Style Transfer Through Back-Translation

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, Alan W Black
Carnegie Mellon University, Pittsburgh, PA, USA
{sprabhum,ytsvetko,rsalakhu,awb}@cs.cmu.edu

Abstract

Style transfer is the task of rephrasing the text to contain specific stylistic properties without changing the intent or affect within the context. This paper introduces a new method for automatic style transfer. We first learn a latent representation of the input sentence which is grounded in a language translation model in order to better preserve the meaning of the sentence while reducing stylistic properties. Then adversarial generation techniques are used to make the output match the desired style. We evaluate this technique on three different style transformations: sentiment, gender and political slant. Compared to two state-of-the-art style transfer modeling techniques we show improvements both in automatic evaluation of style transfer and in manual evaluation of meaning preservation and fluency.

1 Introduction

Intelligent, situation-aware applications must produce naturalistic outputs, lexicalizing the same meaning differently, depending upon the environment. This is particularly relevant for language generation tasks such as machine translation (Sutskever et al., 2014; Bahdanau et al., 2015), caption generation (Karpathy and Fei-Fei, 2015; Xu et al., 2015), and natural language generation (Wen et al., 2017; Kiddon et al., 2016). In conversational agents (Ritter et al., 2011; Sordoni et al., 2015; Vinyals and Le, 2015; Li et al., 2016), for example, modulating the politeness style, to sound natural depending upon a situation: at a party with friends "Shut up! the video is starting!", or in a professional setting "Please be quiet, the video will begin shortly.".

These goals have motivated a considerable amount of recent research efforts focused at "controlled" language generation—aiming at separating the semantic content of *what* is said from the stylistic dimensions of *how* it is said. These include approaches relying on heuristic substitutions, deletions, and insertions to modulate demographic properties of a writer (Reddy and Knight, 2016), integrating stylistic and demographic speaker traits in statistical machine translation (Rabinovich et al., 2016; Niu et al., 2017), and deep generative models controlling for a particular stylistic aspect, e.g., politeness (Sennrich et al., 2016), sentiment, or tense (Hu et al., 2017; Shen et al., 2017). The latter approaches to style transfer, while more powerful and flexible than heuristic methods, have yet to show that in addition to transferring style they effectively preserve meaning of input sentences.

This paper introduces a novel approach to transferring style of a sentence while better preserving its meaning. We hypothesize—relying on the study of Rabinovich et al. (2016) who showed that author characteristics are significantly obfuscated by both manual and automatic machine translation—that grounding in back-translation is a plausible approach to rephrase a sentence while reducing its stylistic properties. We thus first use back-translation to rephrase the sentence and reduce the effect of the original style; then, we generate from the latent representation, using separate style-specific generators controlling for style (§2).

We focus on transferring author attributes: (1) gender and (2) political slant, and (3) on sentiment modification. The second task is novel: given a sentence by an author with a particular political leaning, rephrase the sentence to preserve its meaning but to confound classifiers of political slant (§3). The task of sentiment modification enables us to compare our approach with state-of-

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 866–876
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

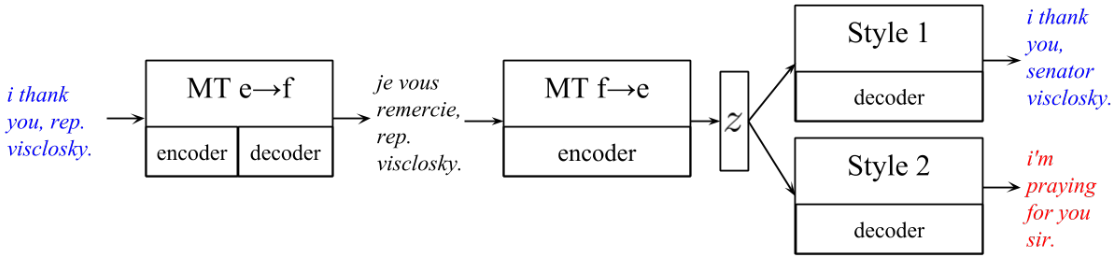
i thank you, rep. visclosky.
MT e→f
encoder
decoder
je vous remercie, rep. visclosky.
MT f→e
encoder
z
Style 1
decoder
Style 2
decoder
i thank you, senator visclosky.
i'm praying for you sir.

Figure 1: Style transfer pipeline: to rephrase a sentence and reduce its stylistic characteristics, the sentence is back-translated. Then, separate style-specific generators are used for style transfer.

the-art models (Hu et al., 2017; Shen et al., 2017).

Style transfer is evaluated using style classifiers trained on held-out data. Our back-translation style transfer model outperforms the state-of-the-art baselines (Shen et al., 2017; Hu et al., 2017) on the tasks of political slant and sentiment modification; 12% absolute improvement was attained for political slant transfer, and up to 7% absolute improvement in modification of sentiment (§5). Meaning preservation was evaluated manually, using A/B testing (§4). Our approach performs better than the baseline on the task of transferring gender and political slant. Finally, we evaluate the fluency of the generated sentences using human evaluation and our model outperforms the baseline in all experiments for fluency.

The main contribution of this work is a new approach to style transfer that outperforms state-of-the-art baselines in both the quality of input–output correspondence (meaning preservation and fluency), and the accuracy of style transfer. The secondary contribution is a new task that we propose to evaluate style transfer: transferring political slant.

2 Methodology

Given two datasets $\boldsymbol{X}_1 = \{\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_1^{(n)}\}$ and $\boldsymbol{X}_2 = \{\boldsymbol{x}_2^{(1)}, \ldots, \boldsymbol{x}_2^{(n)}\}$ which represent two different styles $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, respectively, our task is to generate sentences of the desired style while preserving the meaning of the input sentence. Specifically, we generate samples of dataset $\boldsymbol{X}_1$ such that they belong to style $\boldsymbol{s}_2$ and samples of $\boldsymbol{X}_2$ such that they belong to style $\boldsymbol{s}_1$. We denote the output of dataset $\boldsymbol{X}_1$ transfered to style $\boldsymbol{s}_2$ as $\hat{\boldsymbol{X}}_1 = \{\hat{\boldsymbol{x}}_2^{(1)}, \ldots, \hat{\boldsymbol{x}}_2^{(n)}\}$ and the output of dataset $\boldsymbol{X}_2$ transferred to style $\boldsymbol{s}_1$ as $\hat{\boldsymbol{X}}_2 = \{\hat{\boldsymbol{x}}_1^{(1)}, \ldots, \hat{\boldsymbol{x}}_1^{(n)}\}$.

Hu et al. (2017) and Shen et al. (2017) introduced state-of-the-art style transfer models that use variational auto-encoders (Kingma and Welling, 2014, VAEs) and cross-aligned auto-encoders, respectively, to model a latent content variable $\boldsymbol{z}$. The latent content variable $\boldsymbol{z}$ is a code which is not observed. The generative model conditions on this code during the generation process. Our aim is to design a latent code $\boldsymbol{z}$ which (1) represents the meaning of the input sentence grounded in back-translation and (2) weakens the style attributes of author's traits. To model the former, we use neural machine translation. Prior work has shown that the process of translating a sentence from a source language to a target language retains the meaning of the sentence but does not preserve the stylistic features related to the author's traits (Rabinovich et al., 2016). We hypothesize that a latent code $\boldsymbol{z}$ obtained through back-translation will normalize the sentence and devoid it from style attributes specific to author's traits.

Figure 1 shows the overview of the proposed method. In our framework, we first train a machine translation model from source language e to a target language f. We also train a back-translation model from f to e. Let us assume our styles $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ correspond to DEMOCRATIC and REPUBLICAN style, respectively. In Figure 1, the input sentence *i thank you, rep. visclosky.* is labeled as DEMOCRATIC. We translate the sentence using the $e \rightarrow f$ machine translation model and generate the parallel sentence in the target language f: *je vous remercie, rep. visclosky.* Using the fixed encoder of the $f \rightarrow e$ machine translation model, we encode this sentence in language f. The hidden representation created by this encoder of the back-translation model is used as $\boldsymbol{z}$. We condition our generative models on this $\boldsymbol{z}$. We then train two separate decoders for each style $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ to generate samples in these respective styles in source language e. Hence the sentence could be translated to the REPUBLICAN style using the decoder for $\boldsymbol{s}_2$. For example, the sentence *i'm praying for you sir.* is the REPUBLICAN ver-

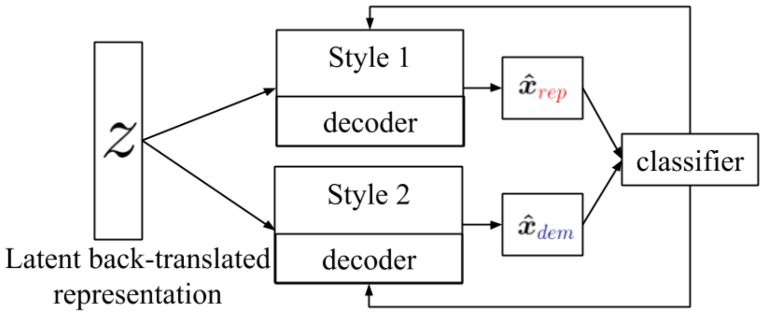

Figure 2: The latent representation from back-translation and the style classifier feedback are used to guide the style-specific generators.

sion of the input sentence and *i thank you, senator visclosky.* is the more DEMOCRATIC version of it.

Note that in this setting, the machine translation and the encoder of the back-translation model remain fixed. They are not dependent on the data we use across different tasks. This facilitates reusability and spares the need of learning separate models to generate $\boldsymbol{z}$ for a new style data.

2.1 Meaning-Grounded Representation

In this section we describe how we learn the latent content variable $\boldsymbol{z}$ using back-translation. The $e \rightarrow f$ machine translation and $f \rightarrow e$ back-translation models are trained using a sequence-to-sequence framework (Sutskever et al., 2014; Bahdanau et al., 2015) with style-agnostic corpus. The style-specific sentence *i thank you, rep. visclosky.* in source language e is translated to the target language f to get *je vous remercie, rep. visclosky.* The individual tokens of this sentence are then encoded using the encoder of the $f \rightarrow e$ back-translation model. The learned hidden representation is $\boldsymbol{z}$.

Formally, let $\boldsymbol{\theta}_E$ represent the parameters of the encoder of $f \rightarrow e$ translation system. Then $\boldsymbol{z}$ is given by:

$$\boldsymbol{z} = Encoder(\boldsymbol{x}_f; \boldsymbol{\theta}_E) \tag{1}$$

where, $\boldsymbol{x}_f$ is the sentence $\boldsymbol{x}$ in language f. Specifically, $\boldsymbol{x}_f$ is the output of $e \rightarrow f$ translation system when $\boldsymbol{x}_e$ is given as input. Since $\boldsymbol{z}$ is derived from a non-style specific process, this Encoder is not style specific.

2.2 Style-Specific Generation

Figure 2 shows the architecture of the generative model for generating different styles. Using the encoder embedding $\boldsymbol{z}$, we train multiple decoders for each style. The sentence generated by a decoder is passed through the classifier. The loss of the classifier for the generated sentence is used as feedback to guide the decoder for the generation process. The target attribute of the classifier is determined by the decoder from which the output is generated. For example, in the case of DEMOCRATIC decoder, the target attribute is DEMOCRATIC and for the REPUBLICAN decoder the target is REPUBLICAN.

2.2.1 Style Classifiers

We train a convolutional neural network (CNN) classifier to accurately predict the given style. We also use it to evaluate the error in the generated samples for the desired style. We train the classifier in a supervised manner. The classifier accepts either discrete or continuous tokens as inputs. This is done such that the generator output can be used as input to the classifier. We need labeled examples to train the classifier such that each instance in the dataset $\boldsymbol{X}$ should have a label in the set $\boldsymbol{s} = \{\boldsymbol{s}_1, \boldsymbol{s}_2\}$. Let $\boldsymbol{\theta}_C$ denote the parameters of the classifier. The objective to train the classifier is given by:

$$\mathcal{L}_{class}(\boldsymbol{\theta}_C) = \mathbb{E}_{\boldsymbol{X}}[\log q_C(\boldsymbol{s}|\boldsymbol{x})]. \tag{2}$$

To improve the accuracy of the classifier, we augment classifier's inputs with style-specific lexicons. We concatenate binary style indicators to each input word embedding in the classifier. The indicators are set to 1 if the input word is present in a style-specific lexicon; otherwise they are set to 0. Style lexicons are extracted using the log-odds ratio informative Dirichlet prior (Monroe et al., 2008), a method that identifies words that are statistically overrepresented in each of the categories.

2.2.2 Generator Learning

We use a bidirectional LSTM to build our decoders which generate the sequence of tokens $\hat{\boldsymbol{x}} = \{x_1, \cdots x_T\}$. The sequence $\hat{\boldsymbol{x}}$ is conditioned on the latent code $\boldsymbol{z}$ (in our case, on the machine translation model). In this work we use a corpus translated to French by the machine translation system as the input to the encoder of the back-translation model. The same encoder is used to encode sentences of both styles. The representation created by this encoder is given by Eq 1. Samples are generated as follows:

$$\begin{aligned} \hat{\boldsymbol{x}} \sim \boldsymbol{z} &= p(\hat{\boldsymbol{x}}|\boldsymbol{z}) && (3) \\ &= \prod_t p(\hat{x}_t|\hat{\boldsymbol{x}}^{<t}, \boldsymbol{z}) && (4) \end{aligned}$$

where, $\hat{\boldsymbol{x}}^{<t}$ are the tokens generated before $\hat{x}_t$.

Tokens are discrete and non-differentiable. This makes it difficult to use a classifier, as the generation process samples discrete tokens from the multinomial distribution parametrized using softmax function at each time step t. This non-differentiability, in turn, breaks down gradient propagation from the discriminators to the generator. Instead, following Hu et al. (2017) we use a continuous approximation based on softmax, along with the temperature parameter which anneals the softmax to the discrete case as training proceeds. To create a continuous representation of the output of the generative model which will be given as an input to the classifier, we use:

$$\hat{x}_t \sim \text{softmax}(\boldsymbol{o}_t/\tau),$$

where, $\boldsymbol{o}_t$ is the output of the generator and τ is the temperature which decreases as the training proceeds. Let $\boldsymbol{\theta}_G$ denote the parameters of the generators. Then the reconstruction loss is calculated using the cross entropy function, given by:

$$\mathcal{L}_{recon}(\boldsymbol{\theta}_G; \boldsymbol{x}) = \mathbb{E}_{q_E(\boldsymbol{z}|\boldsymbol{x})}[\log p_{gen}(\boldsymbol{x}|\boldsymbol{z})] \quad (5)$$

Here, the back-translation encoder E creates the latent code $\boldsymbol{z}$ by:

$$\boldsymbol{z} = E(\boldsymbol{x}) = q_E(\boldsymbol{z}|\boldsymbol{x}) \quad (6)$$

The generative loss $\mathcal{L}_{gen}$ is then given by:

$$\min_{\theta_{gen}} \mathcal{L}_{gen} = \mathcal{L}_{recon} + \lambda_c \mathcal{L}_{class} \quad (7)$$

where $\mathcal{L}_{recon}$ is given by Eq. (5), $\mathcal{L}_{class}$ is given by Eq (2) and λ_c is a balancing parameter.

We also use global attention of (Luong et al., 2015) to aid our generators. At each time step t of the generation process, we infer a variable length alignment vector $\boldsymbol{a}_t$:

$$\boldsymbol{a}_t = \frac{\exp(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s))}{\sum_{s'} \exp(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'}))} \quad (8)$$

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \textbf{dot}(\boldsymbol{h}_t^T, \bar{\boldsymbol{h}}_s), \quad (9)$$

where $\boldsymbol{h}_t$ is the current target state and $\bar{\boldsymbol{h}}_s$ are all source states. While generating sentences, we use the attention vector to replace unknown characters (UNK) using the copy mechanism in (See et al., 2017).

3 Style Transfer Tasks

Much work in computational social science has shown that people's personal and demographic characteristics—either publicly observable (e.g., age, gender) or private (e.g., religion, political affiliation)—are revealed in their linguistic choices (Nguyen et al., 2016). There are practical scenarios, however, when these attributes need to be modulated or obfuscated. For example, some users may wish to preserve their anonymity online, for personal security concerns (Jardine, 2016), or to reduce stereotype threat (Spencer et al., 1999). Modulating authors' attributes while preserving meaning of sentences can also help generate demographically-balanced training data for a variety of downstream applications.

Moreover, prior work has shown that the quality of language identification and POS tagging degrades significantly on African American Vernacular English (Blodgett et al., 2016; Jørgensen et al., 2015); YouTube's automatic captions have higher error rates for women and speakers from Scotland (Rudinger et al., 2017). Synthesizing balanced training data—using style transfer techniques—is a plausible way to alleviate bias present in existing NLP technologies.

We thus focus on two tasks that have practical and social-good applications, and also accurate style classifiers. To position our method with respect to prior work, we employ a third task of sentiment transfer, which was used in two state-of-the-art approaches to style transfer (Hu et al., 2017; Shen et al., 2017). We describe the three tasks and associated dataset statistics below. The methodology that we advocate is general and can be applied to other styles, for transferring various social categories, types of bias, and in multi-class settings.

Gender. In sociolinguistics, gender is known to be one of the most important social categories driving language choice (Eckert and McConnell-Ginet, 2003; Lakoff and Bucholtz, 2004; Coates, 2015). Reddy and Knight (2016) proposed a heuristic-based method to obfuscate gender of a writer. This method uses statistical association measures to identify gender-salient words and substitute them with synonyms typically of the opposite gender. This simple approach produces highly fluent, meaning-preserving sentences, but does not allow for more general rephrasing of sentence beyond single-word substitutions. In our work, we adopt this task of transferring the author's gender and adapt it to our experimental settings.

We used Reddy and Knight's (2016) dataset of reviews from Yelp annotated for two genders corresponding to markers of sex.[1] We split the reviews to sentences, preserving the original gender labels. To keep only sentences that are strongly indicative of a gender, we then filtered out gender-neutral sentences (e.g., *thank you*) and sentences whose likelihood to be written by authors of one gender is lower than 0.7.[2]

Political slant. Our second dataset is comprised of top-level comments on Facebook posts from all 412 current members of the United States Senate and House who have public Facebook pages (Voigt et al., 2018).[3] Only top-level comments that directly respond to the post are included. Every comment to a Congressperson is labeled with the Congressperson's party affiliation: democratic or republican. Topic and sentiment in these comments reveal commenter's political slant. For example, *defund them all, especially when it comes to the illegal immigrants .* and *thank u james, praying for all the work u do .* are republican, whereas *on behalf of the hard-working nh public school teachers- thank you !* and *we need more strong voices like yours fighting for gun control .* represent examples of democratic sentences. Our task is to preserve intent of the commenter (e.g., to thank their representative), but to modify their observable political affiliation, as in the example in Figure 1. We preprocessed and filtered the comments similarly to the gender-annotated corpus above.

Style	*class*	*train*	*dev*	*test*
gender	2.57M	2.67M	4.5K	535K
political	80K	540K	4K	56K
sentiment	2M	444K	63.5K	127K

Table 1: Sentence count in style-specific corpora.

Sentiment. To compare our work with the state-of-the-art approaches of style transfer for non-parallel corpus we perform sentiment transfer, replicating the models and experimental setups of Hu et al. (2017) and Shen et al. (2017). Given a positive Yelp review, a style transfer model will generate a similar review but with an opposite sentiment. We used Shen et al.'s (2017) corpus of reviews from Yelp. They have followed the standard practice of labeling the reviews with rating of higher than three as positive and less than three as negative. They have also split the reviews to sentences and assumed that the sentence has the same sentiment as the review.

Dataset statistics. We summarize below corpora statistics for the three tasks: transferring gender, political slant, and sentiment. The dataset for sentiment modification task was used as described in (Shen et al., 2017). We split Yelp and Facebook corpora into four disjoint parts each: (1) a training corpus for training a style classifier (*class*); (2) a training corpus (*train*) used for training the style-specific generative model described in §2.2; (3) development and (4) test sets. We have removed from training corpora *class* and *train* all sentences that overlap with development and test corpora. Corpora sizes are shown in Table 1.

Table 2 shows the approximate vocabulary sizes used for each dataset. The vocabulary is the same for both the styles in each experiment.

Style	gender	political	sentiment
Vocabulary	20K	20K	10K

Table 2: Vocabulary sizes of the datasets.

Table 3 summarizes sentence statistics. All the

[1] We note that gender may be considered along a spectrum (Eckert and McConnell-Ginet, 2003), but use gender as a binary variable due to the absence of corpora with continuous-valued gender annotations.

[2] We did not experiment with other threshold values.

[3] The posts and comments are all public; however, to protect the identity of Facebook users in this dataset Voigt et al. (2018) have removed all identifying user information as well as Facebook-internal information such as User IDs and Post IDs, replacing these with randomized ID numbers.

sentences have maximum length of 50 tokens.

Style	Avg. Length	%data
male	18.08	50.00
female	18.21	50.00
republican	16.18	50.00
democratic	16.01	50.00
negative	9.66	39.81
positive	8.45	60.19

Table 3: Average sentence length and class distribution of style corpora.

4 Experimental Setup

In what follows, we describe our experimental settings, including baselines used, hyperparameter settings, datasets, and evaluation setups.

Baseline. We compare our model against the "cross-aligned" auto-encoder (Shen et al., 2017), which uses style-specific decoders to align the style of generated sentences to the actual distribution of the style. We used the off-the-shelf sentiment model released by Shen et al. (2017) for the sentiment experiments. We also separately train this model for the gender and political slant using hyper-parameters detailed below.[4]

Translation data. We trained an English–French neural machine translation system and a French–English back-translation system. We used data from Workshop in Statistical Machine Translation 2015 (WMT15) (Bojar et al., 2015) to train our translation models. We used the French–English data from the Europarl v7 corpus, the news commentary v10 corpus and the common crawl corpus from WMT15. Data were tokenized using the Moses tokenizer (Koehn et al., 2007). Approximately 5.4M English–French parallel sentences were used for training. A vocabulary size of 100K was used to train the translation systems.

Hyperparameter settings. In all the experiments, the generator and the encoders are a two-layer bidirectional LSTM with an input size of 300 and the hidden dimension of 500. The generator samples a sentence of maximum length 50. All the generators use global attention vectors of size 500. The CNN classifier is trained with 100 filters of size 5, with max-pooling. The input to CNN is of size 302: the 300-dimensional word embedding plus two bits for membership of the word in our style lexicons, as described in §2.2.1. Balancing parameter λ_c is set to 15. For sentiment task, we have used settings provided in (Shen et al., 2017).

5 Results

We evaluate our approach along three dimensions. (1) Style transfer accuracy, measuring the proportion of our models' outputs that generate sentences of the desired style. The style transfer accuracy is performed using classifiers trained on held-out train data that were not used in training the style transfer models. (2) Preservation of meaning. (3) Fluency, measuring the readability and the naturalness of the generated sentences. We conducted human evaluations for the latter two.

In what follows, we first present the quality of our neural machine translation systems, then we present the evaluation setups, and then present the results of our experiments.

Translation quality. The BLEU scores achieved for English–French MT system is 32.52 and for French–English MT system is 31.11; these are strong translation systems. We deliberately chose a European language close to English for which massive amounts of parallel data are available and translation quality is high, to concentrate on the style generation, rather than improving a translation system. [5]

5.1 Style Transfer Accuracy

We measure the accuracy of style transfer for the generated sentences using a pre-trained style classifier (§2.2.1). The classifier is trained on data that is not used for training our style transfer generative models (as described in §3). The classifier has an accuracy of 82% for the gender-annotated corpus, 92% accuracy for the political slant dataset and 93.23% accuracy for the sentiment dataset.

[4] In addition, we compared our model with the current state-of-the-art approach introduced by Hu et al. (2017); Shen et al. (2017) use this method as baseline, obtaining comparable results. We reproduced the results reported in (Hu et al., 2017) using their tasks and data. However, the same model trained on our political slant datasets (described in §3), obtained an almost random accuracy of 50.98% in style transfer. We thus omit these results.

[5] Alternatively, we could use a pivot language that is typologically more distant from English, e.g., Chinese. In this case we hypothesize that stylistic traits would be even less preserved in translation, but the quality of back-translated sentences would be worse. We have not yet investigated how the accuracy of the translation model, nor the language of translation affects our models.

We transfer the style of test sentences and then test the classification accuracy of the generated sentences for the opposite label. For example, if we want to transfer the style of male Yelp reviews to female, then we use the fixed common encoder of the back-translation model to encode the test male sentences and then we use the female generative model to generate the female-styled reviews. We then test these generated sentences for the *female* label using the gender classifier.

Experiment	CAE	BST
Gender	**60.40**	57.04
Political slant	75.82	**88.01**
Sentiment	80.43	**87.22**

Table 4: Accuracy of the style transfer in generated sentences.

In Table 4, we detail the accuracy of each classifier on generated style-transfered sentences.[6] We denote the Shen et al.'s (2017) **C**ross-aligned **A**uto-**E**ncoder model as CAE and our model as **B**ack-translation for **S**tyle **T**ransfer (BST).

On two out of three tasks our model substantially outperforms the baseline, by up to 12% in political slant transfer, and by up to 7% in sentiment modification.

5.2 Preservation of Meaning

Although we attempted to use automatics measures to evaluate how well meaning is preserved in our transformations; measures such as BLEU (Papineni et al., 2002) and Meteor (Denkowski and Lavie, 2011), or even cosine similarity between distributed representations of sentences do not capture this distance well.

Meaning preservation in style transfer is not trivial to define as literal meaning is likely to change when style transfer occurs. For example "My girlfriend loved the desserts" vs "My partner liked the desserts". Thus we must relax the condition of literal meaning to *intent* or *affect* of the utterance within the context of the discourse. Thus if the intent is to criticize a restaurant's service in a review, changing "salad" to "chicken" could still have the same effect but if the intent is to order food that substitution would not be acceptable. Ideally we wish to evaluate transfer within some downstream task and ensure that the task has the same outcome even after style transfer. This is a hard evaluation and hence we resort to a simpler evaluation of the "meaning" of the sentence.

Experiment	CAE	No Pref.	BST
Gender	15.23	41.36	**43.41**
Political slant	14.55	**45.90**	39.55
Sentiment	35.91	**40.91**	23.18

Table 5: Human preference for meaning preservation in percentages.

We set up a manual pairwise comparison following Bennett (2005). The test presents the original sentence and then, in random order, its corresponding sentences produced by the baseline and our models. For the gender style transfer we asked "Which transferred sentence maintains the same sentiment of the source sentence in the same semantic context (i.e. you can ignore if food items are changed)". For the task of changing the political slant, we asked "Which transferred sentence maintains the same semantic intent of the source sentence while changing the political position". For the task of sentiment transfer we have followed the annotation instruction in (Shen et al., 2017) and asked "Which transferred sentence is semantically equivalent to the source sentence with an opposite sentiment"

We then count the preferences of the eleven participants, measuring the relative acceptance of the generated sentences.[7] A third option "=" was given to participants to mark no preference for either of the generated sentence. The "no preference" option includes choices both are equally bad and both are equally good. We conducted three tests one for each type of experiment - gender, political slant and sentiment. We also divided our annotation set into short (#tokens $\leq$ 15) and long ($15 <$ #tokens ≤ 30) sentences for the gender and the political slant experiment. In each set we had 20 random samples for each type of style transfer. In total we had 100 sentences to be annotated. Note that we did not ask about appropriateness of the style transfer in this test, or fluency of outputs, only about meaning preservation.

The results of human evaluation are presented in Table 5. Although a no-preference option was chosen often—showing that state-of-the-art systems are still not on par with hu-

[6] In each experiment, we report aggregated results across directions of style transfer; same results broke-down to style categories are listed in the Supplementary Material.

[7] None of the human judges are authors of this paper

man expectations—the BST models outperform the baselines in the gender and the political slant transfer tasks.

Crucially, the BST models significantly outperform the CAE models when transferring style in longer and harder sentences. Annotators preferred the CAE model only for 12.5% of the long sentences, compared to 47.27% preference for the BST model.

5.3 Fluency

Finally, we evaluate the fluency of the generated sentences. Fluency was rated from 1 (unreadable) to 4 (perfect) as is described in (Shen et al., 2017). We randomly selected 60 sentences each generated by the baseline and the BST model.

The results shown in Table 6 are averaged scores for each model.

Experiment	CAE	BST
Gender	2.42	**2.81**
Political slant	2.79	**2.87**
Sentiment	3.09	**3.18**
Overall	2.70	**2.91**
Overall Short	3.05	**3.11**
Overall Long	2.18	**2.62**

Table 6: Fluency of the generated sentences.

BST outperforms the baseline overall. It is interesting to note that BST generates significantly more fluent longer sentences than the baseline model. Since the average length of sentences was higher for the gender experiment, BST notably outperformed the baseline in this task, relatively to the sentiment task where the sentences are shorter. Examples of the original and style-transfered sentences generated by the baseline and our model are shown in the Supplementary Material.

5.4 Discussion

The loss function of the generators given in Eq. 5 includes two competing terms, one to improve meaning preservation and the other to improve the style transfer accuracy. In the task of sentiment modification, the BST model preserved meaning worse than the baseline, on the expense of being better at style transfer. We note, however, that the sentiment modification task is not particularly well-suited for evaluating style transfer: it is particularly hard (if not impossible) to disentangle the sentiment of a sentence from its propositional content, and to modify sentiment while preserving meaning or intent. On the other hand, the style-transfer accuracy for gender is lower for BST model but the preservation of meaning is much better for the BST model, compared to CAE model and to ”No preference” option. This means that the BST model does better job at closely representing the input sentence while taking a mild hit in the style transfer accuracy.

6 Related Work

Style transfer with non-parallel text corpus has become an active research area due to the recent advances in text generation tasks. Hu et al. (2017) use variational auto-encoders with a discriminator to generate sentences with controllable attributes. The method learns a disentangled latent representation and generates a sentence from it using a code. This paper mainly focuses on sentiment and tense for style transfer attributes. It evaluates the transfer strength of the generated sentences but does not evaluate the extent of preservation of meaning in the generated sentences. In our work, we show a qualitative evaluation of meaning preservation.

Shen et al. (2017) first present a theoretical analysis of style transfer in text using non-parallel corpus. The paper then proposes a novel cross-alignment auto-encoders with discriminators architecture to generate sentences. It mainly focuses on sentiment and word decipherment for style transfer experiments.

Fu et al. (2018) explore two models for style transfer. The first approach uses multiple decoders for each type of style. In the second approach, style embeddings are used to augment the encoded representations, so that only one decoder needs to be learned to generate outputs in different styles. Style transfer is evaluated on scientific paper titles and newspaper tiles, and sentiment in reviews. This method is different from ours in that we use machine translation to create a strong latent state from which multiple decoders can be trained for each style. We also propose a different human evaluation scheme.

Li et al. (2018) first extract words or phrases associated with the original style of the sentence, delete them from the original sentence and then replace them with new phrases associated with the target style. They then use a neural model to fluently combine these into a final output. Junbo et al. (2017) learn a representation which is style-agnostic, using adversarial training of the auto-encoder.

Our work is also closely-related to a problem of paraphrase generation (Madnani and Dorr, 2010; Dong et al., 2017), including methods relying on (phrase-based) back-translation (Ganitkevitch et al., 2011; Ganitkevitch and Callison-Burch, 2014). More recently, Mallinson et al. (2017) and Wieting et al. (2017) showed how neural back-translation can be used to generate paraphrases. An additional related line of research is machine translation with non-parallel data. Lample et al. (2018) and Artetxe et al. (2018) have proposed sophisticated methods for unsupervised machine translation. These methods could in principle be used for style transfer as well.

7 Conclusion

We propose a novel approach to the task of style transfer with non-parallel text.[8] We learn a latent content representation using machine translation techniques; this aids grounding the meaning of the sentences, as well as weakening the style attributes. We apply this technique to three different style transfer tasks. In transfer of political slant and sentiment we outperform an off-the-shelf state-of-the-art baseline using a cross-aligned autoencoder. The political slant task is a novel task that we introduce. Our model also outperforms the baseline in all the experiments of fluency, and in the experiments for meaning preservation in generated sentences of gender and political slant. Yet, we acknowledge that the generated sentences do not always adequately preserve meaning.

This technique is suitable not just for style transfer, but for enforcing style, and removing style too. In future work we intend to apply this technique to *debiasing* sentences and *anonymization* of author traits such as gender and age.

In the future work, we will also explore whether an enhanced back-translation by pivoting through several languages will learn better grounded latent meaning representations. In particular, it would be interesting to back-translate through multiple target languages with a single source language (Johnson et al., 2016).

Measuring the separation of style from content is hard, even for humans. It depends on the task and the context of the utterance within its discourse. Ultimately we must evaluate our style transfer within some down-stream task where our style transfer has its intended use but we achieve the same task completion criteria.

Acknowledgments

This work was funded by a fellowship from Robert Bosch, and in part by the National Science Foundation through award IIS-1526745. We would like to thank Sravana Reddy for sharing the Yelp corpus used in gender transfer experiments, Zhiting Hu for providing an implementation of a VAE-based baseline, and the 11 CMU graduate students who helped with annotation and manual evaluations. We are also grateful to the anonymous reviewers for their constructive feedback, and to Dan Jurafsky, David Jurgens, Vinod Prabhakaran, and Rob Voigt for valuable discussions at earlier stages of this work.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proc ICLR*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. ICLR*.

Christina L Bennett. 2005. Large scale evaluation of corpus-based synthesizers: Results and lessons from the blizzard challenge 2005. In *Ninth European Conference on Speech Communication and Technology*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proc. EMNLP*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proc. WMT*, pages 1–46.

Jennifer Coates. 2015. *Women, men and language: A sociolinguistic account of gender differences in language*. Routledge.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and

[8] All the code and data used in the experiments will be released to facilitate reproducibility at https://github.com/shrimai/Style-Transfer-Through-Back-Translation

evaluation of machine translation systems. In *Proc. WMT*, pages 85–91.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886. Association for Computational Linguistics.

Penelope Eckert and Sally McConnell-Ginet. 2003. *Language and gender*. Cambridge University Press.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style Transfer in Text: Exploration and Evaluation. In *Proc. AAAI*.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In *Proc. LREC*, pages 4276–4283.

Juri Ganitkevitch, Chris Callison-Burch, Courtney Napoles, and Benjamin Van Durme. 2011. Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proc. EMNLP*, pages 1168–1179.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proc. ICML*, pages 1587–1596.

Eric Jardine. 2016. Tor, what is it good for? political repression and the use of online anonymity-granting technologies. *New Media & Society*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proc. of the Workshop on Noisy User-generated Text*, pages 9–18.

Junbo, Zhao, Y. Kim, K. Zhang, A. M. Rush, and Y. LeCun. 2017. Adversarially Regularized Autoencoders for Generating Discrete Structures. *ArXiv e-prints*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*, pages 3128–3137.

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *Proc. EMNLP*, pages 329–339.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL (demonstration sessions)*, pages 177–180.

Robin Tolmach Lakoff and Mary Bucholtz. 2004. *Language and woman's place: Text and commentaries*, volume 3. Oxford University Press, USA.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proc. ICLR*.

J. Li, R. Jia, H. He, and P. Liang. 2018. Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer. *ArXiv e-prints*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proc. ACL*.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proc. EMNLP*.

Nitin Madnani and Bonnie J Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proce. EACL*, volume 1, pages 881–893.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proc. EMNLP*, pages 2804–2809.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. 2016. Personalized machine translation: Preserving original author traits. In *Proc. EACL*.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proc. of Workshop on Natural Language Processing and Computational Social Science*, pages 17–26.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proc. EMNLP*, pages 583–593.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proc. of the First Workshop on Ethics in Natural Language Processing*, page 74.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proc. ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proc. NAACL*, pages 35–40.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proc. NIPS*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proc. NAACL*.

Steven J. Spencer, Claude M. Steele, and Diane M. Quinn. 1999. Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, 35:4–28.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proc. NIPS*, pages 3104–3112.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proc. ICML Deep Learning Workshop*.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proc. LREC*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proc. EACL*.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proc. EMNLP*, pages 274–285.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*, pages 2048–2057.