
Proceedings Paper:
Li, X., Chen, G., Lin, C. orcid.org/0000-0003-3454-2468 et al. (1 more author) (2020) DGST : a dual-generator network for text style transfer. In: Webber, B., Cohn, T., He, Y. and Liu, Y., (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 16-20 Nov 2020, Virtual conference. Association for Computational Linguistics (ACL) , pp. 7131-7136. ISBN 9781952148606

10.18653/v1/2020.emnlp-main.578

DGST: a Dual-Generator Network for Text Style Transfer

Xiao Li[♠], **Guanyi Chen**[♡], **Chenghua Lin**[♠*], **Ruizhe Li**[♠]
[♠]Department of Computer Science, University of Sheffield
[♡]Department of Information and Computing Sciences, Utrecht University
{xiao.li, c.lin, r.li}@sheffield.ac.uk, g.chen@uu.nl

Abstract

We propose DGST, a novel and simple Dual-Generator network architecture for text Style Transfer. Our model employs two generators only, and does not rely on any discriminators or parallel corpus for training. Both quantitative and qualitative experiments on the Yelp and IMDb datasets show that our model gives competitive performance compared to several strong baselines with more complicated architecture designs.

1 Introduction

Attribute style transfer is a task which seeks to change a stylistic attribute of text, while preserving its attribute-independent information. Sentiment transfer is a typical example of such kind, which focuses on controlling the sentiment polarity of the input text (Shen et al., 2017). Given a review *"the service was very poor"*, a successful sentiment transferrer should covert the negative sentiment of the input to positive (e.g., replacing the phrase *"very poor"* with *"pretty good"*), while keeping all other information unchanged (e.g., the aspect *"service"* should not being changed to *"food"*).

Without supervised signals from parallel data, a transferrer must be supervised in a way to ensure that the generated texts belongs to a certain style category (i.e., transfer intensity). There is a growing body of studies to intensify the target style by means of adversarial training (Fu et al., 2018), variational autoencoder (John et al., 2019; Li et al., 2019; Fang et al., 2019), generative adversarial nets (Shen et al., 2017; Zhao et al., 2018; Yang et al., 2018), or subspace matrix projection (Li et al., 2020)

Furthermore, in order to boost the preservation of non-attribute information during style transformation, some works explicitly focus on modifying sentiment words, which is so-called the "pivot word" (Li et al., 2018; Wu et al., 2019). There are also works which add extra components for constraining the content from being changed too much. These include models like autoencoder (Lample et al., 2019; Dai et al., 2019), part-of-speech preservation, and the content conditional language model (Tian et al., 2018). In order to achieve high-quality style transfer, existing works normally resort to adding additional inner or outer structures such as additional adversarial networks or data preprocessing steps (e.g. generating pseudo-parallel corpora). This inevitably increases the complexity of the model and raises the bar of training data requirement.

In this paper, we propose a novel and simple model architecture for text style transfer, which employs two generators only. In contrast to some of the dominant approaches to style transfer such as CycleGAN (Zhu et al., 2017), our model does not employ any discriminators and yet can be trained without requiring any parallel corpus. We achieve this by developing a novel sentence noisification approach called *neighbourhood sampling*, which can introduce noise to each input sentence dynamically. The nosified sentences are then used to train our style transferrers in the way similar to the training of denoising autoencoders (Vincent et al., 2008). Both quantitative and qualitative evaluation on the Yelp and IMDb benchmark datasets show that DGST gives competitive performance compared to several strong baselines which have more complicated model design. The code of DGST is available at: https://xiao.ac/proj/dgst.

2 Methodology

Suppose we have two non-parallel corpora X and Y with style S_x and S_y, the goal is training two transferrers, each of which can (i) transfer a sen-

*Corresponding author

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7131–7136,
November 16–20, 2020.

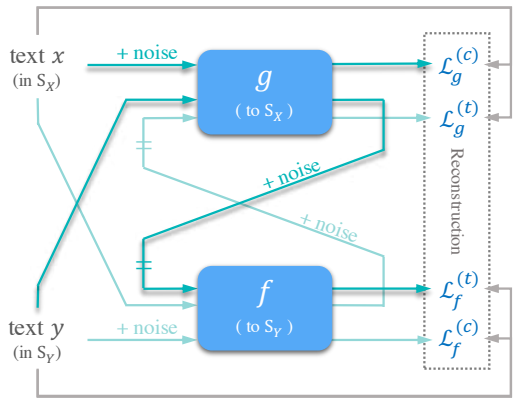

Figure 1: The general architecture of DGST, in which "=" means no back-propagation of gradients.

tence from one style (either S_x and S_y) to another (i.e., transfer intensity); and (ii) preserve the style-independent context during the transformation (i.e., preservation). Specifically, we denote the two transferrers f and g. $f : \mathcal{X} \rightarrow \mathcal{Y}$ transfers a sentence $x \in \mathcal{X}$ with style S_x to y^* with style S_y. Likewise, $g : \mathcal{Y} \rightarrow \mathcal{X}$ transfers a sentence $y \in \mathcal{Y}$ with style S_y to x^* with S_x. To obtain good style transfer performance, f and g need to achieve both a high transfer intensity and a high preservation, which can be formulated as follows:

$$\forall x, \forall x' \in \mathcal{X},\ \forall y, \forall y' \in \mathcal{Y}$$
$$y^* = f(x) \in \mathcal{Y}, x^* = g(y) \in \mathcal{X} \quad (1)$$
$$D(y^*||x) \leq D(y'||x), D(x^*||y) \leq D(x'||y) \quad (2)$$

Here $D(x||y)$ is a function that measures the abstract distance between sentences in terms of the minimum edit distance, where the editing operations Φ includes word-level replacement, insertion, and deletion (i.e., the Hamming distance or the Levenshtein distance). On the one hand, Eq. 1 requires the transferred text should fall within the target style spaces (i.e., $\mathcal{X}$ or $\mathcal{Y}$). On the other hand, Eq. 2 constrains the transferred text from changing too much, i.e., to preserve the style-independent information.

Inspired by CycleGAN (Zhu et al., 2017), our model (sketched in Figure 1) is trained by a cyclic process: for each transferrer, a text is transferred to the target style, and then back-transferred to the source style using another transferrer. In order to transfer a sentence to a target style while preserving the style-independent information, we formulate two sets of training objectives: one set ensures that the generated sentences is preserved as much as possible (detailed in §2.1) and the other set is responsible for transferring the input text to the target style (detailed in §2.2).

2.1 Preserving the Content of Input Text

This section discusses our loss function which enforces our transferrers to preserve the style-independent information of the input. A common solution to this problem is to use the reconstruction loss of the autoencoders (Dai et al., 2019), which is also known as the identity loss (Zhu et al., 2017). However, too much emphasis on preserving the content would hinder the style transferring ability of the transferrers. To balance our model's capability in content preservation and transfer intensity, we instead first train our transferrers in the way of training denoising autoencoders (DAE, Vincent et al., 2008), which has been proved to help preserving the style independent content of input text (Shen et al., 2020). More specifically, we train f (or g; we use f as an example in the rest of this section) by feeding it with a noisy sentence $\mathring{y}$ as input, where $\mathring{y}$ is noisified from $y \in \mathcal{Y}$ and f is expected to reconstruct y.

Different from previous works which use DAE in style transfer or MT (Artetxe et al., 2018; Lample et al., 2019), we propose a novel sentence noisification approach, named *neighbourhood sampling*, which introduces noise to each sentence dynamically. For a sentence y, we define $U_\alpha(y, \gamma)$ as a neighbourhood of y, which is a set of sentences consisting of y and all variations of noisified y with the same noise intensity γ (which will be explained later). The size of the neighbourhood $U_\alpha(y, \gamma)$ is determined by the proportion (denoted by m) of tokens in y that are modified using the editing operations in Φ. Here the proportion m is sampled from a Folded Normal Distribution $\mathcal{F}$. We hereby define that the average value of m (i.e., the mean of $\mathcal{F}$) is the noise intensity γ. Formally, m is defined as:

$$m \sim \mathcal{F}(m'; \gamma) = \frac{2}{\pi\gamma} e^{-\frac{m'^2}{\pi\gamma^2}} \quad (3)$$

That said, a neighbourhood $U_\alpha(y, \gamma)$ would be constructed using y and all sentences that are created by modifying $(m \times \text{length}(y))$ words in y, from which we sample $\mathring{y}$, i.e., a noisified sentence of y: $\mathring{y} \sim U_\alpha(y, \gamma)$. Analogously, we could also construct a neighbourhood $U_\beta(x, \gamma)$ for $x \in \mathcal{X}$ and sample $\mathring{x}$ from it. Using these noisified data as inputs, we then train our transferrers f and g in the way of DAE by optimising the following recon-

struction objectives:

$$\mathcal{L}_f^{(c)} = \mathbb{E}_{y\sim Y, \mathring{y}\sim U_\alpha(y,\gamma)} D(y||f(\mathring{y}))$$
$$\mathcal{L}_g^{(c)} = \mathbb{E}_{x\sim X, \mathring{x}\sim U_\beta(x,\gamma)} D(x||g(\mathring{x})) \quad (4)$$

With Eq. 4, we essentially encourages the generator to preserve the input as much as possible.

2.2 Transferring Text Styles

Making use of non-parallel datasets, we train f and g in an iterative process. Let $M = \{g(y)|y \in Y\}$ be the range of g when the input is all sentences in the training set Y. Similarly, we can define $N = \{f(x)|x \in X\}$. During the training cycle of f, g will be kept unchanged. We first feed each sentence y ($y \in Y$) to g, which tries to transfer y to the target style $\mathcal{X}$ (i.e. ideally $x^* = g(y) \in \mathcal{X}$). In this way, we obtain M which is composed of all x^* for each $y \in Y$. Next, we sample $\mathring{x}^*$ (a noised sentence of x^*) based on x^* via the neighbourhood sampling, i.e., $\mathring{x}^* \sim U_\alpha(x^*, \gamma) = U_\alpha(g(y), \gamma)$. We use $\mathring{M}$ to represent the collection of $\mathring{x}^*$. Similarly, we obtains N and $\mathring{N}$ using the aforementioned procedures during the training cycle for g.

Instead of directly using the sentences from X for training, we use $\mathring{M}$ to train f by forcing f to transfer each $\mathring{x}^*$ back to the corresponding original y. In parallel, $\mathring{N}$ is utilised to train g. We represent the aforementioned operation as the *transfer objective*.

$$\mathcal{L}_f^{(t)} = \mathbb{E}_{\alpha, y\sim Y, \mathring{x}^*\sim U_\alpha(g(y),\gamma)} D(y||f(\mathring{x}^*))$$
$$\mathcal{L}_g^{(t)} = \mathbb{E}_{\beta, x\sim X, \mathring{y}^*\sim U_\beta(f(x),\gamma)} D(x||g(\mathring{y}^*)) \quad (5)$$

The main difference between Eq. 4 and Eq. 5 is how $U_\alpha(\cdot, \gamma)$ and $U_\beta(\cdot, \gamma)$ are constructed, i.e., $U_\alpha(y, \gamma)$ and $U_\beta(x, \gamma)$ in Eq. 4 compared to $U_\alpha(g(y), \gamma)$ and $U_\beta(f(x), \gamma)$ in Eq. 5. Finally, the overall loss of DGST is the sum of the four partial losses:

$$\mathcal{L} = \mathcal{L}_f^{(c)} + \mathcal{L}_f^{(t)} + \mathcal{L}_g^{(c)} + \mathcal{L}_g^{(t)} \quad (6)$$

During optimisation, we freeze g when optimising f, and vice versa. Also with the reconstruction objective, x^* must to be sampled first, and then passed $\mathring{x}^*$ into f; in contrast, it is not necessary to sample according to y when we obtain $x^* = g(y)$.

3 Experiment

3.1 Setup

Dataset. We evaluated our model on two benchmark datasets, namely, the Yelp review dataset (Yelp), which consists of restaurants and business reviews together with their sentiment polarity (i.e., positive or negative), and the IMDb Movie Review Dataset (IMDb), which consists of online movie reviews. For Yelp, we split the dataset following Li et al. (2018), who also provided human produced reference sentences for evaluation. For IMDb, we follow the pre-processing and data splitting protocol of Dai et al. (2019). Detailed dataset statistics is given in Table 1.

Dataset	Yelp		IMDb	
	Positive	Negative	Positive	Negative
Train	266,041	177,218	178,869	187,597
Dev	2,000	2,000	2,000	2,000
Test	500	500	1,000	1,000

Table 1: Statistics of Datasets.

Evaluation Protocol. Following the standard evaluation practice, we evaluate the performance of our model on the textual style transfer task from two aspects: (1) **Transfer Intensity**: a style classifier is employed for quantifying the intensity of the transferred text. In our work, we use FastText (Joulin et al., 2017) trained on the training set of Yelp; (2) **Content Preservation**: to validate whether the style-independent context is preserved by the transferrer, we calculate *self*-BLEU, which computes a BLEU score (Papineni et al., 2002) by comparing inputs and outputs of a system. A higher *self*-BLEU score indicates more tokens from the sources are retained, henceforth, better preservation of the contents. In addition, we also use *ref*-BLEU, which compares the system outputs and the references written by human beings.

3.2 Experimental Results

In our experiment, the two transferrers (f and g) are Stacked BiLSTM-based sequence-to-sequence models, i.e., both 4-layer BiLSTM for the encoder and decoder. The noise intensity γ is set to 0.3 in the first 50 epochs and 0.03 in the following epochs.

As shown in Table 2, for the Yelp dataset our model defeats all baselines models (apart from StyleTransformer (Multi-Class)) on both *ref*-BLEU and *self*-BLEU. In addition, as shown in Table 2, our model works remarkably well on both transfer intensity and preservation without requiring adversarial training or reinforcement learning, or external offline sentiment classifiers (as in Dai et al. (2019)). Besides, the current version of our

Model	Yelp			IMDb	
	acc.	*ref*-BLEU	*self*-BLEU	acc.	*self*-BLEU
RetrieveOnly (Li et al., 2018)	92.6	0.4	0.7	n/a	n/a
TemplateBased (Li et al., 2018)	84.3	13.7	44.1	n/a	n/a
DeleteOnly (Li et al., 2018)	85.7	9.7	28.6	n/a	n/a
DeleteAndRetrieve (Li et al., 2018)	87.7	10.4	29.1	55.8	55.4
ControlledGen (Hu et al., 2017)	88.8	14.3	45.7	94.1	62.1
CycleRL (Xu et al., 2018)	88.0	2.8	7.2	**97.8**	4.9
StyleTransformer (Conditional) (Dai et al., 2019)	**93.7**	17.1	45.3	86.6	66.2
StyleTransformer (Multi-Class) (Dai et al., 2019)	87.7	**20.3**	**54.9**	80.3	**70.5**
DGST	88.0	18.7	54.5	70.1	70.2

Table 2: Automatic evaluation results on Yelp and IMDb corpora, most of which are from Dai et al. (2019).

Yelp	**positive → negative**
input	this golf club is one of the best in my opinion .
output	this golf club is one of the worst in my opinion .
input	i definitely recommend this place to others !
output	i do not recommend this to anyone !
Yelp	**negative → positive**
input	the garlic bread was bland and cold .
output	the garlic bread was tasty and fresh .
input	my dish was pretty salty and could barely taste the garlic crab .
output	my dish was pretty good and could even taste the garlic crab .
IMDb	**positive → negative**
input	a timeless classic , one of the best films of all time .
output	a complete disaster , one of the worst films of all time .
input	and movie is totally backed up by the excellent music both in background and in songs by monty .
output	the movie is totally messed up by the awful music both in background and in songs by chimps .
IMDb	**negative → positive**
input	this one is definitely one for my " worst movies ever " list .
output	this one is definitely one of my " best movies ever " list .
input	i found this movie puerile and silly , as well as predictable .
output	i found this movie credible and funny , as well as tragic .

Table 3: Example results from our model for the sentiment style transfer on the Yelp and IMDb datasets.

model is built upon fundamental BiLSTM, which is a likely explanation of why we lose to the SOTA (i.e., StyleTransformer (Multi-Class)) for a small margin, which are based on the Transformer architecture (Vaswani et al., 2017) with much higher capacity. For the IMDb dataset, comparing to other systems, our model obtained moderate accuracy but competitive *self*-BLEU score (70.2), i.e., slightly lower than StyleTransformer. Table 3 lists several examples for style transfer in sentiment for both datasets. By examining the results, we can see that DGST is quite effective in transferring the sentiment polarity of the input sentence while maintaining the non-sentiment information.

3.3 Ablation Study

To confirm the validity of our model, we did an ablation study on Yelp by eliminating or modifying a certain component (e.g., objective functions, or sampling neighbourhood). We tested the following variations: 1) **full-model**: the proposed model; 2) **no-tran**: the model without the transfer objective; 3) **no-rec**: the model without the reconstruction objective; 4) **rec-no-noise**: the model adding no noise when optimising the reconstruction objective; 5) **tran-no-noise**: the model adding no noise when optimising the transfer objective; 6) **pre-noise**: the model trained by adding noise to y first and then feeding the nosified sentences $\mathring{y}$ to g (or $\mathring{x}$ to f) in Eq. 5. In this study, the transferrers are the simplest LSTM-based sequence-to-sequence models. The hidden size and γ are set to 256 and 0.3, respec-

	positive → negative	negative → positive
input	*it is a cool place , with lots to see and try .*	*so , that was my one and only time ordering the benedict there .*
full-model	it is a sad place , with lots to see and something .	so , that was my one and best time in the shopping there .
no-rec	no no , , _num_ .	so , that was my one and time time over the there there .
rec-no-noise	it is a cool place , with me to see and try .	service was very friendly .
no-tran	it is a loud place , with lots to try and see .	so , that was my only and first visit ordering the there) .
tran-no-noise	it is a noisy place , with lots to try and see .	so , that was my one and time time ordering the ordering there .
pre-noise	it is a cool place , with lots to see to try .	so , that 's one one and my only the the day there .
input	*it is the most authentic thai in the valley .*	*even if i was insanely drunk , i could n't force this pizza down .*
full-model	it is the most overrated thai in the valley .	even if i was n't hungry , i 'll definitely enjoy this pizza here .
no-rec	i was in the the the the food .	she was perfect .
rec-no-noise	it is the most authentic thai in the valley .	even if i was n't , , i could n't recommend this pizza . .
no-tran	it is the most authentic thai in the valley .	even if i was n't , , i could n't get this pizza down .
tran-no-noise	it is the most common thai in the valley .	even if i was hungry hungry , i could n't love this pizza shop .
pre-noise	it is the most thai thai in the valley .	even if i was n't hungry , i could n't recommend this pizza down .

Table 4: Example transferred from the ablation study.

Model Variants	*self*-BLEU	acc.
no-rec	0.0	98.9
rec-no-noise	41.9	73.1
no-tran	98.0	4.2
tran-no-noise	35.6	82.9
pre-noise	38.9	76.8
full-model	37.2	86.3

Table 5: Evaluation results for the ablation study.

tively.

Results. Table 5 depicts the results of the ablation study. As we can see, eliminating the reconstruction or transfer objectives would damage preservation and transfer intensity, respectively. As for the use of noise, the results of the rec-no-noise model shows that the noise in the reconstruction objective helps balance our model's ability in content preservation and transfer intensity. For the transfer objective, omitting noise (tran-no-sp) would reduce the transfer intensity while placing noise in the wrong position (pre-noise) reduces it yet again.

Case Study. Transferred sentences produced by each model variant in the ablation study are listed in Table 4. The model without correction objective (no-corr) collapsed and as a result it generates irrelevant sentences to the inputs most of the time. When neighbourhood sampling is dropped in either corrective or transfer objectives, the transfer intensity is reduced. These models, including rec-no-noise, tran-no-noise, and pre-noise, tend to substitute random words, and result in reduced transfer intensity (i.e., style words are either not modified or still express the same sentiment after modification) and preservation. For example, when transferring from negative to positive, rec-no-noise replace *"force"* to *"recommend"* resulting *"I couldn't recommend this pizza"*, which is still a negative review.

4 Conclusion

In this paper, we propose a novel and simple dual-generator network architecture for text style transfer, which does not rely on any discriminators or parallel corpus for training. Extensive experiments on two public datasets show that our model yields competitive performance compared to several strong baselines, despite of our simpler model architecture design.

Acknowledgements

We would like to thank all the anonymous reviewers for their insightful comments. This work is supported by the award made by the UK Engineering and Physical Sciences Research Council (Grant number: EP/P011829/1).

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. Association for Computational Linguistics.

Le Fang, Chunyuan Li, Jianfeng Gao, Wen Dong, and Changyou Chen. 2019. Implicit deep latent variable models for text generation. 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)*, pages 3937–3947.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596. JMLR. org.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

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 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Ruizhe Li, Xiao Li, Chenghua Lin, Matthew Collinson, and Rui Mao. 2019. A stable variational autoencoder for text modelling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 594–599.

Xiao Li, Chenghua Lin, Ruizhe Li, Chaozheng Wang, and Frank Guerin. 2020. Latent space factorisation and manipulation via matrix subspace projection. In *Proceedings of Machine Learning and Systems 2020*, pages 3211–3221.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.

Tianxiao Shen, Jonas Mueller, Regina Barzilay, and Tommi Jaakkola. 2020. Educating text autoencoders: Latent representation guidance via denoising. In *Proceedings of Machine Learning and Systems 2020*, pages 9129–9139.

Youzhi Tian, Zhiting Hu, and Zhou Yu. 2018. Structured content preservation for unsupervised text style transfer. *arXiv preprint arXiv:1810.06526*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Mask and infill: Applying masked language model for sentiment transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5271–5277. International Joint Conferences on Artificial Intelligence Organization.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia. Association for Computational Linguistics.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, pages 7287–7298.

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. volume 80 of *Proceedings of Machine Learning Research*, pages 5902–5911, Stockholmsmässan, Stockholm Sweden. PMLR.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.