Adapting Language Models for Non-Parallel Author-Stylized Rewriting

by

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu N, Vasudeva Varma

in

34th AAAI Conference on Artificial Intelligence
(*AAAI-2020*)

New York, USA

Report No: IIIT/TR/2020/-1

Centre for Search and Information Extraction Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
February 2020

Adapting Language Models for Non-Parallel Author-Stylized Rewriting

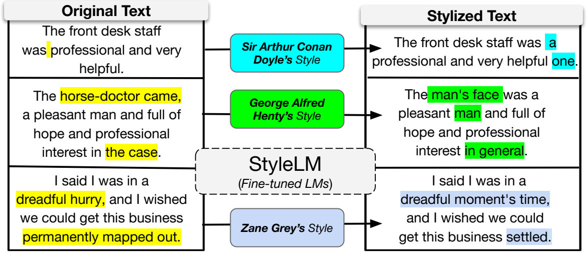
Bakhtiyar Syed[‡], **Gaurav Verma**[†], **Balaji Vasan Srinivasan**[†]
Anandhavelu N[†], **Vasudeva Varma**[‡]
[‡]IIIT Hyderabad, [†]Adobe Research
syed.b@research.iiit.ac.in
{gaverma, balsrini, anandvn}@adobe.com
vv@iiit.ac.in

Abstract

Given the recent progress in language modeling using Transformer-based neural models and an active interest in generating stylized text, we present an approach to leverage the generalization capabilities of a language model to rewrite an input text in a target author's style. Our proposed approach adapts a pre-trained language model to generate author-stylized text by fine-tuning on the author-specific corpus using a denoising autoencoder (DAE) loss in a cascaded encoder-decoder framework. Optimizing over DAE loss allows our model to learn the nuances of an author's style *without* relying on parallel data, which has been a severe limitation of the previous related works in this space. To evaluate the efficacy of our approach, we propose a linguistically-motivated framework to quantify stylistic alignment of the generated text to the target author at lexical, syntactic and surface levels. The evaluation framework is both interpretable as it leads to several insights about the model, and self-contained as it does not rely on external classifiers, e.g. sentiment or formality classifiers. Qualitative and quantitative assessment indicates that the proposed approach rewrites the input text with better alignment to the target style while preserving the original content better than state-of-the-art baselines.

Introduction

There has been a growing interest in studying style in natural language and solving tasks related to it (Hu et al. 2017; Shen et al. 2017; Subramanian et al. 2018; Fu et al. 2018; Vadapalli et al. 2018; Niu and Bansal 2018). Tasks like genre classification (Kessler, Numberg, and Schütze 1997), author profiling (Garera and Yarowsky 2009), sentiment analysis (Wilson, Wiebe, and Hoffmann 2005), social relationship classification (Peterson, Hohensee, and Xia 2011) have been of active interest to the community. Recently, stylized text generation (Hovy 1990; Inkpen and Hirst 2006) and style transfer (Li et al. 2018; Prabhumoye et al. 2018; Fu et al. 2018) have gained traction; both these tasks aim to generate realizations of an input text that align to a target style. A majority of the work here is focused around generating text with different levels of sentiment (Shen et al. 2017; Ficler and Goldberg 2017) and formality (Jain et al. 2019) and also a combination of these attributes (Subramanian et al. 2018). The interest along these lines has given rise to annotated and parallel data that comprise of paired realizations that lie on opposite ends of formality and sentiment spectrum (Rao and Tetreault 2018; Mathews, Xie, and He 2016). The dimensions of style considered across all these works are psycholinguistic aspects of text and the aim is to transfer the text across different levels of the chosen aspect.



The front desk staff was professional and very helpful.

The horse-doctor came, a pleasant man and full of hope and professional interest in the case.

I said I was in a dreadful hurry, and I wished we could get this business permanently mapped out.

The front desk staff was a professional and very helpful one.

The man's face was a pleasant man and full of hope and professional interest in general.

I said I was in a dreadful moment's time, and I wished we could get this business settled.

Figure 1: An overview of generating author-stylized text using *StyleLM*, our proposed model.

However, there has been lack of explorations that aim to generate text across author styles – wherein the notion of style is not a specific psycholinguistic aspect but an amalgam of the author's linguistic choices expressed in their writing (Jhamtani et al. 2017; Tikhonov and Yamshchikov 2018). While the work by Jhamtani et al. (2017) tries to generate "Shakespearized" text from Modern English and is in a similar vein, it relies on the availability of parallel data. Since the availability of parallel data is not always guaranteed and it is arduous to curate one, such an approach cannot scale for different authors. We therefore propose a novel framework for author-stylized rewriting without relying on parallel data with source-text to target-text mappings. Figure 1 shows a few examples where an input text is rewritten in the style of a chosen author by our model.

Our approach for generating author-stylized text involves leveraging the generalization capabilities of state-of-the-art language models and adapting them to incorporate the stylistic characteristics of a target author without the need of parallel data. We first pre-train a language model on a combination of author corpus (Lahiri 2014) and Wikipedia data using the masked language modeling objective (Devlin et al. 2019). Drawing inspiration from the unsupervised machine translation setup of Lample and Conneau (2019), we cascade two copies of this pre-trained language model into an encoder-decoder framework, where the parameters of the encoder and decoder are initialized with the pre-trained language model. This cascaded framework is fine-tuned on a specific target author's corpus of text by reconstructing the original text from its noisy version and optimizing on a denoising autoencoder loss. The fine-tuned model thus adapts itself towards the style of the target author as we show via our experimental analysis.

Author-stylized rewriting of text takes a text, which may or may not have a distinctive style, and rewrites it in a style that can be attributed to a target author. Since the writing style of authors is determined by several linguistically active elements that are expressed at lexical, syntactic, and semantic levels, it is challenging to evaluate the stylistic alignment of rewritten text to target author's style. To this end, we propose a novel and interpretable framework that is linguistically motivated, to quantify the extent of stylistic alignment at multiple levels. As we elaborate upon in the later sections, our evaluation suggests that the proposed approach performs better than three relevant and competitive baselines – showing significant adaption to the writing style of target authors, both qualitatively and quantitatively. Notably, our approach performs on par (and better in certain dimensions) with state-of-the-art method for stylistic rewriting *using parallel data*, *without* leveraging the parallel nature of underlying data.

The **key contributions of this work** are threefold.

1. We propose and evaluate an approach to generate author-stylized text without relying on parallel data by adapting state-of-the-art language models.
2. We propose an evaluation framework to assess the efficacy of stylized text generation that accounts for alignment of lexical and syntactic aspects of style. Contrary to existing evaluation techniques, our evaluation framework is linguistically-aware and easily interpretable.
3. Our proposed approach shows significant improvement in author-stylized text generation over baselines, both in quantitative and qualitative evaluations.

Related Work

Stylized Text Generation: In recent times, several explorations that aim to generate stylized text define a psycholinguistic aspect, like, formality or sentiment (Shen et al. 2017; Ficler and Goldberg 2017; Jain et al. 2019) and transfer text along this dimension. The approaches themselves can range from completely supervised, which is contingent on the availability of parallel data (Ficler and Goldberg 2017), to unsupervised (Shen et al. 2017; Li et al. 2018; Jain et al. 2019). Some of the influential unsupervised approaches include *(a)* using readily available classification-based discriminators to guide the process of generation (Fu et al. 2018), *(b)* using simple linguistic rules to achieve alignment with the target style (Li et al. 2018), or *(c)* using auxiliary modules (called *scorers*) that score the generation process on aspects like fluency, formality and semantic relatedness while deciding on the learning scheme of the encoder-decoder network (Jain et al. 2019). However, in the context of our setting, it is not possible to build a classification-based discriminator or scorers to generate author-stylized text. Moreover, linguistic-rule based generations are intractable given the large number of rules required to define a target author's style. To this end, we aim to adapt state-of-the-art language models to generate author-stylized text from non-parallel data. The choice of using language models is motivated by the fact that stylistic *re*writing *builds on* the task of simple text generation (i.e., writing).

There are some works that adapt an input text to the writing style of a specific author (Jhamtani et al. 2017; Tikhonov and Yamshchikov 2018). While Tikhonov and Yamshchikov (2018) generate author-stylized poetry by learning the style end-to-end using conditioning and concatenated embeddings of corresponding stylistic variables, theirs is not a *rewriting* task. Jhamtani et al. (2017) aim to generate Shakespearized version of modern English language using parallel data. Our proposed approach aims to overcome this shortcoming by only relying on non-parallel data and only requires the corpus of the target author text for stylistic rewriting. As we show later, the proposed framework is comparable (even better in some of the dimensions) to Jhamtani et al.'s approach across content preservation and style transmission metrics without utilizing the parallel corpus.

Language Models: Generative pre-training of sentence encoders (Radford et al. 2018; Devlin et al. 2019; Howard and Ruder 2018) has led to strong improvements on several natural language tasks. Their approach is based on learning a Transformer (Vaswani et al. 2017) language model on a large unsupervised text corpus and then fine-tuning on classification and inference-based natural language understanding (NLU) tasks. Building up on this, Lample and Conneau (2019) extend this approach to learn cross-lingual language models. Taking inspiration from this, we extend the generative pre-training for our task of author-stylized rewriting.

The recently proposed language model GPT-2 (Radford et al. 2019) is pre-trained on a large and diverse dataset (WebText) and is shown to perform well across several domains and datasets including natural language generation. The unsupervised pre-training is setup to model the generation probability of the next word, given the previous words, i.e., $P(y_t \mid y_{1:t-1}, \mathbf{x})$ – more generally referred to as the causal language modeling (CLM) objective. Specific to the task of text generation, it takes an input prompt ($\mathbf{x}$) and aims to generate text that adheres to the input context. As substantiated in the later sections, GPT-2, when fine-tuned on author-specific corpus, shows significant stylistic alignment with the writing style of target author. However, given the inherent differences involved in the setup of stylistic *rewriting* and stylized text generation, it performs poorly on content preservation. While in stylistic rewriting, the objective is to retain the information in the input text in the stylized generation, stylistic generation by GPT-2 generates the content that is *related* to the input *prompt* and hence fine-tuned GPT-2 cannot address the task of stylistic rewriting. In the cross-lingual language modeling literature, a recent exploration by Lample and Conneau (2019) learns cross-lingual language models by first pre-training on 3 different language modelling objectives: *(i)* causal language model (CLM), *(ii)* masked language model (MLM) – similar to BERT (Devlin et al. 2019), and *(iii)* translation language model (TLM) - which is a supervised setup leveraging parallel corpora. Following the pre-training, Lample and Conneau cascade the encoder and decoder to address the tasks of *supervised* cross-lingual classification and machine translation by fine-tuning on a combination of denoising auto-encoder (DAE) and back-translation losses. Taking inspiration from this work, we pre-train a language model on a large corpus using MLM objective and then fine-tune it on author-specific corpus using DAE loss in an encoder-decoder setup. Using DAE loss ensures that we don't rely on availability of parallel corpora, while the pre-trained language model facilitates the task of rewriting by building a firm substratum.

Evaluating Stylized Generation: Fu et al. (2018) propose an evaluation framework to assess the efficacy of style transfer models on two axes: *(i) content preservation* and *(ii) transfer strength*. While the former caters to the content overlap between input and generated text (quantified using BLEU (Papineni et al. 2002)), the latter takes into account the alignment of generated text with target style. In their setup, as it is with many others, the notion of target style is a psycholinguistic aspect (formality or sentiment) for which classifiers or scorers are readily available and are hence used to quantify the transfer strength (Jain et al. 2019; Li et al. 2018; Mir et al. 2019). However, for evaluating *author*-stylized text generations the evaluation frameworks are not well established. Jhamtani et al. (2017) and Tikhonov and Yamshchikov (2018) overcome this by using the content preservation metrics as a proxy of transfer strength, leveraging the availability of the ground-truth stylized text. The unavailability of a suitable metric for transfer strength is particularly pronounced in evaluating unsupervised approaches as there is no target data to compare the generations against. To this end, we propose a linguistically-aware and interpretable evaluation framework which quantifies alignment of multiple lexical and syntactic aspects of style in the generated text with respect to the target author's style.

Proposed Approach: StyleLM

There are two key aspects to our approach – pre-training a Transformer-based language model on a large dataset that acts as a substratum and fine-tuning on author-specific corpus using DAE loss to enable stylized rewriting. The entire approach is **not** contingent on the availability of parallel data and the models are learned in a self-supervised manner.

Figure 2 illustrates the proposed framework for stylistic rewriting. We first pre-train the Transformer-based language model on a large unsupervised corpus using the masked language modeling (MLM) objective (Devlin et al. 2019). The choice of using a Transformer-based architecture is based on their recent success in language modeling (Vaswani et al. 2017; Devlin et al. 2019; Radford et al. 2018; 2019). The MLM objective encourages the LM to predict the masked word(s) from the input sequence of words leveraging bidirectional context information of the input.

Given a source sentence $\mathbf{x}$, $\mathbf{x}^{\setminus u}$ is a modified version of $\mathbf{x}$ where its token from position u is masked by replacing it with a mask token $[\mathbf{MASK}]$ - thus keeping the length of the masked sentence unchanged. The MLM objective pre-trains the language model by predicting the original token x^u, taking the masked sequence $x^{\setminus u}$ as input, while learning the parameters θ for the conditional probability of the language model. We minimize the log-likelihood given by,

$$L(\theta; \mathcal{X}) = \frac{1}{|\mathcal{X}|} \Sigma_{\mathbf{x} \in \mathcal{X}} \log P(\mathbf{x}^u \mid \mathbf{x}^{\setminus u}; \theta) \quad ^1 \qquad (1)$$

where, $\mathcal{X}$ denotes the entire training corpus. For pre-training the language model using the MLM objective, following Devlin et al. (2019), we randomly mask 15% of the tokens in each input sequence, 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. A difference between our model and the MLM proposed by Devlin et al. (2019) is the use of text streams of sentences (truncated at 256 tokens) in contrast to pairs of sentences. This has been shown to give considerable gains for text generation tasks (Lample and Conneau 2019). Also, unlike Devlin et al. (2019), we do not use the *Next Sentence Prediction* (NSP) objective.

The language model (LM) above learns to predict the masked words over a large corpus, but does not incorporate any style-related fine-tuning that facilitates rewriting the input text in a given target author's style. To achieve this, we cascade two instances of the pre-trained LM in an encoder-decoder setup where one instance acts as the encoder and the other acts as a decoder. In other words, the learnable parameters of both encoder and decoder are initialized using the pre-trained LM. Note that the architecture of Transformer-based language models allows two exact instances of the pre-trained LM to be cascaded, without explicitly aligning the encoder's output and the decoder's input (Bahdanau, Cho, and Bengio 2014) since the attention-mechanism is inherent in the design of Transformers (Vaswani et al. 2017). Lample and Conneau (2019) successfully used such a cascading to bootstrap the iterative process of the model initialization for the unsupervised machine translation task. Taking inspiration from this, we fine-tune the encoder-decoder on the DAE loss, given by,

$$\mathcal{L}^{DAE} \quad = \quad \mathbf{E}_{\mathbf{x} \sim \mathcal{S}}[-\log P(\mathbf{x} \mid C(\mathbf{x}))] \qquad (2)$$

where, $C(x)$ is the noisy version of the input sentence $\mathbf{x}$ and $\mathcal{S}$ are the sentences in target author's corpus. To obtain a noisy version $C(\mathbf{x})$ of input text $\mathbf{x}$, we drop every word in $\mathbf{x}$ with a probability p_{drop} and also blank the input words with a probability p_{blank}[2].

[1] The equation given here describes MLM for one token (Song et al. 2019). In practice, multiple tokens are masked in the original BERT architecture and for our experiments, which is just an extension of the above idea for training speed-up.

[2] replace the word with $[BLANK]$.

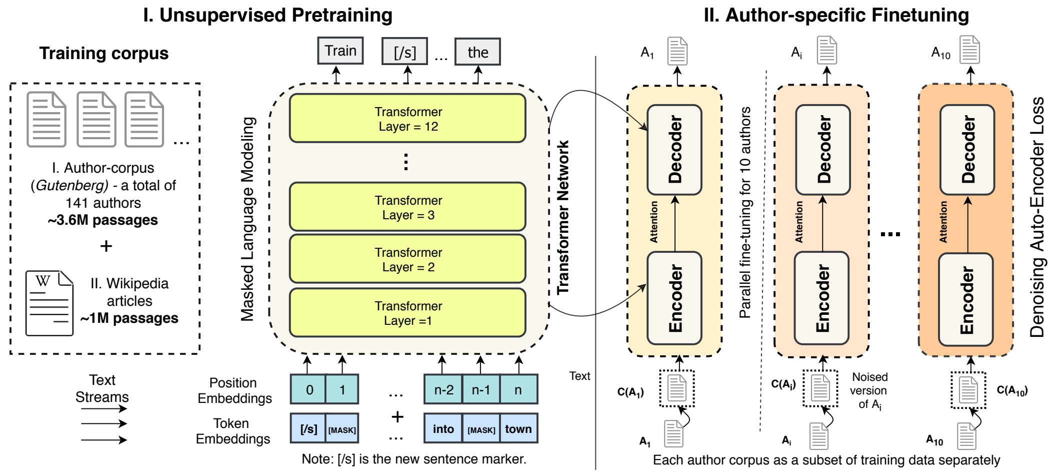

I. Unsupervised Pretraining
II. Author-specific Finetuning
Training corpus
I. Author-corpus (Gutenberg) - a total of 141 authors ~3.6M passages
+
II. Wikipedia articles ~1M passages
Text Streams
Masked Language Modeling
Train
[/s]
...
the
Transformer Layer = 12
Transformer Layer = 3
Transformer Layer = 2
Transformer Layer =1
Transformer Network
Position Embeddings
Token Embeddings
0
1
n-2
n-1
n
[/s]
[MASK]
into
[MASK]
town
+
Note: [/s] is the new sentence marker.
A1
Ai
A10
Decoder
Attention
Encoder
Parallel fine-tuning for 10 authors
Denoising Auto-Encoder Loss
Text
C(A1)
C(Ai)
Noised version of Ai
C(A10)
Each author corpus as a subset of training data separately

Figure 2: **Proposed *StyleLM* model**. We first pre-train a language model on large English corpus (*I. Unsupervised Pretraining*) and then cascade the pre-trained LMs into an encoder-decoder like framework (as represented by the curved arrows). The encoder-decoder is fine-tuned separately on each of the target author's corpus using DAE loss (*II. Author-specific fine-tuning*).

When the pre-trained language model is cascaded as the encoder and decoder, and further fine-tuned with a noisy version of the text, the encoder generates the masked words (since that is the original objective of the pre-trained LM). However, since the input to the decoder, which is same as the output of the encoder, has no masked words, it tries to reconstruct the clean version of the noisy input text. In other words, fine-tuning the encoder-decoder on target author's corpus using the DAE loss (equation 2) pushes the model's decoder towards inducing target author's style while rewriting the input text from the encoder.

Implementation Details During pre-training with MLM, we use the Transformer encoder (Vaswani et al. 2017) (12-layer) with GELU activations (Hendrycks and Gimpel 2017), 512 hidden units, 16 heads, a dropout rate of 0.1 and learned positional embeddings. We train our models with the Adam optimizer (Kingma and Ba 2014), and a learning rate of 10^{-4} . We use streams of 256 tokens and a mini-batches of size 32. We train our model on the MLM objective until the language model's perplexity shows no improvement over the validation dataset. For fine-tuning on a target author, which involves reconstruction of the *whole* input passage[3] from its noisy version we use the same pre-trained MLM Transformer initialization for both the encoder and decoder, similar to Lample and Conneau (2019), with the same hyperparameters used for pre-training. p_{drop} and p_{blank} are set to 0.1 and the model is fine-tuned until convergence.

To handle the vocabulary size for such a huge dataset, we use Byte Pair Encoding (BPE) (Sennrich, Haddow, and Birch 2015) on the combined training dataset and learn $80k$ BPE codes on the dataset. Since we use BPE codes on the combination of the training dataset of the 141 authors, we can scale these for any author at will – thus the ability to adapt to any author in the Gutenberg corpus or beyond.

Evaluation Framework

Dataset We collated a subset of the Gutenberg corpus (Lahiri 2014) consisting of 142 authors and $2,857$ books written by them. For evaluating on a completely unseen author (a zero-shot setting), we set aside the writings by Mark Twain from the training corpus. The remaining authors are used as training corpus during pre-training resulting in $\sim$ 3.6M passages. To diversify the pre-training dataset, we use 1 million passages from Wikipedia (Radford et al. 2018) along with $\sim$ 3.6M passages from the Gutenberg corpus – leading to a total of $\sim$ 4.6M passages for pre-training the LM. Of these, we set aside 5000 passages for validation and 5000 for test during the pre-training stage.

To fine-tune the encoder-decoder framework from the pre-trained LM, we pick a subset of 10 authors from the Gutenberg corpus and independently treat them as target authors to generate author-stylized text. The 10 chosen authors are: Sir Arthur Conan Doyle, Charles Dickens, George Alfred Henty, Nathaniel Hawthorne, Robert Louis Stevenson, Rudyard Kipling, Thomas Hardy, William Makepeace Thackeray, and Zane Grey. We fine-tune independently for each of the 10 target authors and evaluate the efficacy of our proposed approach using a novel evaluation framework with roots in linguistic literature, described in a later section.

For inference during test-time, we use the following three corpora to obtain our source sentences : (a) texts from books written by Mark Twain, (b) Opinosis Review dataset (Ganesan, Zhai, and Han 2010), (c) a Wikipedia article on *Artificial Intelligence* (https://en.wikipedia.org/wiki/Artificial_intelligence) which does not appear in the original mix of the Wikipedia training corpus. Texts from these sources span a diverse range of topics and writing styles – while Mark Twain's writings are literary, Opinosis reviews are everyday, the Wikipedia article on AI presents an interesting scenario where many of the words in the source text are not present in target author's corpus, given the different timelines.

[3] Unlike the MLM which predicts only a part of the input.

We evaluate our performance against 4 baselines - 3 of which are trained on non-parallel data, while the 4th one uses parallel data.

1. Vanilla GPT-2 based generation: Radford et al. (2019) show that language models present considerable promise as unsupervised multi-task learners. We use their vanilla GPT-2 pre-trained Transformer decoder (Radford et al. 2019) as our first baseline.[4]. The GPT-2 is fed a prompt directly during inference and the generated outputs are compared against other generations.

2. Author fine-tuned GPT-2: The second baseline is the fine-tuned GPT-2 model for the cross-entropy loss on each of the target author's corpus separately. We use the stylized text generated by providing a prompt to the fine-tuned model for comparisons.

3. Denoising-LM : no author-specific fine-tuning: This baseline is similar to our StyleLM network, but fine-tuned on the entire corpora using the DAE loss (as opposed to just the author-specific corpus). The purpose of this baseline is to evaluate the content preservation capabilities of our setup.

4. Supervised Stylized Rewriting: Jhamtani et al. (2017) propose an LSTM-based encoder-decoder architecture for generating a "Shakespearized" text originally written in modern English, by leveraging parallel data. We compare this baseline only for generating Shakespearized text (using their data). We train the other three baselines and StyleLM by treating Shakespeare's corpus as the target author's corpus (without using the parallel nature of the data).

Proposed Evaluation Methodology

Following existing literature on style transfer and stylized text generation, we evaluate our proposed frameworks along two axes: *content preservation* and *stylistic alignment*.

Content preservation aims to measure the degree to which the generated stylized outputs have the same meaning as the corresponding input sentences. Following existing literature, we use the BLEU metric[5] (Papineni et al. 2002) and the ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-3 and ROUGE-L (Lin 2004)).

The core contribution of our evaluation framework is in the linguistic-motivation used to quantify the **stylistic alignment** of a generated piece of text with the target style we wish to achieve. While there have been several studies around formality and sentiment transfer on text, the same evaluation criteria does not apply to our setting because of two reasons: *(a)* the classifier-based evaluation, which is facilitated by readily available classifiers for aspects like sentiment and formality, cannot be used to evaluate stylistic alignment with respect to an author's style, and *(b)* author style is an amalgam of several linguistic aspects which are much more granular than the psycholinguistic concepts. To this end, taking motivation from Verma and Srinivasan (2019), we formulate a multi-level evaluation scheme that identifies and quantifies stylistic expression at *surface, lexical and syntactic* level. Once we quantify the stylistic expression, we use standard distance metrics to measure the stylistic alignment with target.

Linguists have identified style, especially in English language, to be expressed at three levels – surface, lexical and syntactic (Strunk 2007; DiMarco and Hirst 1988; Crystal and Davy 2016). We first discuss the expression of stylistic elements as well their quantification. After quantifying the stylistic expressions at these levels, we discuss their incorporation into out evaluation framework.

Lexical elements of style are expressed at the *word-level*. For instance, an authors choice words may be more subjective than objective (*home* vs. *residence*), or more formal than informal (*palatable* vs. *tasty*). For instance, we found that Rudyard Kipling, known for his classics of children's literature, had a higher tendency to use more concrete words (like, *gongs*, *rockets*, *torch*, etc.) unlike Abraham Lincoln, who being a political writer, used more abstract words (like freedom, patriotism, etc.). Inspired from Brooke and Hirst (2013), we consider four different spectrums to take lexical-style into account: *(i)* subjective-objective, *(ii)* concrete-abstract, *(iii)* literary-colloquial, and *(iv)* formal-informal.

For quantifying these lexical elements, we use a list of seed words for each of the eight categories above, viz. subjective, objective, concrete, abstract, literary, colloquial, formal and informal (Brooke and Hirst 2013). Following Brooke and Hirst (2013), we compute normalized pointwise mutual information index (PMI) to obtain a raw style score for each dimension, by leveraging co-occurrences of words in the large corpus. The raw scores are normalized to obtain style vectors for every word, followed by a transformation of style vectors into k-Nearest Neighbor (kNN) graphs, where label propagation is applied. Since the eight original dimensions lie on the two extremes of four different spectrums, i.e., subjective-objective, concrete-abstract, literary-colloquial, and formal-informal, we compute 4 averages across the entire author-specific corpus. The averages, in the range $[0, 1]$, denote the tendency of author using subjective, concrete, literary, or formal words, in contrast to using objective, abstract, colloquial, or informal words, as evidenced in their historical works[6].

Syntactic elements relate to the syntax of the sentence – while some authors construct complex sentences, others construct simple sentences. For instance, as per the writings of Abraham Lincoln available in the Gutenberg corpus, a majority of his sentences can be categorized as compound-complex, while those of Rudyard Kipling's are mostly simple sentences (which are better suited to children). Taking inspiration from Feng, Banerjee, and Choi (2012), we categorize syntactic style into 5 different categories – (a) simple (b) compound (c) complex (d) complex-compound sentences, (e) others. For quantifying these stylistic elements, we compute the fraction of sentences that are categorized into the 5 categories by the algorithm proposed by Feng, Banerjee, and Choi (2012). Since any given sentence will definitely lie

[4] In our experimental setup, we utilise the pre-trained $124M$ parameter model for generation - https://github.com/openai/gpt-2

[5] BLEU score is measured with *multi-bleu-detok.perl*

[6] The final output is a 4 dimensional vector with each of the elements, let's say $l_{sub} \in [0, 1]$.. The value of l_{sub} will denote the tendency of the author to choose subjective words instead of their objective counterparts, which can be given by $1 - l_{sub}$

Data Source	**Model**	**Content Preservation** (↑)					**Stylistic Alignment** (↓)		
		BLEU	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-L	Lexical (MSE)	Syntactic (JSD)	Surface (MSE)
Opinosis	GPT-2	18.3±2.3	0.51±0.06	0.29±0.09	0.20±0.06	0.36±0.08	0.48±0.06	0.27±0.09	0.45±0.01
	GPT-2 (FT)	24.3±1.6	0.58±0.07	0.36±0.08	0.27±0.11	0.42±0.09	0.32±0.08	0.23±0.02	0.40±0.03
	LM + DAE	41.1±1.3	**0.77**±0.11	0.49±0.05	0.39±0.07	0.61±0.08	0.33±0.05	0.23±0.01	0.38±0.02
	StyleLM	**43.4**±1.7	0.73±0.13	**0.53**±0.06	**0.41**±0.08	**0.68**±0.07	**0.29**±0.04	**0.19**±0.01	**0.31**±0.04
Mark Twain	GPT-2	16.7±2.4	0.43±0.03	0.26±0.07	0.16±0.04	0.29±0.09	0.41±0.08	0.29±0.03	0.42±0.05
	GPT-2 (FT)	22.9±1.6	0.49±0.06	0.38±0.08	0.21±0.06	0.37±0.08	0.35±0.07	0.25±0.02	0.39±0.06
	LM + DAE	31.7±1.5	**0.68**±0.14	0.44±0.07	0.27±0.07	0.45±0.10	0.37±0.03	0.24±0.01	0.37±0.03
	StyleLM	**34.4**±1.8	0.61±0.16	**0.48**±0.06	**0.31**±0.06	**0.53**±0.08	**0.32**±0.03	**0.21**±0.02	**0.33**±0.03
AI Wiki	GPT-2	12.6±2.1	0.37±0.04	0.19±0.09	0.09±0.05	0.25±0.08	0.49±0.07	0.31±0.02	0.46±0.05
	GPT-2 (FT)	15.4±1.5	0.43±0.09	0.23±0.06	0.13±0.04	0.29±0.07	0.40±0.03	0.28±0.03	0.42±0.05
	LM + DAE	23.7±1.6	**0.59**±0.12	0.31±0.08	0.18±0.06	0.37±0.09	0.41±0.04	0.26±0.02	0.41±0.03
	StyleLM	**26.7**±1.9	0.54±0.13	**0.34**±0.08	**0.23**±0.08	**0.46**±0.09	**0.34**±0.02	**0.22**±0.01	**0.36**±0.04

Table 1: **Evaluating content preservation and stylistic alignment.** We evaluate the performance of *StyleLM* against three baselines and on three test sets across multiple content preservation and stylistic alignment metrics. The reported numbers are mean and standard deviations ($\mu \pm \sigma$) across all the 10 target authors. FT denotes author-specific fine-tuning; ↑ / ↓ indicates that higher / lower is better, respectively.

Model	**Content Preservation** (↑)					**Stylistic Alignment** (↓)		
	BLEU	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-L	Lexical (MSE)	Syntactic (JSD)	Surface (MSE)
GPT-2	18.1	0.43	0.18	0.11	0.22	0.41	0.30	0.43
GPT-2 (FT)	21.3	0.48	0.24	0.16	0.28	0.36	0.26	0.39
LM + DAE	30.2	0.55	0.30	0.19	0.38	0.32	0.24	0.36
Jhamtani et al. (2017)	31.3	**0.57**	**0.33**	0.23	0.43	0.29	**0.17**	**0.33**
StyleLM	**33.8**	0.53	0.31	**0.25**	**0.44**	**0.28**	0.21	0.34

Table 2: **Comparison against supervised baseline.** Similar to Table 1, we evaluate the performance of all the models against the approach of (Jhamtani et al. 2017) which relies on parallel data. For author-specific fine-tuning of *StyleLM* and GPT-2 (FT), we use Shakespeare's corpus but without exploiting its parallel nature with modern English corpus.

in only one of the 5 categories, the 5 dimensional vector averaged across the sentences in a corpus can be thought of as probability distribution over the 5 categories.

Surface elements relate to statistical observations concerning aspects like the average number of *(i)* commas, *(ii)* semicolons, *(iii)* colons per sentence, *(iv)* sentences in a paragraph, and *(v)* number of words in a sentence. We quantify the surface-level elements into a 5 dimensional vector.

Although the above enumerations of stylistic elements within a level, whether lexical, syntactic or surface, are not exhaustive, they are indicative of the stylistic expression at different levels. Computing the above statistics on an author-specific corpus gives an interpretable notion of the concerned author's writing style. Such a notion of style spans across multiple linguistic levels and has a considerable granularity. To this end, to quantify the stylistic alignment between generated text and the target text, we first compute these statistics for both the generated corpus and the target author's corpus. Then, we use standard distance metrics to obtain the extent of stylistic alignment at different linguistic levels. For lexical and surface-level alignment, we use mean squared error (MSE). Since syntactic style vector is a probability distribution over different syntactic categories, we use Jensen-Shannon divergence (otherwise known as symmetric KL divergence) to measure the alignment.

Results and Analysis

Qualitative Evaluation Table 3 presents samples of author-stylized text generated using StyleLM for some of the authors. Key highlights include the switch between 'kind', 'obliged' and 'ready to accept' for the source word 'catered'. The modification of the word 'super' – which is used in a colloquial sense, to 'extra' without sacrificing the semantic meaning, demonstrates author-specific adaptation across different time frames. Similar observation can be made by noting the adaptation of 'AI is programmed' to 'brain is to learn' and 'rewarding' to 'gratification' on fine-tuning for Charles Dickens' writing style. Qualitative assessment of the generated samples depict the efficacy of our approach by illustrating alignment with the target author's style as well as significant content preservation.

Quantitative Evaluation Our evaluation framework assesses the capability of our proposed *StyleLM* model across both content preservation and stylistic alignment metrics.

The results for stylized rewriting of the test corpus to the various author's style (10 in total) are presented in in Table 1. All the fine-tuned *StyleLM* models are tested on a test set that spans different domains – *(a)* Opinosis (Ganesan, Zhai, and Han 2010) which contains sentences extracted from user reviews on a variety of topics from Tripadvisor (hotels), Edmunds.com (cars) and Amazon.com (various electronics), *(b)* text from Mark Twain's books, and *(c)* a Wikipedia page on *Artificial Intelligence*[7]. To reiterate, the objective is to *rewrite* the above test corpora into a style that reflects the style of target author we fine-tuned for. The averaged values for all 10 authors, as well as the standard deviation, across both content preservation as well as stylistic alignment metrics, are given in Table 1.

It can be inferred from Table 1 that in terms of stylistic alignment, GPT-2 (FT), i.e., author fine-tuned GPT-2, performs comparable to LM + DAE, i.e., denoising LM with

[7] We did not include any of these in the pre-training nor in the fine-tuning stage. As such, our model has never seen this data.

Source	Original Text	NH's Style	CD's Style	GAH's Style
Opinosis	The staff was so polite and **catered** to our every need .	The staff was so polite and kind to our every need .	The staff was so polite and obliged to our every need .	The staff was so polite and ready to accept our every need .
	Front desk staff were **not super easy** to work with but ...	Western desk , the staff were not abilities easy to work with , but...	front desk and staff were not extra easy to work with , but...	The won desk staff were not force easy to work with , but...
Mark Twain	I asked him **if he** learned to talk out of a book, and if I could **borrow it anywhere?**	I asked him whether he had learned to talk of a dream , and if I could borrow it.	I asked him if he had learned to talk out of a book ; and if I could borrow it .	I asked him if he learned to talk out of a man's mind and if I could borrow it
	Meanwhile, if we understand each other now, I will go to work again.	And if we understand each other's, I go to work.	And if we understand each other , I will go to work.	Then if we understand each other's words I will go to work.
AI Wiki	If the **AI is programmed** for "reinforcement learning", goals can be implicitly induced by **rewarding** some types of behavior or punishing others.	If the human mind is bosoms for Heaven's sake , he can be implicitly induced by rewarded some types of behavior or punishment.	If the brain is learn for men's object can be implicitly induced by gratification some kind of behaviour or punishment's punish's	If the round is turn for one's point he can be implicitly induced by done some type of conduct or punishing.

Table 3: **Samples of stylized text generated by *StyleLM*.** The target authors are Nathaniel Hawthorne (NH), Charles Dickens (CD) and George Alfred Henty (GAH). The source text has been taken from Opinosis, Mark Twain and AI Wiki, as indicated.

no author-specific fine-tuning, across all the three datasets and on each of the three stylistic levels. However, the content preservation for LM + DAE is better than that of GPT-2 (FT). The vanilla GPT-2, however, shows the least impressive in terms of both content preservation as well stylistic alignment. Specifically, the poor performance on content preservation can be attributed to the fact that GPT-2 and GPT-2 (FT) are both trained for generating continuations of input prompts and not for the task of stylistic *rewriting*. It is nonetheless encouraging to see that fine-tuning the GPT-2 language model on author-specific corpus, i.e., GPT-2 (FT), increases the extent of stylistic alignment with target author's style, establishing GPT-2 (FT) as a competitive baseline to compare stylistic alignment against.

While LM + DAE, i.e., denoising LM without author-specific fine-tuning, shows good performance in terms of content preservation and stylistic alignment, our proposed approach, *StyleLM*, shows considerable gains across all the metrics, against the LM + DAE. This observation confirms our hypothesis that the author-specific fine-tuning using DAE loss teaches the model to better learn the stylistic characteristics of the target author. Consistency of results across the diverse test sets shows a broader coverage in terms of applicability of the presented results.

Interestingly, we notice that ROUGE-1 scores for the baseline LM + DAE (without author fine-tuning) are slightly higher than those for *StyleLM*. A closer inspection of the generated samples from the two models reveals that this is because the stylized generations of the former are not as structurally coherent as those of the latter; i.e., while the predicted words are more accurate, they are not predicted in the correct order. This is further substantiated by the higher values for ROUGE-2, ROUGE-3 and ROUGE-L scores.

Comparison with Supervised Approach While *StyleLM* performs better than the other unsupervised stylized generation models as shown in Table 1, it is critical to determine its performance w.r.t. the supervised approach proposed by Jhamtani et al. (2017). We compare their LSTM-based encoder-decoder approach with GPT-2, GPT-2 (FT), LM + DAE and *StyleLM* after fine-tuning them on Shakespeare's corpus. As it can be inferred from the results presented in Table 2, *StyleLM* performs better than the supervised approach in terms of BLEU, ROUGE-3, ROUGE-L, and lexical stylistic alignment. The performance, as quantified by rest of the metrics, is comparable to that of (Jhamtani et al. 2017). Given that *StyleLM* was trained without leveraging the parallel nature of the data, the results are promising and demonstrate the abilities of our proposed model in generating author-stylized text while preserving the original content.

Conclusion & Future Work

In this work, we address the task of author-stylized rewriting by proposing a novel approach that leverages the generalization capabilities of language models. Building on the top of language models, we fine-tune on target author's corpus using denoising autoencoder loss to allow for stylistic adaptation in the process of reconstruction, without relying on parallel data. We also propose a new interpretable framework to evaluate stylistic alignment at multiple linguistic levels. We show that our proposed approach is able to capture the stylistic characteristics of target authors while rewriting the input text and performs not only better than other relevant and competitive baselines, but is also competent to an entirely supervised approach that relies on parallel data.

The linguistic understanding of style, on which the proposed evaluation framework is based, can be used to guide the process of generating stylized text. The process of generation can be tuned to comply with attributes of style at different levels by penalizing or rewarding the (mis)alignment with these elemental attributes of style. Our plan is to explore this in further details, as part of future work.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*.

Brooke, J., and Hirst, G. 2013. A multi-dimensional bayesian approach to lexical style. In *NAACL-HLT*.

Crystal, D., and Davy, D. 2016. *Investigating english style*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

DiMarco, C., and Hirst, G. 1988. Stylistic grammars in language translation. In *Proceedings of CoLing*. ACL.

Feng, S.; Banerjee, R.; and Choi, Y. 2012. Characterizing stylistic elements in syntactic structure. In *EMNLP-CoNLL*.

Ficler, J., and Goldberg, Y. 2017. Controlling linguistic style aspects in neural language generation. *arXiv:1707.02633*.

Fu, Z.; Tan, X.; Peng, N.; Zhao, D.; and Yan, R. 2018. Style transfer in text: Exploration and evaluation. In *AAAI*.

Ganesan, K.; Zhai, C.; and Han, J. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of CoLing*. ACL.

Garera, N., and Yarowsky, D. 2009. Modeling latent biographic attributes in conversational genres. In *ACL-IJCNLP*.

Hendrycks, D., and Gimpel, K. 2017. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *ArXiv* abs/1606.08415.

Hovy, E. H. 1990. Pragmatics and natural language generation. *Artificial Intelligence*.

Howard, J., and Ruder, S. 2018. Universal language model fine-tuning for text classification. In *Proceedings of ACL*.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward controlled generation of text. In *ICML*.

Inkpen, D., and Hirst, G. 2006. Building and using a lexical knowledge base of near-synonym differences. *Computational linguistics*.

Jain, P.; Mishra, A.; Azad, A.; and Sankaranarayanan, K. 2019. Unsupervised controllable text formalization. In *AAAI*.

Jhamtani, H.; Gangal, V.; Hovy, E.; and Nyberg, E. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *EMNLP*.

Kessler, B.; Numberg, G.; and Schütze, H. 1997. Automatic detection of text genre. In *Proceedings of EACL*. ACL.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Lahiri, S. 2014. Complexity of Word Collocation Networks: A Preliminary Structural Analysis. In *SRW EACL*.

Lample, G., and Conneau, A. 2019. Cross-lingual language model pretraining. *arXiv:1901.07291*.

Li, J.; Jia, R.; He, H.; and Liang, P. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv:1804.06437*.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Mathews, A. P.; Xie, L.; and He, X. 2016. Senticap: Generating image descriptions with sentiments. In *AAAI*.

Mir, R.; Felbo, B.; Obradovich, N.; and Rahwan, I. 2019. Evaluating style transfer for text. In *NAACL-HLT*.

Niu, T., and Bansal, M. 2018. Polite dialogue generation without parallel data. *Transactions of ACL* 6.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Peterson, K.; Hohensee, M.; and Xia, F. 2011. Email formality in the workplace: A case study on the enron corpus. In *Workshop on Languages in Social Media*.

Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; and Black, A. W. 2018. Style transfer through back-translation. *arXiv:1804.09000*.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners.

Rao, S., and Tetreault, J. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *NAACL-HLT*.

Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural machine translation of rare words with subword units. *ArXiv* abs/1508.07909.

Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. 2017. Style transfer from non-parallel text by cross-alignment. In *NeurIPS*.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv:1905.02450*.

Strunk, W. 2007. *The elements of style*.

Subramanian, S.; Lample, G.; Smith, E.; Denoyer, L.; Ranzato, M.; and Boureau, Y. 2018. Multiple-attribute text style transfer.

Tikhonov, A., and Yamshchikov, I. P. 2018. Guess who? multilingual approach for the automated generation of author-stylized poetry. In *IEEE SLT*.

Vadapalli, R.; Syed, B.; Prabhu, N.; Srinivasan, B.; and Varma, V. 2018. Sci-blogger: A step towards automated science journalism. In *CIKM*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Verma, G., and Srinivasan, B. V. 2019. A lexical, syntactic, and semantic perspective for understanding style in text. *ArXiv* abs/1909.08349.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of EMNLP*. ACL.