this is the matter of future work.

Subramanian et al. (2018) state that learning a latent representation, which is independent of the attributes specifying its style, is rarely attainable. There are other works on style transfer that are based on the ideas of neural machine translation with (Carlson et al., 2018) and without parallel corpora (Zhang et al., 2018) in line with (Lample et al., 2017) and (Artetxe et al., 2017).

It is important to underline here that majority of the papers dedicated to style transfer for texts treat sentiment of a sentence as a stylistic rather than semantic attribute despite particular concerns (Tikhonov and Yamshchikov, 2018a). It is also crucial to mention that in line with (Fu et al., 2018) majority of the state of the art methods for style transfer use an external pre-trained classifier to measure the accuracy of the style transfer. BLEU computes the harmonic mean of precision of exact matching n-grams between a reference and a target sentence across the corpus. It is not sensitive to minute changes, but BLEU between input and output is often used as the coarse measure of the semantics preservation. For the corpora that have human written reformulations, BLEU between the output of the model and human text is used. These metrics are used alongside with a handful of others such as PINC (Paraphrase In N-gram Changes) score (Carlson et al., 2018), POS distance (Tian et al., 2018), language fluency (John et al., 2018), etc. Figure 2 shows self-reported results of different models in terms of two most frequently measured performance metrics, namely, BLEU and Accuracy of the style transfer.

This paper focuses on Yelp! reviews dataset[1] that was lately enhanced with human written reformulations by (Li et al., 2018). These are Yelp! reviews, where each short English review of a place is labeled as a negative or as a positive once. This paper studies three metrics that are most common in the field at the moment and questions to which extent can they be used for the performance assessment. These metrics are the accuracy of an external style classifier that is trained to measure the accuracy of the style transfer, BLEU between input and output of a system, and BLEU between output and human-written texts.

[1]https://www.yelp.com/dataset

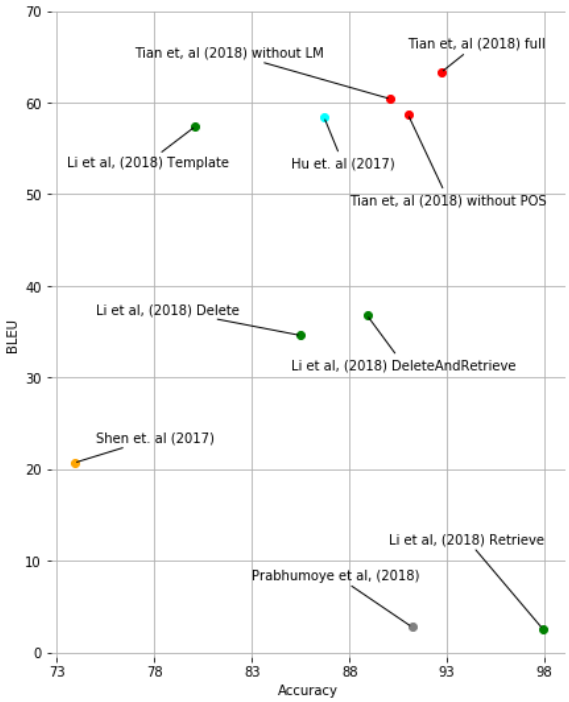

Figure 2: Overview of the self-reported results for sentiment transfer on Yelp! reviews. Results of (Romanov et al., 2018) are not displayed due to the absence of self-reported BLEU scores. Later in the paper we show that on different reruns BLEU and accuracy can vary from these self-reported single results.

3 Style transfer

In this work we experiment with extensions of a model, described in (Hu et al., 2017a), using Texar (Hu et al., 2018) framework. To generate plausible sentences with specific semantic and stylistic features every sentence is conditioned on a representation vector z which is concatenated with a particular code c that specifies desired attribute, see Figure 3. Under notation introduced in (Hu et al., 2017a) the base autoencoder (AE) includes a conditional probabilistic encoder E defined with parameters θ_E to infer the latent representation z given input x

$$z \sim E(x) = q_E(z, c|x).$$

Generator G defined with parameters θ_G is a GRU-RNN for generating and output $\hat{x}$ defined as a sequence of tokens $\hat{x} = \hat{x}_1, ..., \hat{x}_T$ conditioned on the latent representation z and a stylistic component c that are concatenated and give rise to a generative distribution

$$\hat{x} \sim G(z, c) = p_G(\hat{x}|z, c).$$

These encoder and generator form an AE with the following loss

$$\mathcal{L}_{ae}(\theta_G, \theta_E; x, c) = -\mathbb{E}_{q_E(z,c|x)} \left[\log q_G(x|z,c)\right]. \quad (1)$$

This standard reconstruction loss that drives the generator to produce realistic sentences is combined with two additional losses. The first discriminator provides extra learning signals which enforce the generator to produce coherent attributes that match the structured code in c. Since it is impossible to propagate gradients from the discriminator through the discrete sample $\hat{x}$, we use a deterministic continuous approximation a "soft" generated sentence, denoted as $\tilde{G} = \tilde{G}_\tau(z, c)$ with "temperature" τ set to $\tau \to 0$ as training proceeds. The resulting soft generated sentence is fed into the discriminator to measure the fitness to the target attribute, leading to the following loss

$$\mathcal{L}_c(\theta_G, \theta_E; x) = -\mathbb{E}_{q_E(z,c|x)} \left[\log q_D(c|\tilde{G})\right]. \quad (2)$$

Finally, under the assumption that each structured attribute of generated sentences is controlled through the corresponding code in c and is independent from z one would like to control that other not explicitly modelled attributes do not entangle with c. This is addressed by the dedicated loss

$$\mathcal{L}_z(\theta_G; x) = -\mathbb{E}_{q_E(z,c|x)q_D(c|x)} \left[\log q_E(z|\tilde{G})\right]. \quad (3)$$

The training objective for the baseline, shown in Figure 3, is therefore a sum of the losses from Equations (1) – (3) defined as

$$min_{\theta_G} \mathcal{L}_{baseline} = \mathcal{L}_{ae} + \lambda_c \mathcal{L}_c + \lambda_z \mathcal{L}_z, \quad (4)$$

where λ_c and λ_z are balancing parameters.

Let us propose two further extensions of this baseline architecture. To improve reproducibility of the research the code of the studied models is open[2]. Both extensions aim to improve the quality of information decomposition within the latent representation. In the first one, shown in Figure 4, a special dedicated discriminator is added to the model to control that the latent representation does not contain stylistic information. The loss of this discriminator is defined as

$$\mathcal{L}_{D_z}(\theta_G; x, c) = -\mathbb{E}_{q_E(z|x)} \left[\log q_{D_z}(c|z)\right]. \quad (5)$$

[2]https://github.com/VAShibaev/text_style_transfer

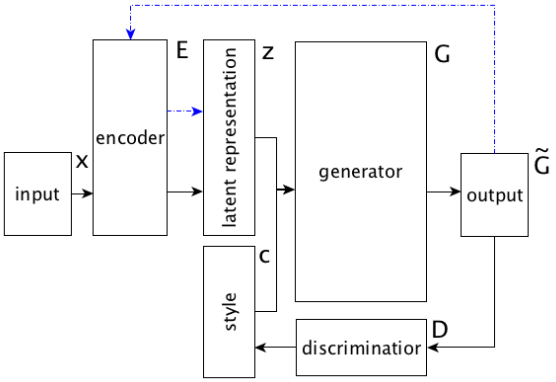

Figure 3: The generative model, where style is a structured code targeting sentence attributes to control. Blue dashed arrows denote the proposed independence constraint of latent representation and controlled attribute, see (Hu et al., 2017a) for the details.

Here a discriminator denoted as D_z is trying to predict code c using representation z. Combining the loss defined by Equation (4) with the adversarial component defined in Equation (5) the following learning objective is formed

$$min_{\theta_G} \mathcal{L} = \mathcal{L}_{baseline} - \lambda_{D_z} \mathcal{L}_{Dz}, \quad (6)$$

where $\mathcal{L}_{baseline}$ is a sum defined in Equation (4), λ_{D_z} is a balancing parameter.

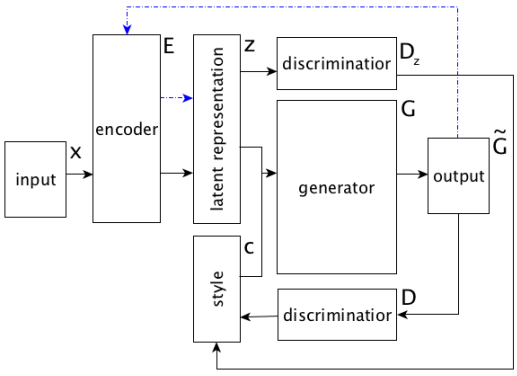

Figure 4: The generative model with dedicated discriminator introduced to ensure that semantic part of the latent representation does not have information on the style of the text.

The second extension of the baseline architecture does not use an adversarial component D_z that is trying to eradicate information on c from component z. Instead, the system, shown in Figure 5 feeds the "soft" generated sentence $\tilde{G}$ into encoder E and checks how close is the representation $E(\tilde{G})$

to the original representation $z = E(x)$ in terms of the cosine distance. We further refer to it as *shifted autoencoder* or SAE. Ideally, both $E(\tilde{G}(E(x), c))$ and $E(\tilde{G}(E(x), \bar{c}))$, where $\bar{c}$ denotes an inverse style code, should be both equal to $E(x)$[3]. The loss of the shifted autoencoder is

$$min_{\theta_G}\mathcal{L} = \mathcal{L}_{baseline} + \lambda_{cos}\mathcal{L}_{cos} + \lambda_{cos^-}\mathcal{L}_{cos^-}, \tag{7}$$

where λ_{cos} and λ_{cos^-} are two balancing parameters, with two additional terms in the loss, namely, cosine distances between the softened output processed by the encoder and the encoded original input, defined as

$$\mathcal{L}_{cos}(x, c) = \cos\left(E(\tilde{G}(E(x), c)), E(x)\right),$$
$$\mathcal{L}_{cos^-}(x, c) = \cos\left(E(\tilde{G}(E(x), \bar{c})), E(x)\right). \tag{8}$$

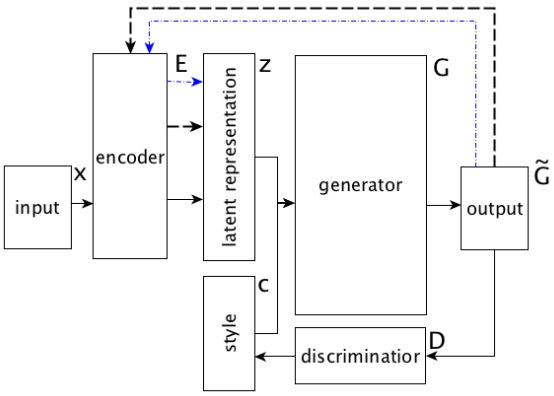

Figure 5: The generative model with a dedicated loss added to control that semantic representation of the output, when processed by the encoder, is close to the semantic representation of the input.

We also study a combination of both approaches described above, shown on Figure 6.

In Section 4 we describe a series of experiments that we have carried out for these architectures using Yelp! reviews dataset.

4 Experiments

We have found that the baseline, as well as the proposed extensions, have noisy outcomes, when retrained from scratch, see Figure 1. Most of the papers mentioned in Section 2 measure the performance of the methods proposed for the sentiment transfer with two metrics: accuracy of the external sentiment classifier measured on test data, and BLEU between the input and output that is regarded as a coarse metric for semantic similarity.

[3]This notation is valid under the assumption that every stylistic attribute is a binary feature

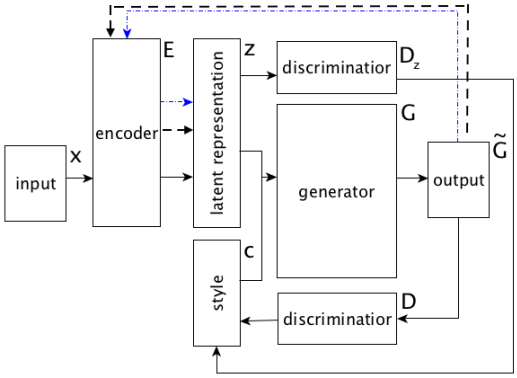

Figure 6: A combination of an additional discriminator used in Figure 4 with a shifted autoencoder shown in Figure 5

In the first part of this section, we demonstrate that reporting error margins is essential for the performance assessment in terms that are prevalent in the field at the moment, i.e., BLEU between input and output and accuracy of the external sentiment classifier. In the second part, we also show that both of these two metrics after a certain threshold start to diverge from an intuitive goal of the style transfer and could be manipulated.

4.1 Error margins matter

On Figure 1 one can see that the outcomes for every single rerun differ significantly. Namely, accuracy can change up to 5 percentage points, whereas BLEU can vary up to 8 points. This variance can be partially explained with the stochasticity incurred due to sampling from the latent variables. However, we show that results for state of the art models sometimes end up within error margins from one another, so one has to report the margins to compare the results rigorously. More importantly, one can see that there is an inherent trade-off between these two performance metrics. This trade-off is not only visible across models but is also present for the same retrained architecture. Therefore, improving one of the two metrics is not enough to confidently state that one system solves the style-transfer problem better than the other. One has to report error margins after several consecutive retrains and instead of comparing one of the two metrics has to talk about Pareto-like optimization that would show confident improvement of both.

To put obtained results into perspective, we have retrained every model from scratch five times in a row. We have also retrained the models of Tian et al. (2018) five times since their code is published online. Figure 7 shows the results of all models with error margins. It is also enhanced with other self-reported results on the same Yelp! review dataset for which no code was published.

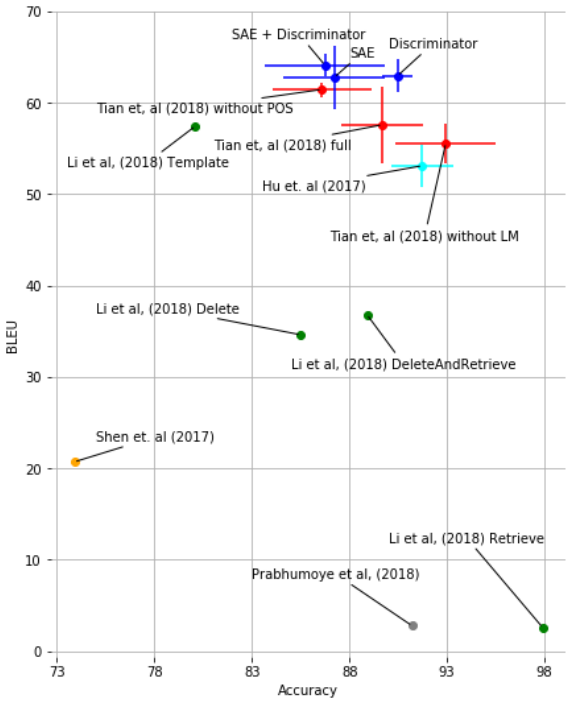

Figure 7: Overview of the self-reported results for sentiment transfer on Yelp! reviews alongside with the results for the baseline model (Hu et al., 2017a), architecture with additional discriminator, shifted autoencoder (SAE) with additional cosine losses, and a combination of these two architectures averaged after five re-trains alongside with architectures proposed by (Tian et al., 2018) after five consecutive re-trains. Results of (Romanov et al., 2018) are not displayed due to the absence of self-reported BLEU scores.

One can see that error margins of the models, for which several reruns could be performed, overlap significantly. In the next subsection, we carefully study BLEU and accuracy of the external classifier and discuss their aptness to measure style transfer performance.

4.2 Delete, duplicate and conquer

One can argue that as there is an inevitable entanglement between semantics and stylistics in natural language, there is also an apparent entanglement between BLEU of input and output and accuracy estimation of the style. Indeed, the output that copies input gives maximal BLEU yet clearly fails in terms of the style transfer. On the other hand, a wholly rephrased sentence could provide a low BLEU between input and output but high accuracy. These two issues are not problematic when both BLEU between input and output and accuracy of the transfer are relatively low. However, since style transfer methods have significantly evolved in recent years, some state of the art methods are now sensitive to these issues. The trade-off between these two metrics can be seen in Figure 1 as well as in Figure 7.

As we have mentioned above, the accuracy of an external classifier and BLEU between output and input are the most widely used methods to assess the performance of style transfer at this moment. However, both of these metrics can be manipulated in a relatively simple manner. One can extend the generative architecture with internal pre-trained classifier of style and then perform the following heuristic procedure:

- measure the style accuracy on the output for a given batch;
- choose the sentences that style classifier labels as incorrect;
- replace them with duplicates of sentences from the given batch that have correct style according to the internal classifier and show the highest BLEU with given inputs.

This way One can replace all sentences that push measured accuracy down and boost reported accuracy to 100%. To see the effect that this manipulation has on the key performance metric we split all sentences with wrong style in 10 groups of equal size and replaces them with the best possible duplicates of the stylistically correct sentences group after group. The results of this process are shown in Figure 8.

This result is disconcerting. Simply replacing part of the output with duplicates of the sentences that happen to have relatively high BLEU with given inputs allows to "boost" accuracy to 100% and "improve" BLEU. The change of BLEU during such manipulation stays within error margins of the architecture, but accuracy is significantly manipulated. What is even more disturbing is that BLEU between such manipulated output of the batch and human-written reformulations provided in (Tian

et al., 2018) also grows. Figure 8 shows that for SAE but all four architectures described in Section 3 demonstrate similar behavior.

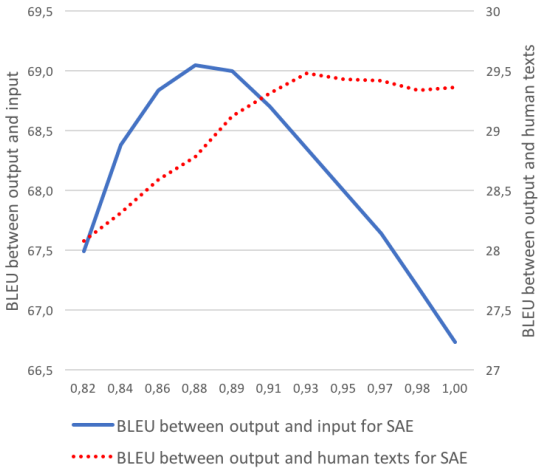

Figure 8: Manipulating the generated output in a way that boosts accuracy one can change BLEU between output and input. Moreover, such manipulation increases BLEU between output and human written reformulations. The picture shows behavior of SAE, but other architectures demonstrate similar behavior. The results are an average of four consecutive retrains of the same architecture.

Our experiments show that though we can manipulate BLEU between output and human-written text, it tends to change monotonically. That might be because of the fact that this metric incorporates information on stylistics and semantics of the text at the same time, preserving inevitable entanglement that we have mentioned earlier. Despite being costly, human-written reformulations are needed for future experiments with style transfer. It seems that modern architectures have reached a certain level of complexity for which naive proxy metrics such as accuracy of an external classifier or BLEU between output and input are already not enough for performance estimation and should be combined with BLEU between output and human-written texts. As the quality of style transfer grows further one has to improve the human-written data sets: for example, one would like to have data sets similar to the ones used for machine translation with several reformulations of the same sentence.

On Figure 9 one can see how new proposed architectures compare with another state of the art approaches in terms of BLEU between output and human-written reformulations.

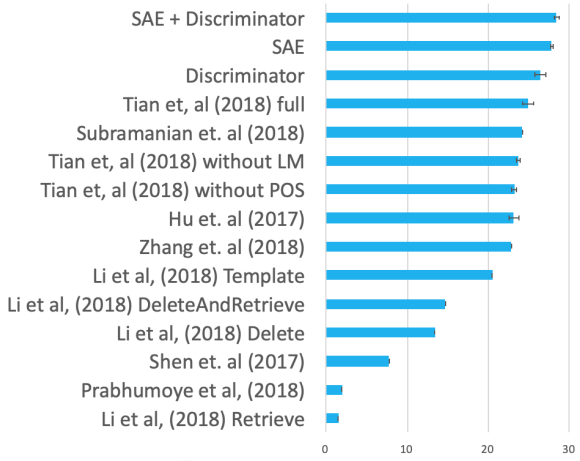

Figure 9: Overview of the BLEU between output and human-written reformulations of Yelp! reviews. Architecture with additional discriminator, shifted autoencoder (SAE) with additional cosine losses, and a combination of these two architectures measured after five re-runs outperform the baseline by (Hu et al., 2017a) as well as other state of the art models. Results of (Romanov et al., 2018) are not displayed due to the absence of self-reported BLEU scores

5 Conclusion

Style transfer is not a rigorously defined NLP problem. Starting from definitions of style and semantics and finishing with metrics that could be used to evaluate the performance of a proposed system. There is a surge of recent contributions that work on this problem. This paper highlights several issues connected with this lack of rigor. First, it shows that the state of the art algorithms are inherently noisy on the two most widely accepted metrics, namely, BLEU between input and output and accuracy of the external style classifier. This noise can be partially attributed to the adversarial components that are often used in the state of the art architectures and partly due to certain methodological inconsistencies in the assessment of the performance. Second, it shows that reporting error margins of several consecutive retrains for the same model is crucial for the comparison of different architectures, since error margins for some of the models overlap significantly. Finally, it demonstrates that even BLEU on human-written reformulations can be manipulated in a relatively simple way.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. In *arXiv preprint*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society open science*, 5(10):171920.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, pages 2172–2180.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, volume 94-104.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. *AAAI*.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural network. *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference*, pages 2414–2423.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, pages 2672–2680.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association of Computational Linguistics*, 6:437–450.

Zhiting Hu, Haoran Shi, Zichao Yang, Bowen Tan, Tiancheng Zhao, Junxian He, Wentao Wang, Xingjiang Yu, Lianhui Qin, Di Wang, Xuezhe Ma, Hector Liu, Xiaodan Liang, Wanrong Zhu, Devendra Singh Sachan, and Eric P. Xing. 2018. Texar: A modularized, versatile, and extensible toolkit for text generation. In *arXiv preprint*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017a. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596.

Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric Xing. 2017b. On unifying deep generative models. In *arXiv preprint*.

James M. Hughes, Nicholas J. Foti, David C. Krakauer, and Daniel N. Rockmore. 2012. Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences*, 109(20):7682–7686.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2018. Disentangled representation learning for text style transfer. In *arXiv preprint*.

Jad Kabbara and Jackie Chi Kit Cheung. 2016. Stylistic transfer in natural language generation systems using recurrent neural networks. *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 43–47.

Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *arXiv preprint*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. In *arXiv preprint*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. 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*, volume 1, pages 865–1874.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Gbleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2015. Ghostwriter: Using an lstm for automatic rap lyric generation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1919–1924. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Alan W. Black, and Ruslan Salakhutdinov. 2018. Style transfer through back-translation. In *arXiv preprint*.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual lstm networks. In *arXiv preprint*.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. In *arXiv preprint*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 129–140.

Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2018. Adversarial decomposition of text representation. In *arXiv preprint*.

Aurko Roy and David Grangier. 2019. Unsupervised paraphrasing without translation. In *arXiv preprint*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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*, pages 35–40.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *31st Conference on Neural Information Processing Systems*, pages 6833–6844.

Sandeep Subramanian, Guillaume Lample, Eric M. Smith, Ludovic Denoyer, Marc Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. In *arXiv preprint*.

Youzhi Tian, Zhiting Hu, and Zhou Yu. 2018. Structured content preservation for unsupervised text style transfer. In *arXiv preprint*.

Alexe Tikhonov and Ivan P. Yamshchikov. 2018a. What is wrong with style transfer for texts? In *arXiv preprint*.

Alexey Tikhonov and Ivan P. Yamshchikov. 2018b. Guess who? Multilingual approach for the automated generation of author-stylized poetry. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 787–794.

Alexey Tikhonov and Ivan P. Yamshchikov. 2018c. Sounds Wilde. Phonetically extended embeddings for author-stylized poetry generation. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–124.

Olga Vechtomova, Hareesh Bahuleyan, Amirpasha Ghabussi, and Vineet John. 2018. Generating lyrics with variational autoencoder and multi-modal artist embeddings. In *arXiv preprint*.

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. In *arXiv preprint*.

Wei Xu. 2017. From shakespeare to twitter: What are language styles all about? *Proceedings of the Workshop on Stylistic Variation*, pages 1–9.

Wei Xu, Alan Ritter, William B. Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. *Proceedings of COLING*, pages 2899–2914.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. In *arXiv preprint*.

Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2017. Adversarially regularized autoencoders for generating discrete structures. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. CoRR.

A Supplemental Material

Here are some examples characteristic for different systems. An output of a system follows the input. Here are some successful examples produced by the system with additional discriminator:

- it's not much like an actual irish pub, which is depressing. → it's definitely much like an actual irish pub, which is grateful.
- i got a bagel breakfast sandwich and it was delicious! → i got a bagel breakfast sandwich and it was disgusting!
- i love their flavored coffee. → i dumb their flavored coffee.
- i got a bagel breakfast sandwich and it was delicious! → i got a bagel breakfast sandwich and it was disgusting!
- i love their flavored coffee. → i dumb their flavored coffee.
- nice selection of games to play. → typical selection of games to play.
- i'm not a fan of huge chain restaurants. → i'm definitely a fan of huge chain restaurants.

Here are some examples of typical faulty reformulations:

- only now i'm really hungry, and really pissed off. → kids now i'm really hungry, and really extraordinary off.
- what a waste of my time and theirs. → what a wow. of my time and theirs.
- cooked to perfection and very flavorful. → cooked to pain and very outdated.

- the beer was nice and cold! → the beer was nice and consistant!
- corn bread was also good! → corn bread was also unethical bagged

Here are some successful examples produced by the SAE:

- our waitress was the best, very accommodating. → our waitress was the worst, very accommodating.
- great food and awesome service! → horrible food and nasty service!
- their sandwiches were really tasty. → their sandwiches were really bland.
- i highly recommend the ahi tuna. → i highly hated the ahi tuna.
- other than that, it's great! → other than that, it's horrible!

Here are some examples of typical faulty reformulations by SAE:

- good drinks, and good company. → 9:30 drinks, and 9:30 company.
- like it's been in a fridge for a week. → like it's been in a fridge for a true.
- save your money & your patience. → save your smile & your patience.
- no call, no nothing. → deliciously call, deliciously community.
- sounds good doesn't it? → sounds good does keeps it talented

Here are some successful examples produced by the SAE with additional discriminator:

- best green corn tamales around. → worst green corn tamales around.
- she did the most amazing job. → she did the most desperate job.
- very friendly staff and manager. → very inconsistent staff and manager.
- even the water tasted horrible. → even the water tasted great.
- go here, you will love it. → go here, you will avoid it.

Here are some examples of typical faulty reformulations by the SAE with additional discriminator:

- _num_ - _num_ % capacity at most , i was the only one in the pool. → sweetness - stylish % fountains at most, i was the new one in the
- this is pretty darn good pizza! → this is pretty darn unsafe pizza misleading
- enjoyed the dolly a lot. → remove the shortage a lot.
- so, it went in the trash. → so, it improved in the hooked.
- they are so fresh and yummy. → they are so bland and yummy.