Multi-Style Transfer with Discriminative Feedback on Disjoint Corpus

Navita Goyal, Balaji Vasan Srinivasan, Anandhavelu N, Abhilasha Sancheti
Adobe Research, India
{navgoyal, balsrini, anandvn, sancheti}@adobe.com

Abstract

Style transfer has been widely explored in natural language generation with non-parallel corpus by directly or indirectly extracting a notion of style from source and target domain corpus. A common shortcoming of existing approaches is the prerequisite of joint annotations across all the stylistic dimensions under consideration. Availability of such dataset across a combination of styles limits the extension of these setups to multiple style dimensions. While cascading single-dimensional models across multiple styles is a possibility, it suffers from content loss, especially when the style dimensions are not completely independent of each other. In our work, we relax this requirement of jointly annotated data across multiple styles by using independently acquired data across different style dimensions without any additional annotations. We initialize an encoder-decoder setup with transformer-based language model pre-trained on a generic corpus and enhance its re-writing capability to multiple target style dimensions by employing multiple style-aware language models as discriminators. Through quantitative and qualitative evaluation, we show the ability of our model to control styles across multiple style dimensions while preserving content of the input text. We compare it against baselines involving cascaded state-of-the-art uni-dimensional style transfer models.

1 Introduction

Style transfer is a popular task in natural language processing and has been studied on attributes like age or gender (Subramanian et al., 2018), styles emanating from social construct like formality (Rao and Tetreault, 2018) and politeness (Madaan et al., 2020), linguistic styles based on author writing style (Syed et al., 2020), or psycho-linguistic styles based on personality types (Mairesse and Walker, 2011). While early style transfer frameworks were modeled as a supervised learning task on a parallel corpus, state-of-the-art models are semi-supervised/unsupervised and operate on non-parallel corpus. These models achieve style transfer by aligning source and target distribution of sentences from non-parallel corpus (Shen et al., 2017), disentangling content space from style space in latent representation (Hu et al., 2017) or employing self-reconstruction (Dai et al., 2019) and back translation (Lample et al., 2018) objectives to achieve pseudo-supervision with non-parallel corpus. Recent works have also modeled this in a self-supervised manner where rewriting (transfer) is achieved by utilizing corpus from the target style alone (Syed et al., 2020). These wide studies have also led to the curation and benchmarking of non-parallel dataset for various style dimensions, such as sentiment (Li et al., 2018), formality (Rao and Tetreault, 2018), politeness (Danescu-Niculescu-Mizil et al., 2013), excitement (Sancheti et al., 2020), etc. But availability of data with joint tagging across multiple styles is limited and has restricted the ability of existing approaches to scale from single-dimensional transfer to multiple style dimensions. In this paper, we propose a multi-dimensional style transfer approach that can work off partially labelled data for style transfer across multiple dimensions simultaneously.

The work by Subramanian et al. (2018) attempts style transfer with multiple attributes such as age, gender, and sentiment simultaneously. However, their approach avails corpus tagged with each of these three style dimensions. In contrast to this and other similar explorations in multi-style transfer, our approach does not require jointly labelled data across all the stylistic dimensions in source and/or target corpus. We focus on the problem where independent corpus is available across different stylistic dimensions (say *sentiment* and *formality*) and we achieve style transfer spanning different stylistic dimensions (say make a sentence more *positive* and *formal*). While state-of-the-art approaches can be

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3500–3510
June 6–11, 2021.

extended to achieve this by sequentially transferring one style after another, it is limited as different style dimensions are not necessarily independent of each other. In aspects that are not independent, changing one style aspect of the text might affect another aspect considered, making a sequential brute-force approach non-ideal. As we show in our experiments later, the cascaded setup also lacks common grounding between the content from different styles leading to erratic changes in content. We circumvent this by grounding our framework on the linguistic understanding of a large language model. Our model builds understanding of interplay between the different styles by incorporating multiple discriminative language models (LM) with language model-based encoder-decoder setup. The key contributions of this paper are:

1) An encoder-decoder setup with multiple language models as discriminator, with each entity harnessing the language understanding from a large pre-trained transformer model.

2) Relaxing the requirement of jointly labelled data for multi-style transfer, by leveraging independently acquired disjoint corpus for different styles.

3) Achieving better style control with better content preservation in multi-dimensional style transfer than a cascaded setup of state-of-the-art uni-dimensional style transfer models.

2 Related Work

One line of work in **style transfer** attempts to learn disentangled latent representation for style and content, and transfer style by manipulating latent representation of style (Shen et al., 2017). Although these approaches perform well with one style at a time, they do not trivially scale to multi-dimensional style transfer. Several other works develop unsupervised approach for style transfer by employing Denoising Autoencoding (DAE) (Fu et al., 2017) and back-translation (BT) (Lample et al., 2018) loss to develop interaction and hence transfer between the source and target domain. Subramanian et al. (2018) extend this approach to multiple styles by conditioning on average of embedding of each target attribute and using combination of DAE and back-translation techniques. DAE takes as input a sentence x from style s and tries to reconstruct sentence x from its corrupted version $\tilde{x}$. This relies on the assumption that the input sentence x is from a certain style combination $s = \{s_1, s_2, \ldots, s_k\}$. Similarly back translation (BT) objective with input sentence x from style s, first estimates $x' = f(x, s')$, where $s \neq s'$ and then reconstruct x from $\tilde{x} = f(x', s)$. Thus, these approaches are inherently dependent on knowledge of annotation of each sentence with all the style combinations. Dai et al. (2019) achieve state-of-the-art style transfer in single style dimensions by employing transformer-based model in conjunction with classifier-based discriminator. In addition to discriminator losses, their proposed technique uses self-reconstruction and cycle reconstruction losses, which similar to DAE and BT losses are also reliant on availability of jointly annotated data to be extendable to multiple style setup.

Language modeling is integral to several natural language generation (NLG) tasks like text summarization, spelling correction, image captioning, etc. The model architecture for these tasks has evolved from n-gram based methods to Recurrent Neural Networks to transformer architectures. The introduction of Transformer-based architecture accompanied with generative pre-training (Radford, 2018) capabilities have led to strong improvements in many downstream generation and GLUE (Wang et al., 2018) tasks. Generative pre-training aims to adapt a large Transformer language model to large unsupervised corpus. This capability of generative pre-training is exploited in many large language models like BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2018), ERNIE 2.0 (Sun et al., 2020) which have the ability to perform tasks like reading comprehension (Xu et al., 2019), summarization (Liu and Lapata, 2019), question-answering (Rajpurkar et al., 2016) and translation (Clinchant et al., 2019) in zero-shot and few-shot settings.

Recently these pre-trained generative language models have been explored in translation (Conneau and Lample, 2019) and style transfer tasks (Syed et al., 2020). Conneau and Lample (2019) develop cross-lingual models for unsupervised machine translation by initializing encoder and decoder with a pre-trained language model trained on Masked Language Modeling (MLM) (Devlin et al., 2019) objective and fine-tuning the encoder-decoder framework with adversarial training. Syed et al. (2020) extend this to stylized re-writing task by employing DAE during fine-tuning. The joint encoder-decoder framework learns to reconstruct sentences in target-domain from its noisy version using DAE objective. As previously discussed, the DAE objective is reliant on the corpus being tagged

for the target domain style (or combination of style) and restricts the generalization of this setup to multiple attributes. We overcome this by employing discriminative language models to assist the decoder with feedback for various target styles.

Shen et al. (2017) show that even with non-parallel data, the content distribution across source and target style is shared. Based on this, a language model trained on target style will have high perplexity on transferred text if it does not match target style and low perplexity otherwise. Yang et al. (2018) exploit this ability of language models to replace standard binary classifier-based discriminators with an implicitly trained language model as discriminator. They show that using the language model as structured discriminator allows for more stable training by eliminating the adversarial step. We extend this idea to a multi-discriminator approach. Training a LM on combination of target styles is not possible in absence of jointly labelled dataset. Due to this, we attempt to use multiple discriminators for each of the target styles. Since with multiple styles, the underlying corpus is independently acquired, the variation in content distribution across different styles is more noticeable. Consequently, an independently trained LM on one of the target styles might have high perplexity even if the transferred sentence fits in the corresponding target style, due to the content space of source sentence. To equip discriminative LM with more generalized notion of content, we use large transformer-based LM pre-trained on large unsupervised corpus to establish generic content distribution before style-oriented fine-tuning.

3 Approach

Our proposed approach has two key elements — a Transformer-based encoder-decoder model initialized with a pre-trained Transformer Language Model and fine-tuned on DAE loss to achieve style transfer (Section 3.1) and the multiple language models as discriminators stacked together to enable multi-style transfer (Section 3.2).

3.1 Pre-trained LM as Encoder-Decoder

Similar to Syed et al. (2020), we first pre-train a Transformer-based language model with Masked Language Modeling (MLM) objective on English Wikipedia data extracted using WikiExtractor.[1] This equips LM with the ability to predict masked words over a large corpus. Masked Language Modeling leverages bidirectional context of the input, thus enabling better language understanding. Following Masked Language Modeling objective from Devlin et al. (2019), we randomly sample 15% of the tokens from the text stream and replace them with the [MASK] token 80% of the time, by a random token 10% of the time and keep them unchanged 10% of the time, with the objective of predicting the original identity of the masked word based on its bidirectional context. To enable style transfer from a given sentence to target style, we use independently trained language models (LMs) to initialize the encoder and decoder and connect these with randomly initialized attention layers to arrive at a *encoder-decoder setup*. As discussed by Syed et al. (2020), the Transformer architecture (Vaswani et al., 2017) allows such independent initialization by implicitly aligning encoder-decoder layers via attention mechanism.

Pre-training an encoder only transformer on generative task and then leveraging it to initialize as both encoder and decoder as opposed to pre-training a joint encoder-decoder model has several advantages. Transformer-based models with encoder-only (Devlin et al., 2019) or decoder-only (Radford et al., 2018) blocks have been shown to perform well in generative pre-training task. Clearly, pre-training a single transformer block on generative task and then utilizing it as both encoder and decoder blocks has lower computational cost than training the entire encoder-decoder block jointly. Moreover, this also enables us to use the same pre-trained model to initialize both style transfer module and the discriminator models, explained in the following section. This is not only computationally more efficient but it also closely ties the underlying language distribution of the two modules. This is expected to make the discriminative feedback more effective while fine tuning the transfer model for multiple styles.

In Syed et al. (2020)'s setup, both encoder and decoder in the style transfer module are initialized with the pre-trained language model (trained on MLM objective). Instead, we initialize the decoder with the language model fine-tuned with the target style using Causal Language Modeling (CLM) objective, before training the joint encoder-decoder model, as detailed in Section 3.2. The encoder is initialized with the pre-trained model directly. Aligning the decoder to the distribution of the tar-

[1] https://github.com/attardi/wikiextractor

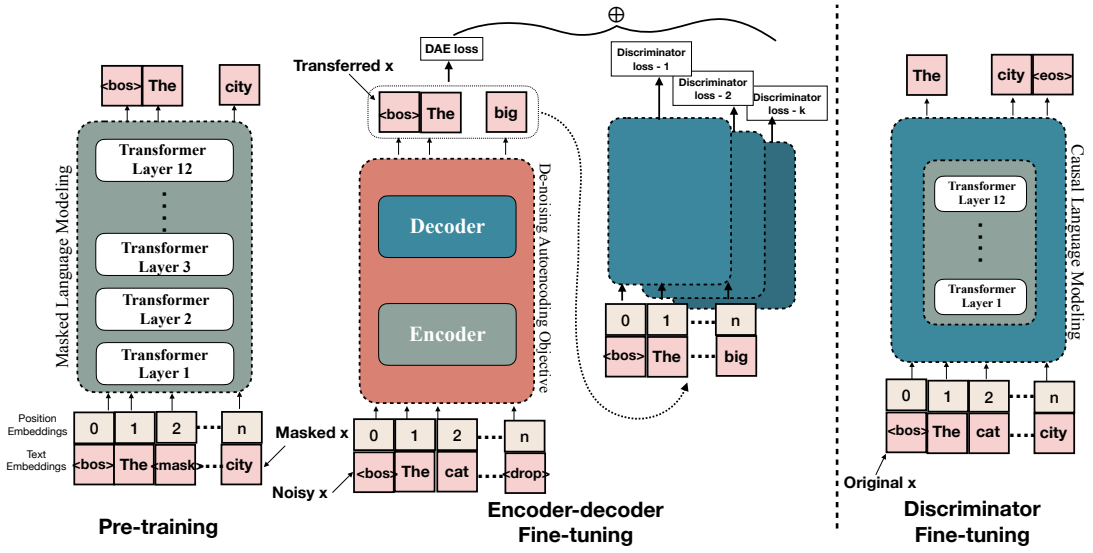

Figure 1: Model Architecture - Left: Generative pre-training using MLM objective, and Fine-tuning encoder-decoder LM with multiple discriminative losses and Right: Discriminator fine-tuning with language modeling (next token prediction) objective. Color for model blocks represents the pre-trained model used for initialization prior to fine-tuning.

get style helps speed up the fine-tuning process as decoder is more adept at generating stylized outputs. This does not add to computational overhead as these fine-tuned models are repurposed as discriminators for stylistic feedback (Section 3.2).

To instill style-awareness to the encoder-decoder setup initialized with pre-trained Transformer models, we fine-tune it with Denoising Autoencoder (DAE) loss using the target-domain corpus. In case of multiple styles, we use a randomized mixture of target-domain corpus from each of the target styles. Under the DAE objective, the encoder takes a noisy masked version $\tilde{x}$ of the text x as input and attempts to fill in the mask token as per the MLM objective that it was pre-trained on. In turn, the decoder re-creates stylistic version of original sentence from this noisy output from the encoder. The overall training objective is

$$\mathcal{L}_{DAE}(\theta_G) = \mathbf{E}_{x \sim T}[-\log P_{\theta_G}(x|\tilde{x})], \quad (1)$$

where θ_G are the trainable parameters of the encoder-decoder model. The noisy version of sentence x from the target corpus T is obtained after dropping tokens from x with probability p_{drop} and masking with a probability of p_{mask}. In conjunction, the encoder and decoder enable style transfer to the target style. The noteworthy aspect here is that the model has no sense of source style and is trained to generate sentences to match the style of the target-domain corpus with which it is trained.

3.2 Fine-tuned LM as discriminators

To extend the single-dimensional style transfer setup above to multi-dimensional setting, we use language models as discriminators to provide the feedback to the model for partially annotated nature of input data. As opposed to a classifier-based discriminator, the language model as discriminator takes into account the wider language distribution of the target style. Additionally, such a setup allows us to use only the target style corpus for training the transfer model, whereas the classifier would require both source and target style corpus to distinguish between a sentence as being from one style or another. Inspired by Yang et al. (2018), we fine-tune a language model on the target style s_i, so that the language model is equipped with language distribution of target domain data. This entails generating the probability of next token, given the previous tokens — also known as Causal Language Modeling objective (Conneau and Lample, 2019). The training loss for the LM for target style s_i with corresponding corpus T_i is

$$\mathbf{E}_{x \sim T_i}\left[\sum_{t=1}^{n}[-\log P_{LM}(x_t|x_1, \ldots, x_{t-1})]\right] \quad (2)$$

We show in our experiments that such a fine-tuning step transforms language distribution of this language model to style s_i and hence serve as soft-discriminator for our framework. We exploit this capability of language models to imbibe style of

fine-tuning corpus by employing language models as style discriminators for transferred sentences. This is based on the idea that if the transferred sentence does not fit well in the target style, then the perplexity of language model fine-tuned on that style will be high (Section 4.1).

For k-dimensional style transfer with target styles $s = \{s_1, s_2, \ldots, s_k\}$, we independently fine-tune k language models on each of the target styles. As discussed in Yang et al. (2018), we are able to forgo the adversarial training for the discriminator, since the fine-tuned discriminative language model is implicitly capable of assigning high perplexity to negative samples (out-of-style samples), as shown in Section 4.1. For the transferred sentence x', the training objective for each target style s_i is,

$$\underset{\theta_G}{\text{argmin}} \mathcal{L}^{s_i} = \mathbf{E}_{x \sim T, x' \sim P_{\theta_G}(x)} \left[\sum_{t=1}^{n} - \log P_{LM_i}(x'_t | x'_1, .., x'_{t-1}) \right] \quad (3)$$

This dictates that transferred sentence x' has low perplexity on the language model fine-tuned on style s_i, for each target style s_i. However, we cannot directly find the argmin_{θ_G} using gradient descent because of discrete sampling of $x' \sim P_{\theta_G}(x)$. To account for this, we use a policy gradient reinforcement learning approach using REINFORCE algorithm (Sutton et al., 1999). The reward for an input sequence x to the style discriminator LM_i is calculated as,

$$r(x) = \sum_{t=1}^{n} \log P_{LM_i}(x_t | x_1, .., x_{t-1}) \quad (4)$$

Using these rewards, the RL objective is to minimize the loss $\mathcal{L}^{s_i}$ given by,

$$\mathcal{L}^{s_i} = \mathbf{E}_{x \sim T, x' \sim P_{\theta_G}(x)} (r(x') - r(x)) [- \log P_{\theta_G}(x' | \tilde{x})] \quad (5)$$

for style s_i, where $P_{\theta_G}(x|\tilde{x})$ is as in Equation 1 and $r(x')$ is the reward in the Equation 4 for the transferred sentence x'. The rewards $r(x)$ represents the baseline reward of greedily sampling the input sequence x by the style discriminator LM_i.

For the style combination $s = \{s_1, s_2, \ldots, s_k\}$, the joint encoder-decoder model is trained on randomized mixture of data from each of the target-domain corpus. The mixture is thus agnostic of individual style of each of the sentence and the discriminative LM for each style guides the generation towards that specific style by rewarding style adherence in the transferred sentence. Randomized mixture of training corpus across styles allows for unified and cohesive understanding of multiple styles by diversifying rewards from different discriminators across samples. The overall training loss for the joint encoder-decoder model is

$$\mathcal{L} = \lambda_{DAE} \mathbf{E}_{x \sim T}[- \log P_\theta(x | \tilde{x})] + \sum_{i=1}^{k} \lambda_i \mathcal{L}^{s_i}, \quad (6)$$

where $\mathcal{L}^{s_i}$ is as defined in Equation 5, and λ_{DAE} and $\{\lambda_i\}_{i=1}^{k}$ are hyper-parameters.

The overall training process is summarized in Figure 1. First, we pre-train a transformer model with Masked language modeling objective as shown in Figure 1(Left). We then initialize discriminator model with this pre-trained language model and fine-tune it with Causal language modeling objective, shown in Figure 1(Right), for each target style. Finally, we initialize the encoder and decoder of the style transfer module with the pre-trained and style-specific fine-tuned language models, respectively. In case of multiple styles, the decoder can be initialized with the language model which is fine-tuned with CLM loss on the mixture of data from target styles, i.e., CLM loss in Equation 2 with $x \sim T$. The joint encoder-decoder model (Figure 1(Centre)) is then trained with a combination of DAE objective and rewards from fine-tuned discriminators of respective target styles.

4 Experiments

We experiment with a combination of sentiment and formality styles. For sentiment, we use a mixture of IMDB (Maas et al., 2011) and Yelp dataset (Li et al., 2018) with $300k$ examples in the positive and negative sentiment each. For formality, we use GYAFC corpus (Rao and Tetreault, 2018) which has $104k$ examples in each formal and informal class. The test set has 3000 and 4849 examples for sentiment and formality respectively, following the data split available in Dai et al. (2019); Rao and Tetreault (2018). For both datasets, the training corpus is non-parallel and the test corpus has human written references available, which we use for content evaluation (Section 4.2).

For pre-training, we use 12-layer Transformer model with 512 hidden units, 16 heads, a dropout rate of 0.1 and learned positional embedding. We train our models with the Adam optimizer, and

Style/Dimension	Sentiment %	Formality %
Positive	71.41	67.09
Negative	76.17	75.59

Table 1: Accuracy of sentences generated by model fine-tuned on style s_i as % of generated sentences labelled as class s_i by the classifier trained on the corresponding style dimension.

Fine-tuning corpus	Test Corpus	
	Same ↓	Opposite ↑
Positive	6.9275	9.6850
Negative	7.7131	9.9637

Table 2: Perplexity of test corpus on models fine-tuned positive and negative corpus (rows). The column *Same* represents that test corpus is same as fine-tuning corpus, leading to lower perplexities and *Opposite* represent test corpus from opposite polarity as fine-tuning corpus leading to higher perplexity.

a learning rate of 10^{-4}. To handle large vocabulary sizes, we use Byte Pair Encoding (BPE) (Sennrich et al., 2016) learned on the Wikipedia dataset. The λs in Equation 6 are determined using hyperparameter tuning on validation set, with style transfer accuracy (Section 4.2) as search criteria.

4.1 Style-awareness of Language Models

To evaluate style variation across language models fine-tuned on different styles, we compare the generations of the fine-tuned models. For single-dimensional style evaluation, we generate sentences from models fine-tuned on negative corpus and positive corpus and compare the style accuracy of generated sentences. The style accuracy is evaluated by employing a FastText (Joulin et al., 2016) classifier trained on the corresponding style dimension. For instance, the classifier for evaluating sentiment accuracy is trained on sentiment corpus tagged with positive and negative class in IMDB and Yelp data. Table 1 shows the accuracy of sentences generated by a model fine-tuned on style s_i as belonging to the class s_i. For both sentiment and formality, the fine-tuned language models are able to generate text faithful to the target style dimension. Thus, we conclude that the language models trained on style s_i are able to capture the essence of the corresponding style reasonably well.

These accuracies are an indication of the style awareness in these fine-tuned LMs. We, therefore, employ the perplexities of these fine-tuned language models to gauge the style of the input text to guide our style transfer model. As discussed in discriminative modeling (Section 3.2), the model fine-tuned with corpus from a certain style is expected to have high perplexity on sentence not from that style and low perplexity otherwise. To this end, we experiment with two models independently fine-tuned on positive and negative corpus. We calculate the perplexity of each of these models on the test corpus from the same style and from the opposite style. As seen in Table 2, the perplexity for each model is substantially lower on the same corpus as compared to that on the opposite corpus. This implies that a language model fine-tuned on positive corpus shows higher perplexity for negative sentences and lower for positive sentences and vice versa. This corroborates the effectiveness of these fine-tuned language models to serve as discriminators for training the style transfer module.

4.2 Evaluation metrics

We measure the performance of our model and the baselines based on the style control, content preservation and fluency. To measure the **accuracy of style transfer**, we train two Fasttext[2] classifiers independently for sentiment and formality using the train corpus, as described in Section 4.1. These classifiers have accuracy of 93.74% and 88.95% respectively on test corpus of respective datasets. We note that formality as a style is more intricately designed, so we also check lexical scoring by Brooke et al. (2010) to evaluate formality, which uses a formality lexicon to assign formality score between -1 (informal) and 1 (formal) to each word and averages it. We scale these scores between 0–100, where higher (100) lexical score signifies formal style and lower (0) score signifies informal style. For informal target style, we report lexical score as $100 - n$, so that a higher average lexical score signifies a better transfer for either polarity.

To measure **content preservation** on transfer, we calculate the BLEU score (Papineni et al., 2002) between the transferred sentence and the input sentence *(self-BLEU)*. Besides this, we also calculate BLEU score between the transferred sentence generated by our model and the corresponding human reference transferred sentence, available for GYAFC and Yelp corpus *(ref-BLEU)*. Since both these corpus account for transfer across only one style dimension each, the provided references are only partial indication of expected outcome. This

[2]https://github.com/facebookresearch/fastText

Model	Style Accuracy			Content Preservation		Fluency
	Classifier ↑		Lexical Scoring ↑	BLEU ↑		Perplexity ↓
	Sentiment	Formality	Formality	*-self*	*-ref*	
Cascaded Style Transformer (Dai et al., 2019)	72.17	64.08	81.29	0.6066	0.3479	8.8657
Adapted Rewriting LM (Syed et al., 2020)	52.59	36.39	72.21	**0.7917**	**0.4259**	6.5963
Cascaded Discriminative LM	69.30	48.18	83.02	0.6634	0.3579	6.6846
Joint Discriminative LM	**79.78**	**65.33**	**85.39**	0.7710	0.4136	**6.4574**

Table 3: Quantitative Comparison of our proposed approach (Joint Discriminative LM) against Cascaded Style Transformer (Dai et al., 2019), Cascaded Discriminative LM method and multi-style transfer using Adapted Rewriting LM (Syed et al., 2020). The upward arrow signifies that higher is better and vice versa. Score of near 100 on formality lexical scoring imply the transferred text is close in formality to the target corpus.

is also apparent from low ref-BLEU scores for our model as well as baselines. Since, the results are presented on aggregated dataset from both these style dimensions, this evaluation is still able to provide reasonable indication of content preservation.

To measure the **fluency** of the text, we calculate perplexity assigned to the generated text sequence by a language model trained on the train corpus, as is standard in style transfer literature (Dai et al., 2019; Subramanian et al., 2018). The perplexity is the measure of log likelihood of the generated sentence on the language model. A lower perplexity is indicative of a more fluent sentence. We use a generative transformer-based language model trained on the dataset combined from two styles.

4.3 Automatic Evaluation

Dai et al. (2019) use transformer-based model (*Style Transformer*) for single-dimensional style transfer. We train two independent Style Transformer models for sentiment and formality transfer and then perform transfer one after another to compare results with our model. We term this as Cascaded Style Transformer setup. The Style Transformer model is shown to have state-of-the-art performance in single-dimensional style transfer; thus it provides an estimate of the performance of sequential single style transfer. We also experiment with Adapted Rewriting LM (Syed et al., 2020) as another baseline. Their work on style rewriting to match author-specific style does not require explicit annotations for the various aspects that constitutes an author's style, but is based on the assumption that the training corpus reflects the target style. In this context, we train their framework on the mixture of data from the respective target styles and report the performance. These are the closest baselines to our proposed approach, since other works dealing with multi-style transfer assume presence of jointly annotated dataset, which is a stronger assumption that we aim to relax. In addition to our proposed model with multiple style transfer, we also train our encoder-decoder architecture with single discriminative LM for one style at a time and perform two stage transfer, similar to one with Cascaded Style Transformer (Dai et al., 2019) setup.

The results in Table 3 show that our model achieves better style control than the Cascaded Style Transformer (Dai et al., 2019) as well as the joint transfer using Syed et al. (2020) for both sentiment and formality. As seen in Table 3, cascaded style transfer models perform poorly on content preservation. This is because transferring style one after other leads to huge loss in content, thus both the two-stage models score lower on content preservation metrics, both w.r.t. the input text and the reference transferred text. This demonstrates the advantage of using single model to control for multiple styles. The effect can also be observed in Table 4 which demonstrates qualitative results for Cascaded Style Transformer model and our model. We can see in many cases content loses the underlying meaning of source sentence during the two-stage transfer, whereas our model is able to retain original meaning of the sentence well, corroborating the findings of automatic evaluation. Among the cascaded models, the Discriminative LM scores marginally better on content preservation than the Style Transformer model. We attribute this to initialization with the same pre-trained LM resulting in shared content space in the underlying single style transfer models. However, due to independent training of the two single style transfer models, they are not able to model interplay between these styles and hence perform worse on style control than our proposed model trained jointly on multiple styles.

Our model also scores better on fluency, as seen in Table 3. This is also apparent from the exam-

Target style	Source sentence	Transferred Sentence	
		Style Transformer	Our model (multi-style)
Positive+Formal	That's not funny. I don't think she'll like it.	So funny movie. I really like it.	That was very funny. I am sure she will **appreciate it**.
	Give your brother some money and tell him to take a hike.	Just give your brother some time and it will be good again.	Give your brother some money and **request him to leave**.
Negative+Formal	An intelligent, rewarding film that I look forward to watching again.	ludicrous, shallow film that look forward to watching again.	An unintelligent, poor film that **I would not look forward** to watching again.
	super friendly staff, quick service and amazing and simple food was done right!	says wait staff, quick not amazing before overcooked food done were okay.	dirty staff and slow service and simple food was not done right.
Positive+Informal	You need to separate the bad thing and move on.	need to the great thing and move on.	You need to enjoy the good stuff and move on.
	The evening started out slow.	The evening spent in professional show.	The evening **began** amazing.
Negative+Informal	Great food recommendations steak and tuna were both great.	terrible food 9am steak and were both terrible.	Disappointing food recommendations steak and tuna were horrible.
	That person in hilarious.	You person in worse!	**That guy** in so boring.

Table 4: Qualitative results for transfer to different target style combination across different models. (Different colors highlight the transferred segments corresponding to underlined input sentence; Text in bold highlights adherence to target *formality* in text generated by our model.)

Model	Style Accuracy		Content Preservation	Fluency	Transfer Quality
	Sentiment	Formality			
Cascaded Style Transformer (Dai et al., 2019)	3.5909	2.7424	3.2803	2.7424	2.9318
Joint Discriminative LM (Our Model)	**3.8561**	**3.0379**	**4.1061**	**4.1894**	**4.1091**

Table 5: Results for Human Evaluation across different metrics. Each value represents the average of rating between 1 (Very bad) and 5 (Very good).

ples in Table 4, where sentences generated by Cascaded Style Transformer are much less coherent. Qualitative experiments also highlight the ability of our model to incorporate intricacies of *formality* stylistic dimension (shown in bold) better than the Cascaded Style Transformer model. Among single step transfer models (Syed et al. (2020) and our proposed approach), we note that content preservation is marginally better for Syed et al. (2020)'s model, however, our model is able to yield much better style transfer owing to feedback on style control by multiple discriminators.

4.4 Human evaluation

To augment automatic evaluation results, we conduct a human study to evaluate the model outputs across various dimensions such as content preservation, style control, fluency, and overall transfer quality. Based on comparable style control in Cascaded Style Transformer and our proposed approach on automatic metrics, we compare the transfer quality across these two models by a small-scale human study. We select 40 sentences, with 10 examples from each combinations of sentiment and formality as target style, and collect annotations from 4–5 participants for each example. Out of resulting annotations, more than 85% annotations favoured our results over baseline. The average participant rating across different dimensions is shown in Table 5. We test the statistical significance of these results using z-test statistic. With $\alpha = 0.05$, the preferences indicated in human study are significant across all metrics. These results are in line with our automatic evaluations and add confidence to the efficacy of our proposed approach in achieving style transfer across multiple dimensions.

5 Conclusion and Future Work

We propose an approach to extend currently existing style transfer work to multiple style setting without imposing any extra constraints on availability of dataset. Our method makes use of disjoint corpus from separate styles to enable one step transfer across multiple target styles. We exploit multiple discriminative language models with an encoder-decoder framework, all emerging from large transformer-based language models pretrained on Masked Language Modeling objective and fine-tuned separately for transfer and discriminative purposes. We show that unified single step transfer approach is able to achieve better transfer while offering much better content preservation which is paramount to any style transfer task.

Further improvements are in scope for adding modularity to the proposed transfer module. In the current setup, each version of model is trained for a specific combination of target style(s). The utility of such a model increases manifold with added ease of transfer across multiple style combinations within a single model. This could be attempted by employing a controlled language model as a unified discriminator for multiple styles, which would be the subject of further research.

Ethics Statement. We recognise the ethical implication of employing large language models trained on data infused with unchecked biases. As with any generative task, style transfer too suffers from the potential misuse for fact distortion, plagiarism and more. The paper aims at establishing academic utility of proposed framework. To meet ethical standards, this solution has to coupled with strict misrepresentation, offensiveness and bias checks.

References

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Coling 2010: Posters*, pages 90–98, Beijing, China. Coling 2010 Organizing Committee.

Stephane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of BERT for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 108–117, Hong Kong. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

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.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2017. Style Transfer in Text: Exploration and Evaluation. *arXiv e-prints*, page arXiv:1711.06861.

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 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness Transfer: A Tag and Generate Approach. *arXiv e-prints*, page arXiv:2004.14257.

François Mairesse and Marilyn A. Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Abhilasha Sancheti, Kundan Krishna, Balaji Vasan Srinivasan, and Anandhavelu Natarajan. 2020. Reinforced rewards framework for text style transfer. In *Advances in Information Retrieval*, pages 545–560, Cham. Springer International Publishing.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6830–6841. Curran Associates, Inc.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. *CoRR*, abs/1811.00552.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8968–8975.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 1057–1063, Cambridge, MA, USA. MIT Press.

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9008–9015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 6000–6010.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. 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 S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7287–7298. Curran Associates, Inc.