Decomposing Textual Information For Style Transfer

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

Viacheslav Shibaev
Ural Federal University
Ekaterinburg, Russia

Aleksander Nagaev
Ural Federal University
Ekaterinburg, Russia

Jürgen Jost
Max Planck Institute
for Mathematics in the Sciences
Leipzig, Germany

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

Abstract

This paper focuses on latent representations that could effectively decompose different aspects of textual information. Using a framework of style transfer for texts, we propose several empirical methods to assess information decomposition quality. We validate these methods with several state-of-the-art textual style transfer methods. Higher quality of information decomposition corresponds to higher performance in terms of bilingual evaluation understudy (BLEU) between output and human-written reformulations.

1 Introduction

The arrival of deep learning seems transformative for many areas of information processing and is especially interesting for generative models (Hu et al., 2017b). However, natural language generation is still a challenging task due to a number of factors that include the absence of local information continuity and non-smooth disentangled representations (Bowman et al., 2015), and discrete nature of textual information (Hu et al., 2017a). If information needed for different natural language processing (NLP) tasks could be encapsulated in independent components of the obtained latent representations, one could have worked with different aspects of text independently. This could also naturally simplify learning transfer for NLP models and potentially make them more interpretable.

Despite the fact that content and style are deeply fused in natural language, 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). A majority of 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. This approach seems intuitive and appealing but has certain difficulties. For example, Subramanian et al. (2018) question the quality and usability of the disentangled representations for texts with an elegant experiment. The authors train a state of the art architecture that relies on disentangled representations and show that an external artificial neural network can predict the style of the input using a semantic component of an obtained latent representation (that supposedly did not incorporate stylistic information).

In this work, we demonstrate that the decomposition of latent representations is, indeed, attainable with encoder-decoder based methods but depends on the used architecture. Moreover, architectures with higher quality of information decomposition perform better in terms of the style transfer task.

The contribution of this paper is threefold: (1) we propose several ways to quantify the quality of the obtained latent semantic representations; (2) we show that the quality of such representation can significantly differ depending on the used architecture; (3) finally we demonstrate that architectures with higher quality of information decomposition perform better in terms of BLEU (Papineni et al., 2002) between output of a model and a human written reformulations.

2 Related Work

It is hard to define style transfer rigorously (Xu, 2017). Therefore recent contributions in the field are mostly motivated by several empirical results and rather address specific narrow aspects of style

---

[*] Equal contribution

that could be empirically measured. Stylistic attributes of text include author-specific attributes (see (Xu et al., 2012) or (Jhamtani et al., 2017) on 'shakespearization'), politeness (Sennrich et al., 2016), the *'style of the time'* (Hughes et al., 2012), gender or political slant (Prabhumoye et al., 2018), and formality of speech (Rao and Tetreault, 2018). All these attributes are defined with varying degrees of rigor. Meanwhile, the general notion of literally style is only addressed in a very broad context. For example, Hughes et al. (2012) shows that the style of a text can be characterized quantitively and not only with an expert opinion; Potash et al. (2015) demonstrate that stylized texts could be generated if a system is trained on a dataset of stylistically similar texts; and literary styles of the authors could be learned end-to-end (Tikhonov and Yamshchikov, 2018a,b; Vechtomova et al., 2018).

In this particular submission we focus on a very narrow framework of sentiment transfer. There is certain controversy whether sentiment of a text could be regarded as its stylistic attribute, see (Tikhonov and Yamshchikov, 2018c). However, there seems to be certain agreement in the field that sentiment could be regarded as a viable attribute to be changed by the style transfer system. Addressing the problem of sentiment transfer Kabbara and Cheung (2016); Li et al. (2018); Xu et al. (2018) estimate the quality of the style transfer with a pre-trained binary sentiment classifier. Fu et al. (2018) and Ficler and Goldberg (2017) generalize this ad-hoc approach and in principle enable the information decomposition approach. They define a style as a set of arbitrary quantitatively measurable categorical or continuous parameters that could be automatically estimated with an external classifier. In this submission we stay within this empirical paradigm of literary style.

Generally speaking, a solution that works for one aspect of a style could not be applied for a different aspect of it. For example, a retrieve-edit approach by (Guu et al., 2018) works for sentiment transfer. A delete-retrieve model shows good results for sentiment transfer in (Li et al., 2018). However, these retrieval approaches could hardly be used for the style of the time or formality or any other case when the system is expected to paraphrase a given sentence to achieve the target style. To address this challenge Hu et al. (2017a) 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 discriminator that assesses generated samples and feedbacks learning signals to the generator. Labels for style were concatenated with the text representation of the encoder and used with "hard-coded" information about the sentiment of the output as the input of the decoder. This approach is promising and is used in many recent contributions. Shen et al. (2017) use an adversarial loss to decompose information about the form of a sentence and apply a GAN to align hidden representations of sentences from two corpora. Fu et al. (2018) use an adversarial network to make sure that the output of the encoder does not include stylistic information. Hu et al. (2017a) also use an adversarial component to ensure there is no stylistic information within the representation. A dedicated component that controls semantic component of the latent representation 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 a bag-of-words prediction. Romanov et al. (2018) also introduce 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.

The framework of information decomposition within latent representations is challenged by an alternative family of neural machine translation approaches. These are works on style transfer 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). In particular, Subramanian et al. (2018) state that learning a latent representation, which is independent of the attributes specifying its style is rarely attainable. They experiment with the model developed in (Fu et al., 2018) where by design the discriminator, which was trained adversarially and jointly with the model, gets worse at predicting the sentiment of the input when the coefficient of the adversarial loss increases. Authors show that a classifier that is separately trained on the resulting encoder representations easily recovers the sentiment of a latent representation produced by the encoder.

In this paper, we show that contrary to (Subramanian et al., 2018) decomposition of the stylistic and semantic information is attainable with

autoencoder-type models and could be quantified. However, the quality of such decomposition severely depends on the particular architecture. We propose three different measures for information decomposition quality and using four different architectures show that models with better information decomposition outperform the state-of-the-art models in terms of BLEU between output and human-written reformulations.

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 1. 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 1, 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.
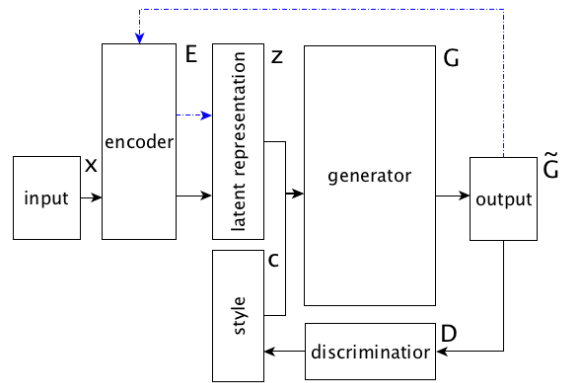

Figure 1: The generative model, where style is a structured code targeting sentence attributes to control. Blue dashed arrows denote the proposed independence constraint of latent representation and controlled attribute, see (Hu et al., 2017a) for the details.

Let us propose two further extensions of this baseline architecture. To improve reproducibility of the research the code of the studied models is open[1]. Both extensions aim to improve the quality of information decomposition within the latent representation. In the first one, shown in Figure 2, 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

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

---

[1]https://github.com/VAShibaev/textstyletransfer

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

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

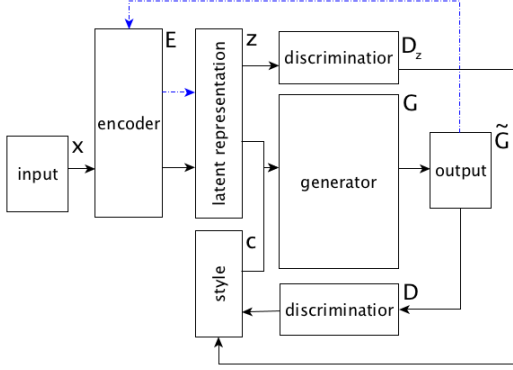where $\mathcal{L}_{baseline}$ is a sum defined in Equation (4), λ_{D_z} is a balancing parameter.

Figure 2: 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 3 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 $E(x)^2$. The loss of the shifted autoencoder is

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

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

$$\mathcal{L}_{cos}(x, c) = \cos\left(E(\tilde{G}(E(x), c)), E(x)\right),$$

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

---

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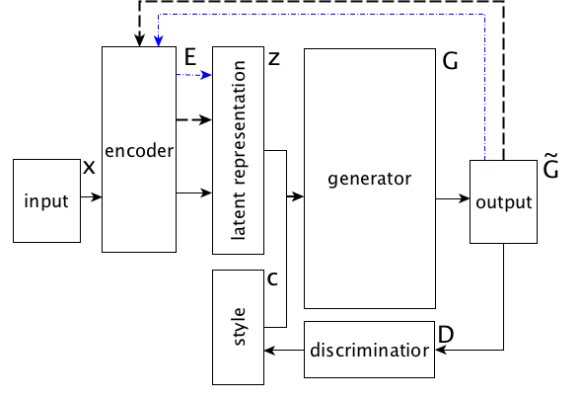

Figure 3: The generative model with a dedicated loss added to control that semantic representation of the output, when processed by the encoder, is close to the semantic representation of the input.

We also study a combination of both approaches described above, shown on Figure 4.
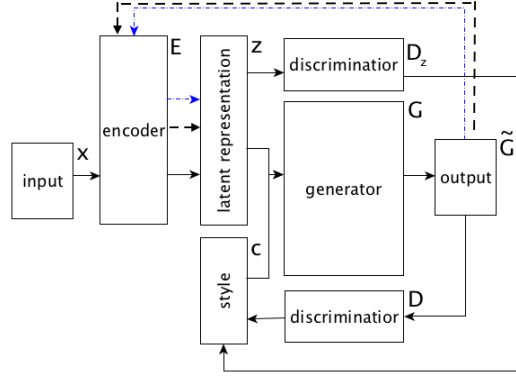

Figure 4: A combination of an additional discriminator used in Figure 2 with a shifted autoencoder shown in Figure 3

Tikhonov et al. (2019) carry out a series of experiments for these architectures. In this contribution, we work with the same data set of human-labeled positive and negative reviews but focus solely on the quality of information decomposition.

4 Information decomposition for texts

As we have mentioned earlier, several recent contributions rely on the idea that decomposing different aspects of textual information into various components of a latent representation might be helpful for a task of style transfer. To our knowledge, this is a supposition that is rarely addressed rigorously. The majority of the arguments in favor of information decomposition based architectures is of an intu-

itive and qualitative rather than quantitative nature. Moreover, there are specific arguments against this idea.

In particular, Subramanian et al. (2018) show that information decomposition does not necessarily occur in autoencoder-based systems using a method developed in (Fu et al., 2018). Subramanian et al. (2018) demonstrate that as training proceeds, the internal discriminator, which was trained adversarially and jointly with the model, gets worse at predicting the sentiment of the input. However, an external classifier that is separately trained on the resulting latent representations easily recovers the sentiment. This is a strong argument in favor of the idea that actual disentanglement does not happen. Instead of decomposing the semantic and stylistic aspects of information, the encoder merely 'tricks' internal classifier and 'hides' stylistic information in the semantic component ending up in some local optimum.

4.1 Empirical measure of information decomposition quality

Yelp![3] reviews dataset that was lately enhanced with human written reformulations by (Tian et al., 2018) is one of the most frequently used baselines for textual style transfer at the moment. It consists of restaurant reviews split into two categories, namely, positive and negative. There is a human written reformulation of every review in which the sentiment is changed that is commonly used as a ground truth for the task performance estimation.

We applied an empirical method to estimate the quality of information decomposition to the architectures described in Section 3 as well as architectures developed by (Tian et al., 2018). An external classifier was trained from scratch to predict a style of a message using component z of a latent representation produced by an encoder. If information decomposition does not happen, one would expect that accuracy of an external classifier would be close to 1. This would mean that despite intuitive expectations, information about the style of a message is present in z. If decomposition were effective, the accuracy of an external classifier would be close to 0.5; in (Tikhonov et al., 2019) it is shown that style transfer methods show varying results in terms of accuracy and BLEU for different retrains, so in this paper the accuracy of an external classifier and BLEU between the system's output and

human-written reformulations was measured after four independent retrains. On Figure 5, one can see the results of these experiments.
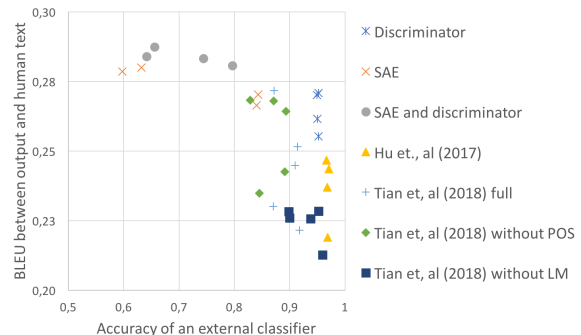

Figure 5: BLEU between system's output and human-written reformulations seems to be higher if accuracy of an external classifier is closer to one half. Systems that decompose information better tend to show higher BLEU.

The fact that the external classifier always predicts style with the probability that is above one half could be partially attributed to the fact that full information decomposition of sentiment and semantics is hardly attainable. For example, such adjectives as "delicious" or "yummy" incorporate positive sentiment with the semantics of taste, whereas "polite" or "friendly" in Yelp! reviews are combining positive sentiment with the semantics of service. This internal entanglement of sentiment and semantics is discussed in detail in (Tikhonov and Yamshchikov, 2018c). It is essential to mention that the very fact that semantics and stylistics are entangled on the level of words does not deny a theoretical possibility to build a latent representation where they are fully disentangled. Anyway, Figure 5 demonstrates that the quality of the disentanglement is much better for SAE-type architectures. Since the shifted autoencoder controls the cosine distance between soft output and input, the encoder has to disentangle the semantic component, rather than "hide" the sentiment information from the discriminator.

On Figure 6 one can see how state of the art approaches compare to each other in terms of BLEU between output and human-written reformulations. All systems were retrained five times from scratch to report error margins of the methods since the results are noisy. BLEU between output and human-written reformulations is higher for lower values of

---

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

external classifier accuracy. Systems that perform better in terms of information decomposition outperform system with lower quality of information decomposition. Moreover, the system that does not rely on an idea of disentangled latent representations at all shows weaker results than systems with high information disentanglement. It is important to note that there is a variety of methods to assess the quality of style transfer 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. The methodology of style transfer quality assessment is addressed in detail in (Tikhonov et al., 2019), but BLEU between output and input is a very natural all-purpose metric for the task of such type that is common in the style transfer literature.
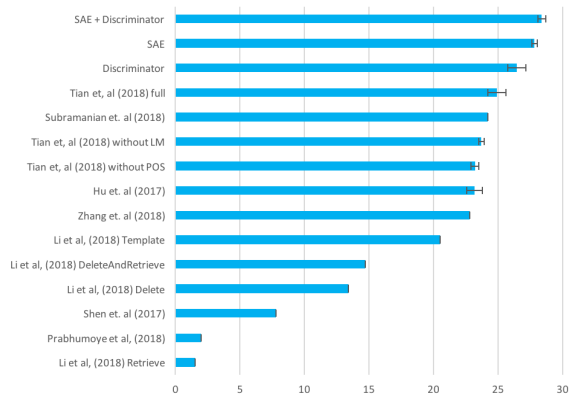

Figure 6: Overview of the BLEU between output and human-written reformulations of Yelp! reviews. Architecture with additional discriminator, shifted autoencoder (SAE) with additional cosine losses, and a combination of these two architectures measured after five re-runs outperform the baseline by (Hu 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

Tables 1 - 2 allow to compare random examples for different architectures. Generally, baseline and discriminator perform poorly once the syntax of a review is irregular or if there are some omissions in the text. SAE-based architectures tend to preserve the semantic component better. They also add sentimentally charged words at random not as often as the baseline and the discriminator-based architecture.

4.2 Preservation of semantic component

Another way to quantify the quality of latent representations is to calculate cosine distance and KL-divergence between semantic components of latent representations for the inputs and corresponding outputs. If we believe that the latent representation captures the semantics of the input that should be preserved in the output, the ideal behavior of the system is to produce equal latent representation for both the input and the output phrase. Indeed, on Figure 7 one can see that SAE manages to learn a space of latent representations in which semantic components of inputs and outputs are always equal to each other. Architecture with additional stylistic discriminator shows lower cosine distances and lower KL-divergences then the baseline yet. This results are in line with the measurements discussed above in Section 4.1.
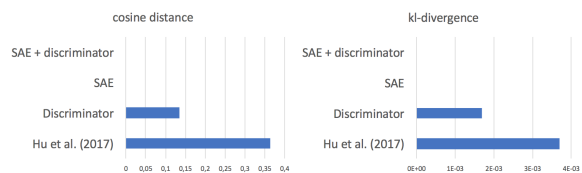

Figure 7: Comparison of cosine distances and KL-divergences between semantic components of latent representation for inputs and outputs. After 12 epochs of training SAE makes semantic component z for every output equal to the semantic component for a corresponding input. Discriminator corresponds to lower values of KL-divergence and cosine distance then baseline (Hu et al., 2017a)

To get an intuition on how the resulting latent space differs for different architectures, one can look at the t-SNE visualizations (Maaten and Hinton, 2008) for the resulting latent representations of the data that different systems produce. In Figure 8, one can see that the baseline latent representations easily allow recovering the sentiment.

In contrast with the baseline, the architecture with additional discriminator obtains better disentanglement. Figure 9 shows that in this case one has a harder time recovering the sentiment of the sentence based on its latent representation.

SAE does not only show a higher level of disentanglement but also produces equal semantic components for the input and the corresponding output. Judging by Figure 10 this makes SAE representations denser in certain areas of the semantic space and sparser in the others.

input	Human	baseline
the carne asada burrito is awesome!	the carne asada burrito is awful!	the worst asada burrito is gross!
the rooms are not that nice and the food is not that good either.	the rooms were spacious and food was very well cooked	the rooms are excellent that nice and the food is not that good either.
it was so delicious; i've never had anything like it!	everything tasted bad, nothing i liked	it was so rude; i've never had anything like it!
so, that was my one and only time ordering the benedict there.	i will be ordering the benedict again very good meal!	so, that was my one and best time ordering the perfect there.
you'll see why once you get there.	you'll see why i don't like it once you get there.	you'll see why once you get there.
i wanted to like this place but it just became a big disappointment.	a place like this is a great value	i helped to like this place it just became a big hidden.
and i had my sugar bowl favorite, the top hat sundae!	my sugar bowl favorite was not in stock.	and i had my sugar bowl worst, the lackluster hat gross!
um... we just told him that we didn't want to finance.	um... we just told him, sure we'd go ahead and finance!	amazing... we just told him that we did definitely want to open.
definitely a place to keep in mind.	not a place i would recommend	disappointing a place to keep in mind.
firstly, their fees are generally higher than other places.	the fees are comparable to other places.	best, their fees are generally higher than other places.
love the afternoon - tea at the phoenician.	hate the aternoon tea at the phoenician	absurd the inappropriate - tea at the insult.

Table 1: Several random input lines alongside with human written reformulation and the reformulation generated by the baseline.

Discriminator	SAE	SAE + Discriminator
the carne asada burrito is absurd!	the carne asada burrito is worst!	the carne asada burrito is sub-par!
the rooms are delicious that nice and the food is delicious that good either.	the rooms are definitely that nice and the food is definitely that good either.	the rooms are consistanly that nice and the food is consistanly that good either.
it was so not; i've never had anything like it!	it was so disgusting; i've never had anything like it!	it was so angry ; i've never had anything like it!
so, that was my one and fam time solid the benedict perfectly.	so, that was my one and kids time ordering the benedict there.	so, that was my one and always time ordering the benedict there.
you'll trash why once you get there.	you'll avoid why once you get there.	you'll see why once you get there.
i wanted to like this place but it just delightful a big genius.	i wanted to like this place but it just became a big midwest.	i wanted to like this place but it just mildly a big stocked.
and i had my sugar bowl favorite, the absurd hurts ache!	and i had my sugar bowl broken, the garage hat holes!	and i had my sugar bowl misleading, the top quesadilla sundae!
expertly... we just delightful him that we did magical want adds marvelous.	um... we just loved him that we did definatly want to finance.	um... we just entertained him that we did perfected want to incredible.
ridiculous a place to keep in mind.	would a place to keep in mind.	wont a place to keep in mind.
firstly, their project are generally higher than other places.	firstly, their draw are generally higher than other places.	sheila, their round are generally higher than other places.
horrific the trap - tea at the gut.	dumb the afternoon - tea at the rabbit.	wtf the afternoon - tea at the slim.

Table 2: Reformulations generated by the baseline with additional discriminator, shifted autoencoder and shifted autoencoder with additional discriminator corresponding to the inputs in Table 1.

Aligning results shown on Figures 5 - 10 one can clearly see several crucial things: (1) architectures based on the idea of disentangled latent representations show varying performance in terms of BLEU between output and human written reformulations; (2) architectures with higher quality of information decomposition in terms of correlation or KL-divergence between representations for input and output, show higher performance; (3) architectures that produce equal semantic components for a given input and corresponding output show the highest performance; (4) these results are aligned with empirical estimation of decomposition quality with external classifiers; it shows that architectures that are more successfully disentangling semantics of the input from its stylistics tend to perform better.

5 Conclusion

This paper addresses the questions of information decomposition for the task of textual style transfer. We propose three new architectures that use latent representations to decompose stylistic and semantics information of input. Two different methods to assess the quality of such decomposition are proposed. It is shown that architectures that produce an equal semantic component of latent representations for input and corresponding output
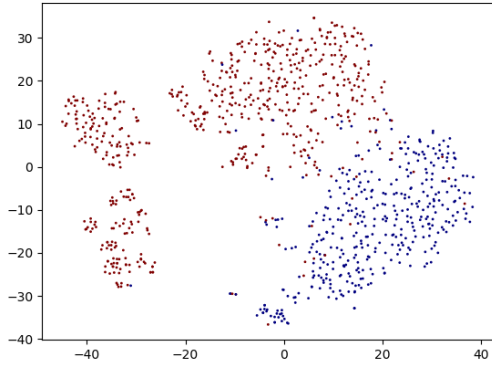
Figure 8: t-SNE visualisation of the obtained latent representations for the baseline architecture proposed in (Hu et al., 2017a). Red dots represent positive reviews. Blue dots represent negative reviews. One can clearly see that stylistic information can be recovered from the representation.

Figure 10: t-SNE visualisation of the obtained latent representations for the shifted autoencoder. Red dots represent positive reviews. Blue dots represent negative reviews. One can see that it is harder to recover stylistic information from the representation and the structure of the differs significantly from the latent representation space obtained by the baseline.
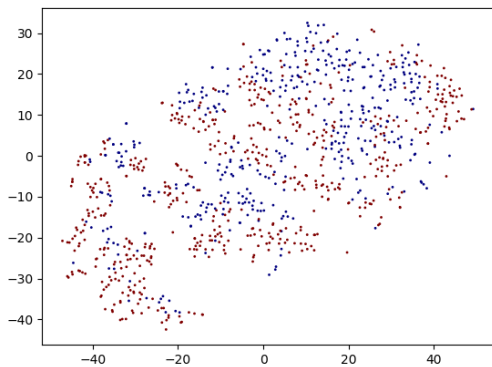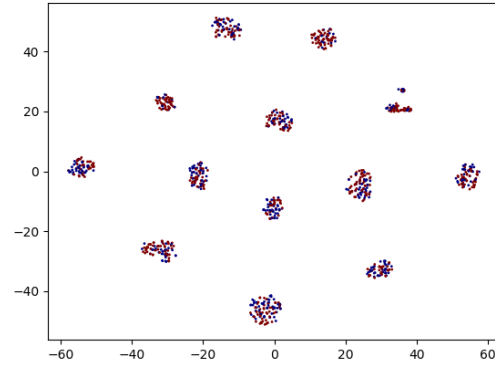

Figure 9: t-SNE visualisation of the obtained latent representations for the architecture with an additional discriminator. Red dots represent positive reviews. Blue dots represent negative reviews. One can see that it is harder to recover stylistic information from the representation.

outperform state of the art architectures in terms of BLEU between output and human written reformulations. An empirical method to assess the quality of information decomposition is proposed. There is a correspondence between higher BLEU between output and human written reformulations and better quality of information decomposition.

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.

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*.

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.

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.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605.

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*.

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*.

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*.

Alexey Tikhonov, Viacheslav Shibaev, Aleksander Nagaev, Aigul Nugmanova, and Ivan Yamshchikov. 2019. Style transfer for texts: to err is human, but error margins matter. In *EMNLP*.

Alexey Tikhonov and Ivan P. Yamshchikov. 2018a. 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. 2018b. 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.

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

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*.