Neural Text Style Transfer via Denoising and Reranking

Joseph Lee*, Ziang Xie*, Cindy Wang, Max Drach, Dan Jurafsky, Andrew Y. Ng
Computer Science Department, Stanford University
{joseph.lee,zxie,cindyw,mdrach,ang}@cs.stanford.edu,
jurafsky@stanford.edu

Abstract

We introduce a simple method for text style transfer that frames style transfer as denoising: we synthesize a noisy corpus and treat the source style as a noisy version of the target style. To control for aspects such as preserving meaning while modifying style, we propose a reranking approach in the data synthesis phase. We evaluate our method on three novel style transfer tasks: transferring between British and American varieties, text genres (formal vs. casual), and lyrics from different musical genres. By measuring style transfer quality, meaning preservation, and the fluency of generated outputs, we demonstrate that our method is able both to produce high-quality output while maintaining the flexibility to suggest syntactically rich stylistic edits.

1 Introduction

Following exciting work on style transfer for images (Gatys et al., 2016), neural style transfer for text has gained research interest as an application and testbed for syntactic and semantic understanding of natural language (Li et al., 2018; Shen et al., 2017; Hu et al., 2017; Prabhumoye et al., 2018). Unfortunately, unlike image style transfer, which often requires only a single reference image in the desired style, neural text style transfer typically requires a large parallel corpus of sentences in the source and target style to train a neural machine translation model (Sutskever et al., 2014; Bahdanau et al., 2014).

One approach to mitigate the need for a large parallel corpus is to develop methods to disentangle stylistic attributes from semantic content, for example by using adversarial classifiers (Shen et al., 2017) or by predefining markers associated with stylistic attributes (Li et al., 2018). However, such approaches can reduce fluency and alter meaning, or make only lexical changes instead of larger, phrase-level edits.

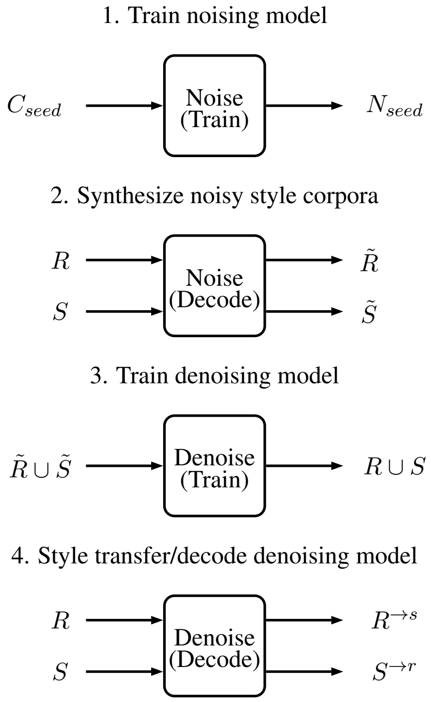

Figure 1: An overview of our method. We assume a seed corpus of parallel (clean, noisy) sentence pairs $(C, N) = \{(C_k, N_k)\}_{k=1}^{M_{\text{seed}}}$, as well as two nonparallel corpora $R = \{R_k\}_{k=1}^{M_R}$ and $S = \{S_k\}_{k=1}^{M_S}$ of different styles. We first use noising to generate synthetic parallel data in both styles, then "denoise" to transfer from one style to the other.

Given the limitations of these techniques, we propose an approach which uses backtranslation (Sennrich et al., 2015a) to synthesize parallel data, starting with nonparallel data in differing styles. We introduce a simple method for unsupervised text style transfer that frames style transfer as a *denoising* problem in which we treat the source style as a noisy version of the target style. By further introducing hypothesis reranking techniques in the data synthesis procedure, our method

[*] Equal contribution.

Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen), pages 74–81
Minneapolis, Minnesota, USA, June 6, 2019. ©2019 Association for Computational Linguistics

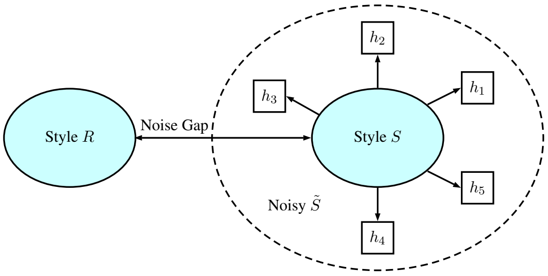

Figure 2: When synthesizing noisy sentences to train the denoising model $\tilde{S} \to S$, we use style reranking to choose the noisy hypothesis h closest to the alternate style R. In this case, h_3 minimizes the "noise gap."

(summarized in Figure 1) allows for rich syntactic modifications while encouraging preservation of meaning.

We evaluate our method on three distinct style transfer tasks, transferring between English varieties (American and British), formal and informal writing (news data and Internet forum data), and lyrics of different musical genres (pop and hip hop). We use three criteria to measure the quality of outputs that have been mapped to the target style: style transfer strength, meaning preservation, and fluency. Despite the simplicity of the method, we demonstrate that it is capable of making syntactically rich suggestions. The proposed reranking technique can also be used to modulate aspects of the style transfer, such as the degree to which the style is applied or the extent to which meaning is changed.

2 Method

We assume a seed corpus of parallel (clean, noisy) sentence pairs $(C, N) = \{(C_k, N_k)\}_{k=1}^{M_{\text{seed}}}$, as well as two non-parallel corpora $R = \{R_k\}_{k=1}^{M_R}$ and $S = \{S_k\}_{k=1}^{M_S}$ of different styles.

2.1 Noising

We first synthesize noisy versions of R and S. We first obtain a seed noise corpus of (clean, noisy) sentence pairs from a language learner forum. Using the seed noise corpus, we train a neural sequence transduction model to learn the mapping from clean to noisy $C \to N$ from our (clean, noisy) sentence pairs. Then, we decode R and S using the noising model to synthesize the corresponding noisy versions, $\tilde{R}$ and $\tilde{S}$.

- **Baseline** As a baseline, we apply the noising method described in Xie et al. (2018). This method utilizes beam search noising techniques to encourage diversity during the noising process in order to avoid copying of the inputs.

- **Style Reranking** A shortcoming of the baseline noising method is that it mimics the noise in the initial seed corpus, which may not match well with the input style. In order to produce noise that better matches the inputs that will later be fed to the denoising model, we perform reranking to bias the synthesized noisy corpora $\tilde{R}$ and $\tilde{S}$ towards the clean corpora S and R, respectively.

  Consider the noise synthesis for S, and denote the noising procedure for a single input as $f_{\text{noise}}(\cdot)$. We generate multiple noise hypotheses, $h_i = f_{\text{noise}}(S_k)$ and select the hypothesis closest to the alternate style R, as ranked by a language model trained on R:

  $$h^* = \arg\max_i p_R(h_i)$$

  Figure 2 illustrates the intuition that the style reranking will result in noised data "closer" to the expected source inputs.

- **Meaning Reranking** Similar to style reranking, we rerank the hypotheses to encourage meaning preservation by ranking the different noise hypotheses according to the cosine similarity of the sum of word embeddings between the hypothesis and the original source input.

2.2 Denoising

After the synthesized parallel corpus is generated, we train a denoising model between the synthesized noisy corpora and the clean counterparts. To encode style information, we prepend a start token to each noisy sentence corresponding to its style, i.e. $\tilde{R}_k = (\langle\text{style}\rangle, w_1, w_2, \ldots, w_T)$.

Besides providing a simple method to specify the desired target style, this also allows us to combine the noisy-clean corpora from each of the two styles and train a single model using both corpora. This provides two benefits. First, it allows us to learn multiple styles in one model. This allows one model to perform style transfer from both $R \to S$ and $S \to R$. Second, multi-task learning often improves the performance for each of the separate tasks (Luong et al., 2016).

We then join the corpora to obtain the (clean, noisy) sentence pairs,

$$(X, \tilde{X}) = \{(X_k, \tilde{X}_k)\}_{k=1}^{M_{R+S}},$$

from which we will learn our denoising model. Our denoising model learns the probabilistic mapping $P(X|\tilde{X})$, obtaining model parameters θ^* by minimizing the loss function:

$$\mathcal{L}(\theta) = -\sum_{k=1}^{M_{R+S}} \log P(X_k|\tilde{X}_k; \theta)$$

For our experiments we use the Transformer encoder-decoder model (Vaswani et al., 2017) with byte-pair encoding (Sennrich et al., 2015b) with vocabulary size of 30000. We follow the usual training procedure of minibatch gradient descent to minimize negative log-likelihood.

The trained denoising model is then applied to the source style—that we treat as the "noisy" corpus—with the start token of the target style to perform style transfer (Figure 1).

3 Experiments

Task	Dataset	Training	LM
US/UK	NYT/BNC	800K	2.5MM
Forum/News	NYT/Reddit	800K	2.5MM
Music Genres	Hip Hop/Pop	500K	400K

Table 1: Style transfer datasets and number of sentences. *Training* refers to examples used to synthesize noisy sentences and train the denoising model. *LM* refers to examples used to train language models for reranking and evaluation. In addition to training and LM data, 20K examples are held out for each of the dev and test sets.

3.1 Data

We evaluate our methods on three different style transfer tasks between the following corpus pairs: (1) American and British English, (2) formal news writing and informal forum writing, and (3) pop and hip hop lyrics. The first task of transferring between American and British English is primarily intended as a preliminary test for our proposed technique by demonstrating that it can capture lexical changes. The latter two tasks require more sophisticated syntactic edits and form the basis of our later analysis.

A summary of the datasets used for the three tasks is provided in Table 1. We use The New York Times for the American English data, the British National Corpus for the British English data, and the Reddit comments dataset for informal forum data. The pop and hip hop lyrics are gathered from MetroLyrics.[1] For the parallel seed corpus used to train the noising model, we use a dataset of roughly 1MM sentences collected from an English language learner forum (Tajiri et al., 2012).

3.2 Evaluation

We define effective style transfer using the following criteria:

1. **Transfer strength** For a given output sentence, effective style transfer should increase the probability under the target style distribution relative to the probability of observing it under the source style distribution. We thus define transfer strength as the ratio of target-domain to source-domain shift in sentence probability. Let R be the source style inputs and $R^{\to\text{tgt}}$ be the target style outputs. Then,

$$\begin{aligned}\textsc{Shift}_{\text{src}} = \exp\Bigg[&\frac{1}{n}\sum_{k=1}^{n}\log(P(R_k^{\to\text{tgt}}|LM_{\text{src}})) \\ &-\frac{1}{n}\sum_{k=1}^{n}\log(P(R_k|LM_{\text{src}}))\Bigg] \\ \textsc{Shift}_{\text{tgt}} = \exp\Bigg[&\frac{1}{n}\sum_{k=1}^{n}\log(P(R_k^{\to\text{tgt}}|LM_{\text{tgt}})) \\ &-\frac{1}{n}\sum_{k=1}^{n}\log(P(R_k|LM_{\text{tgt}}))\Bigg] \\ \textsc{TransferStrength}_{\text{src}\to\text{tgt}} &\overset{\text{def}}{=} \frac{\textsc{Shift}_{\text{tgt}}}{\textsc{Shift}_{\text{src}}}\end{aligned}$$

A positive transfer is any ratio greater than one.

2. **Meaning preservation** The target output should also have similar meaning and intent as the source. To measure this, we compute the cosine similarity between embeddings r of the source and target:

$$\textsc{MeaningPreservation} \overset{\text{def}}{=} \frac{{r_{\text{src}}}^{\top} r_{\text{tgt}}}{\|r_{\text{src}}\|\|r_{\text{tgt}}\|}$$

[1] https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics

	NYT↔BNC		NYT↔Reddit		Pop↔Hip hop	
Method	→	←	→	←	→	←
Baseline	1.315	1.227	1.252	1.202	1.097	1.086
Style Rerank	**1.359**	**1.274**	**1.312**	**1.246**	1.110	1.072
Meaning Rerank	1.285	1.222	1.281	1.145	**1.118**	**1.092**

Table 2: The transfer strength for each style transfer task.

	NYT↔BNC		NYT↔Reddit		Pop↔Hip hop	
Method	→	←	→	←	→	←
Baseline	92.22	92.17	97.00	91.56	96.84	97.22
Style Rerank	91.93	91.66	97.25	91.10	96.84	97.18
Meaning Rerank	**94.40**	**93.47**	**98.34**	**94.18**	**97.29**	**97.48**

Table 3: Meaning preservation for each style transfer task. All reported numbers scaled by 10^2 for display.

	NYT↔BNC		NYT↔Reddit		Pop↔Hip hop	
Method	→	←	→	←	→	←
Pre-Transfer	**5.763**	3.891	**4.609**	**5.763**	**2.470**	**1.453**
Baseline	4.016	**4.012**	3.920	5.506	2.112	1.429
Style Rerank	3.877	3.992	3.603	5.194	1.930	1.310
Meaning Rerank	3.874	3.743	3.808	5.395	1.915	1.284

Table 4: Fluency of each style transfer task. All reported numbers scaled by 10^3 for display.

To compute the embeddings r, we use the sentence encoder provided by the InferSent library, which has demonstrated excellent performance on a number of natural language understanding tasks (Conneau et al., 2017).

3. **Fluency** The post-transfer sentence should remain grammatical and fluent. We use the average log probability of the sentence post-transfer with respect to a language model trained on CommonCrawl as our measure of fluency.

The source and target language models are 4-gram (in the case of music lyrics) or 5-gram (in the case of other datasets) language models trained on a held-out subset of each corpus, estimated with Kneser-Ney smoothing using KenLM (Heafield et al., 2013).

Task	Base	Rerank	No Pref
NYT → BNC	6.00	6.25	87.8
BNC → NYT	10.8	6.5	82.8
NYT → Reddit	6.75	9.5	83.8
Reddit → NYT	9.75	18.3	72.0
Pop → Hip Hop	5.25	6.50	88.3
Hip Hop → Pop	7.5	10.3	82.3

Table 5: Human evaluation results for style transfer strength. Entries give percentage of time where annotator preferred base vs. rerank (combined for 2 annotators).

3.3 Pairwise Human Evaluation of Reranking

While language model likelihood is an established measure of fluency or grammaticality, and InferSent has been used as an effective sentence representation on a number of natural language understanding tasks (Conneau et al., 2017), we wish to validate our transfer strength results for our proposed reranking method using human evaluation as well.

For each of the six tasks (3 pairs crossed with 2 directions), we randomly selected 200 sentences, then took the outputs with models trained using style reranking and without style reranking. We then randomized the outputs such that the human evaluators would not be given the label for which output was produced using reranking.

Two annotators then labeled each (randomized) pair with the sentence that seemed to have higher transfer strength. We allowed for a "No preference" option for cases where neither output seemed to have higher transfer strength. We chose pairwise comparisons as it seemed most robust to sometimes minor changes in the sentences. Results are shown in Table 5. We see that while for Reddit → NYT there seems to be a clear preference, in most cases stylistic differences tend to be subtle given small differences in transfer strength.

3.4 Results

As shown in Table 2, we observed positive style transfer on all six transfer tasks. For the task of British and American English as well as formal news writing and informal forum writing, applying style reranking during the noising process increased the transfer strength across all four of these tasks. On the other hand, applying meaning reranking during the noising process often decreased the transfer strength. For pop and hip hop lyrics, we do not observe the same pattern; this may be due to the lack of data for the language model, thereby leading to less effective style reranking. In Section 4.1, we also address the possibility of a mismatch with the initial seed corpus.

As noted in Table 3, meaning is also well-preserved. On this metric, the meaning rerank method outperformed the other two models across all six tasks, showing the effectiveness of the reranking method.

In all six style transfer tasks in Table 4, the fluency was highest for the baseline model as compared to the reranked models, although fluency is often higher for the original sentence pairs. We suspect that transfer strength and meaning preservation are largely orthogonal to fluency, and hence encouraging one of the metrics can lead to dropoffs in the others.

4 Discussion

After experimental evidence that the proposed method produces reasonable stylistic edits, we wished to better understand the effects of our reranking methods as well as the choice of our initial seed corpus.

4.1 Limitations of Noise Corpus

A key factor in the performance of our style transfer models is the noisy data synthesis. Our method relies on an initial seed corpus of (clean, noisy) sentence pairs to bootstrap training. However, such a corpus is not ideal for the style transfer tasks we consider, as there is mismatch in many cases between the style transfer domains (e.g. news, music lyrics, forum posts) and the seed corpus (language learner posts). We observe in Table 2 that more significant transfer appears to occur for the tasks involving news data, and less for music lyrics.

To examine why this might be the case, we trained a 5-gram LM on the clean portion of the initial seed corpus, corresponding to the input of the noise model. We then measured the perplexity of this language model on the different domains. Results are given in Table 7. This may indicate why style transfer with music lyrics proved most difficult, as there is the greatest domain mismatch between the initial seed corpus and those corpora.

4.2 Comparing with Prior Work on Sentiment Transfer

Prior work on text style transfer has often focused on transferring between positive and negative sentiment (Li et al. (2018), Shen et al. (2017)). When we applied our method and evaluation trained on the same Yelp sentiment dataset as Li et al. (2018), using a subset of the Yelp Dataset for training our language model,[2] we obtained positive style transfer results across all three models (Table 8).

However, on further inspection of our decoded outputs, sentiment did not appear to change despite our evaluation metrics suggesting positive style transfer. This apparent contradiction can be explained by our approach treating sentiment as a *content* attribute instead of a *style* attribute.

The problem of sentiment transfer can be construed as changing certain content attributes while

[2] https://www.kaggle.com/yelp-dataset/yelp-dataset

Task	Source	Target
UK to US	As the BMA's own study of alternative therapy showed, life is not as simple as that.	As the *F.D.A.'s* own study of alternative therapy showed, life is not as simple as that.
US to UK	The Greenburgh Drug and Alcohol Force and investigators in the Westchester District Attorney's Narcotics Initiative Program Participated in the arrest.	The *Royal Commission on Drug and Attache Force* and investigators in the Westchester District Attorney's Initiative Program Participated in the arrest.
NYT to Reddit	The votes weren't there.	*There weren't any upvotes.*
Reddit to NYT	i guess you need to refer to bnet website then.	*I* guess you need to refer to *the* bnet website then.
Pop to Hip Hop	My money's low	My money's *on the low*
Hip Hop to Pop	Yo, where the hell you been?	Yo, where the hell *are you?*

Table 6: Qualitative examples of style transfer results for different tasks. No parallel data outside of the initial noise corpus was used. Note that the style transfer approach can generate targets with significant syntactic changes from the source. All examples shown are without reranking during data synthesis. BMA refers to the British Medical Association.

NYT	686 (460)	BNC	608 (436)
Reddit	287 (215)	Pop	702 (440)
Hip hop	1239 (802)		

Table 7: Perplexities with (and without) OOVs for different datasets under seed corpus language model.

	Yelp Pos ↔ Neg	
Method	→	←
Baseline	1.182	1.184
Style Rerank	1.189	1.198
Meaning Rerank	1.197	1.191

Table 8: Transfer strength for our method on Yelp sentiment transfer task shows positive style transfer (> 1).

	Yelp Pos ↔ Neg	
Method	→	←
Baseline	96.91	97.87
Style Rerank	97.33	97.74
Meaning Rerank	97.17	98.18
Shen et al. (2017)	96.03	96.32
Li et al. (2018)	90.82	92.36

Table 9: Meaning Preservation for our models as well as CrossAlign (Shen et al. (2017)) and DeleteAndRetrieve (Li et al. (2018)) on Yelp Sentiment Transfer Task. All reported numbers scaled by 10^2 for display.

keeping other style and content attributes constant. Meanwhile, style transfer aims to change style attributes while preserving all content attributes and thus preserving semantic meaning. Modifying style attributes include syntactic changes or word choices which might be more appropriate for the target style, but does not fundamentally change the meaning of the sentence.

A look at the meaning preservation metric across our models and across some models from prior work (Table 9) validates this hypothesis. Models that report higher-quality sentiment transfer such as Li et al. (2018) perform more poorly on the metric of meaning preservation, suggesting that changing a Yelp review from a positive review to a negative one fundamentally changes the content and meaning of the review, not just the style. Our model thus performs poorly on sentiment transfer, since our denoising method is limited to modifying style attributes while preserving all content attributes.

5 Related Work

Our work is related to broader work in training neural machine translation models in low-resource settings, work examining effective methods for applying noise to text, as well as work in style transfer.

Machine translation Much work in style transfer builds off of work in neural machine translation, in particular recent work on machine translation without parallel data using only a dictionary or aligned word embeddings (Lample et al., 2017; Artetxe et al., 2017). These approaches also use backtranslation while introducing token-level corruptions to avoid the problem of copying during an initial autoencoder training phase. They additionally use an initial dictionary or embedding alignments which may be infeasible to collect for many style transfer tasks. Finally, our work also draws from work on zero-shot translation between languages given parallel corpora with a pivot language (Johnson et al., 2017).

Noising and denoising To our knowledge, there has been no prior work formulating style transfer as a denoising task outside of using token corruptions to avoid copying between source and target. Our style transfer method borrows techniques from the field of noising and denoising to correct errors in text. We apply the noising technique in Xie et al. (2018) that requires an initial noise seed corpus instead of dictionaries or aligned embeddings. Similar work for using noise to create a parallel corpus includes Ge et al. (2018).

Style transfer Existing work for style transfer often takes the approach of separating *content* and *style*, for example by encoding a sentence into some latent space (Bowman et al., 2015; Hu et al., 2017; Shen et al., 2017) and then modifying or augmenting that space towards a different style. Hu et al. (2017) base their method on variational autoencoders (Kingma and Welling, 2014), while Shen et al. (2017) instead propose two constrained variants of the autoencoder. Yang et al. (2018) use language models as discriminators instead of a binary classifier as they hypothesize language models provide better training signal for the generator. In the work perhaps most similar to the method we describe here, Prabhumoye et al. (2018) treat style transfer as a backtranslation problem, using a pivot language to first transform the original text to another language, then encoding the translation to a latent space where they use adversarial techniques to preserve content while removing style.

However, such generative models often struggle to produce high-quality outputs. Li et al. (2018) instead approaches the style transfer task by observing that there are often specific phrases that define the attribute or style of the text. Their model segments in each sentence the specific phrases associated with the source style, then use a neural network to generate the target sentence with replacement phrases associated with the target style. While they produce higher quality outputs than previous methods, this method requires manual annotation and may be more limited in capturing rich syntactic differences beyond the annotated phrases.

6 Conclusion

In this paper, we propose a *denoising* method for performing text style transfer by treating the source text as a noisy version of the desired target. Our method can generate rich edits to map inputs to the target style. We additionally propose two reranking methods during the data synthesis phase intended to encourage meaning preservation as well as modulate the strength of style transfer, then examine their effects across three varied datasets. An exciting future direction is to develop other noising methods or datasets in order to consistently encourage more syntactically rich edits.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pages 2414–2423.

Tao Ge, Furu Wei, and Ming Zhou. 2018. Fluency boost learning and inference for neural grammatical error correction. *Association for Computational Linguistics*.

K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *ACL-HLT*, pages 690–696, Sofia, Bulgaria.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *ICML*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, and Nikhil Thorat. 2017. Googles multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *ICLR*.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

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. Association for Computational Linguistics.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *ICLR*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for esl learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 198–202. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Y. Ng, and Dan Jurafsky. 2018. Noising and denoising natural language: Diverse backtranslation for grammar correction. *NAACL-HLT 2018*.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P. Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *NIPS*.