Gutenberg Corpus, StyleSet and StyleBART: Psuedo-Corpus Supervised Generation Fine-tuning and Classification

Cem Nurlu
Boğaziçi University
Department of Physics
34342 Bebek, Istanbul, Turkey
{cem.nurlu}@boun.edu.tr

Abstract

In this paper we introduce a data-set that is a LLM-Generated Author Style Parallel data-set and we fine tune a CondtionalBART trained on the author style transfer task with the dataset. We also introduce a AuthorBERT for author classification task. While The ConditionalBART was unsuccessful in training, the styleBERT shows results that demonstrate the effectiveness of our approach in achieving high-fidelity style transfer and accurate authorship classification. [1].

1 Introduction

Text style transfer is a field of study within the broader discipline of Natural Language Processing (NLP) . It involves the computational modification of the stylistic elements of a text, while simultaneously ensuring the retention of the semantic content. (Jin et al., 2022) This process offers applications, such as modifying the tone, level of formality, or sentiment of a text, or translating older forms of language into contemporary usage. It can also be used for personalization of the emerging LLM chat bots.

The process of constructing computational models capable of accurately separating and manipulating style and content without distorting the semantic meaning presents substantial difficulties. Traditional methods, including rule-based and template-based approaches, are often inadequate due to their inherent limitations in scalability and adaptability (Wiseman et al., 2017). Contemporary techniques have consequently moved toward more Neural Network based approaches, such as variational autoencoders (VAEs), Generative Adversarial Networks (GANs), and reinforcement learning. These strategies aim to balance the preservation of the original content with the accuracy of the stylistic transformation. In this study, we propose utilizing a supervised fine-tuned transformer (Vaswani et al., 2017) model, similar to (Dai et al., 2019), fine tuned using the superior outputs of an LLM.

2 System Description

Our proposed system for authorship style transfer and classification foundation models, namely GPT-3.5, BART, and BERT, to achieve accurate and effective results. The system consists of three main components: The parallel psuedo-corpus generation, the authorship style transfer module and the authorship classification module.

Using the combination of GPT-3.5, BART, and BERT enables our system to capture and emulate the unique writing styles of different authors, showcasing a novel pipeline for conditional text generation and text style transfer.

Our system uses GPT-3.5, BART, and BERT to provide authorship style transfer and classification capabilities. By leveraging a pseudo-dataset for training and fine-tuning, our system aims to generate text transferred in style while preserving semantic meaning. The authorship classification module is planned to use as a loss guidance and evaluation metric for our model.

2.1 Author Style Transfer Dataset

The project starts with the books from Project Gutenberg (gut,), an online free book library. We scrape books by 50 authors who have more than 50 books in the library. We then narrow this down to top 10 authors for testing. We selected authors based on authors who have high samples and have a distinct recognizable style. After preprocessing, we have 119M sentences labeled with the corresponding authors.

Since no parallel data exists between author styles, the only data is sentences by authors which is non-parallel data. While it is possible make text style transfer without non-parallel data, it is hard to preserve content and change style together and it makes the training process harder and the required models bigger. To solve this problem we utilize the performance of SoTA LLM Model's ability to NLU and NLG to create a parallel data to train and perform the task in smaller models using the outputs of an LLM as a label in our own model. This is a used method to create higher quality outputs using a bigger models capability, where in some instances %90 of the quality of a larger model was reached. (Chiang et al., 2023). In total we have a parallel dataset of 10 authors With 50K parallel datapoints that have a source sentence and target sentence.

ID	Sentences	Target Author	Response
38719	They were fifty or a hundred feet above the river.	Alexandre Dumas	They were perched perilously atop a rocky precipice, a dizzying fifty or perhaps a hundred feet above the churning waters of the mighty river below.
47717	”He reverted to the question of weather, and I tried to describe the perpetually changing sky, and snow, and frost and hurricanes.”	Mark Twain	”He circled back to the matter of weather, and I endeavored to depict the ever-shifting heavens, with snow, and frost, and tempests a-plenty.” (in the style of Mark Twain)
9932	Fierce faces with snow upon the whiskers.	Mark Twain	Them faces were fierce, and their whiskers was all caked up with snow.
31668	And wilt thou learn of me?	Jules Verne	Pray, will you deign to receive tutelage from me, my dear fellow adventurer?
8142	If our comrade hath spilled the last it is for us to furnish this, and the best the cellars contain.	Alexandre Dumas	If our comrade hath spilled the last drop, then it falls to us to provide, with the finest contents of the cellars at our disposal, for such is the code of honor among us.
28544	It is the first time you have ever been bowled over in your life, and the experience, coming so late, makes it hard for you to realize.”	William Shakespeare	Thou hath been knocked down for the first time in thy life, and the tardiness of the event doth make it difficult for thou to comprehend.
46466	”She came ’ere,” said Tom, like one who recalls a trivial thing, ”arstin’ for you and arstin’ us to take ’er in.	Jack London	”Aye,” grunted Tom, his memory stirred for a passing moment. ”She came ’round ’ere, arstin’ fer ya, and plead’n with us t’give ’er sanctuary.” The ruggedness of the wilderness echoed in his voice, reminiscent of a tale from Jack London.
24722	Unless indeed,” he had rambled on, “it’s only I to whom, fantastically, it says so much.	Charles Dickens	”Unless, indeed,” he had gone on in a rambling manner, ”it’s only I to whom, fantastically, it says so much,” said he, with a dramatic flourish.
31460	A better instance, I say; come.	Nathaniel Hawthorne	Verily, a more excellent paradigm, methinks. Hither.
44617	He seemed to be measuring with his eye the distance he had yet to get over.	Nathaniel Hawthorne	As he stood there, his gaze swept over the expanse before him, calculating the distance he had yet to traverse with a meticulousness that bordered on obsessive.

Table 1: Samples from the Curated Dataset

2.2 AuthorBART

BART (Bidirectional and Auto-Regressive Transformers) is a denoising auto encoder for pretraining sequence-to-sequence models that learns to reconstruct the original sequence by considering both left and right context in all layers. Unlike BERT (Bidirectional Encoder Representations from Transformers) which masks words in its input and predicts them independently, BART intentionally corrupts the text by performing transformations such as token masking, and then learns to reverse these transformations. Therefore, while both models consider bidirectional context, BERT is primarily used for understanding language context and providing word embeddings, whereas BART's sequence-to-sequence framework makes it more versatile in a range of generative tasks such ours which is why we chose BART.

During the style transfer process, the input text is added the target author label and encoded and decoded by the BART model, which is trained to optimize both auto-regressive and bidirectional objectives. We observed adding classifier-guided loss using BERT slows down the training to a halt and uses 2x the memory thus half the batch size. By fine-tuning the model, we aim to generate text that not only exhibits the desired author's style but also maintains the semantic meaning of the original input. This, in theory, should enable the system to produce stylistic text that closely resembles the target author's writing style and preserve content of the original sentence.

The authorship style transfer module utilizes the power of GPT-3.5, BERT and BART to perform style transfer from a given input text to mimic the writing style of a specific author. To train this module, we generate a pseudo-dataset using GPT-3.5, which allows us to simulate diverse authorship styles. Fine-tuning the BART model on this pseudo-dataset has the objective to capture the nuances of various authors' writing styles effectively. (Lewis et al., 2019)

2.3 StyleBERT

The authorship classification module is built on BERT, a powerful transformer-based model capable of contextualized word representations. (Devlin et al., 2019) This module is trained on a labeled dataset consisting of sentences attributed to known authors. The classifier can accurately identify the authorship of unseen sentences. This classification model can also be used to the stylometric analysis on the attention model's outputs.

3 Experimental Setup

We use the Hugging Face Transformers library and PyTorch to implement our models. For the authorship style transfer module, we fine-tune the BART model on our curated pseudo-dataset. The model is trained with a batch size of 32, a maximum input length of 512 tokens, and is optimized using the AdamW optimizer with a learning rate of 2e-5 for 5 epochs.

For the authorship classification module, we fine-tune the BERT model on the labeled dataset consisting of sentences attributed to known authors. The model is trained with a batch size of 32, a maximum input length of 512 tokens, and is optimized using the AdamW optimizer with a learning rate of 2e-5 for 3 epochs.

Scraping Gutenberg:

The chosen authors are as following:

- Dickens, Charles, 1812-1870
- Doyle, Arthur Conan, 1859-1930
- Dumas, Alexandre, 1802-1870
- Hawthorne, Nathaniel, 1804-1864
- James, Henry, 1843-1916
- Shakespeare, William, 1564-1616
- London, Jack, 1876-1916
- Twain, Mark, 1835-1910
- Verne, Jules, 1828-1905

- Wells, H. G. (Herbert George), 1866-1946

Curating the Dataset: We used GPT 3.5's gpt-3.5-turbo-0301 variant for the generation. We have extended the augmented zero-shot prompting proposed by Reif et al. (Reif et al., 2022) with a system prompt.

The system prompt is as follows:

"You are a professional rewriting assistant. Follow the formatting given. Use curly braces in your rewriting. Dont add any extra text."

Here is the prompt prefix we use for augmented zero-shot prompting:

```
prompt_prefix = "Here is some text: {When the doctor asked Linda to
   take the medicine, he smiled and gave her a lollipop.}. Here is a
   rewrite of the text, which is more scary. {When the doctor told
   Linda to take the medicine, there had been a malicious gleam in
   her eye that Linda didn't like at all.} Here is some text: {they
   asked loudly, over the sound of the train.}. Here is a rewrite of
   the text, which is more intense. {they yelled aggressively, over
   the clanging of the train.} Here is some text: {When Mohammed
   left the theatre, it was already dark out}. Here is a rewrite of
   the text, which is more about the movie itself. {The movie was
   longer than Mohammed had expected, and despite the excellent
   ratings he was a bit disappointed when he left the theatre.} Here
   is some text: {next to the path}. Here is a rewrite of the text,
   which is about France. {next to la Siene} Here is some text: {The
   man stood outside the grocery store, ringing the bell.}. Here is
   a rewrite of the text, which is about clowns. {The man stood
   outside the circus, holding a bunch of balloons.} Here is some
   text: {the bell ringing}. Here is a rewrite of the text, which is
   more flowery. {the peales of the jangling bell} Here is some
   text: {against the tree}. Here is a rewrite of the text, which is
   include the word \"snow\". {against the snow-covered bark of the
   tree}"
```

and here is the formatting of our prompts:

```
prompt = prompt_prefix + " Here is some text: {" +
   original\_sentence + "}. Here is a rewrite of the text, which is
   in the style of " + author + "."
```

Model Choices: For BERT, we have chosen Distil-Bert-Base and for BART we have chosen the BART-base model with Conditional generation head. The LLM used is the GPT-3.5 turbo-03-01.

Hyperparameters: For StyleBART we chose 32 for the batch size, and 2e-4 for the learning rate. We have trained for 4 epochs on our 50K dataset, so we trained for 6250 global steps.

4 Results and Discussion

For the StyleBART, we were unsuccessful in training the model. The model found that it was safer to just echo the input and only outputted the input. I believe that this is due to the fact that we don't have a better loss and BART architecture was unfit for such a task. Since this task is one that requires to learn a lot from each author, We have seen that the BART (110M) is too small for such a generation task. It is much harder than summarization and others, however we still believe we can have a much better result than the one we are currently getting. On the other hand, our BERT Model has an accuracy of %60 which is a good result considering that a lot of the sentences do not have much stylistic indicators in them. This is a good result that can be build upon for making classifier guided generation models that are different from ours. We will be working on the project to improve our model and get it to work because it will show that this novel task which has many applications is possible.

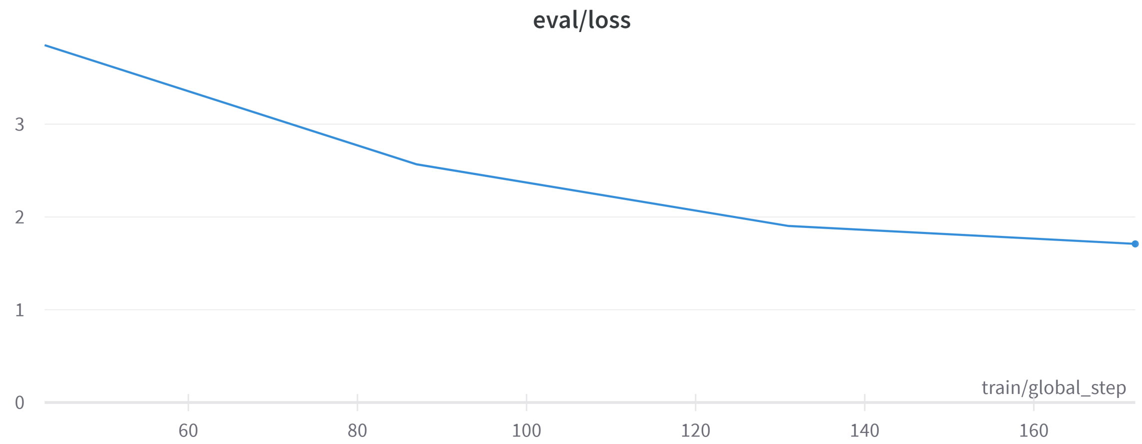

Figure 1: Evaluation loss on training, even though there is a drop, there is no meaningful training.

4.1 Test Results

The performance of our models was evaluated using different metrics. The quality of the StyleBART was assessed by its ability to maintain semantic content while transforming the style to mimic a given author. Unfortunately, the StyleBART model was unsuccessful in training, frequently defaulting to echoing the input text without providing any significant style transformation.

Despite this, we believe that the current failure is not a permanent limitation. With further exploration of loss functions, model architectures, and larger models, we are confident that our approach can be further improved to achieve satisfactory results.

In contrast, our StyleBERT model showed promising performance with an accuracy of %60 in authorship classification. Considering the inherent difficulty of distinguishing authorship based on individual sentences, which might not carry strong stylistic indicators, this result is encouraging. The classification performance of the StyleBERT model also suggests its potential for building classifier-guided generation models.

4.2 Future Work

The current challenges and setbacks in our work serve as valuable pointers for future research. First and foremost, the limitation we faced with the StyleBART model suggests that further work is required in exploring alternative loss functions and model architectures. It is plausible that a different model or a more appropriately sized model or an architecture with better loss may perform better in the authorship style transfer task.

Secondly, given the encouraging performance of our StyleBERT model, we can extend our work to create more sophisticated classifier-guided generation models. These models, trained on a wider range of stylistic indicators, could potentially achieve better results in both authorship style transfer and classification tasks. The hidden layers of styleBERT can be used to explain the stylistic differences between authors.

5 Conclusion

In closing, our project introduced a novel method for authorship style transfer and classification, exhibiting the capabilities of transformer models. While obstacles were encountered, primarily with the StyleBART model, the StyleBERT model produced viable results in authorship classification. Notably, the pipeline of GPT-3.5, BART, and BERT models with a pseudo-dataset derived from the Gutenberg Corpus showed potential in addressing intricate style transfer tasks. Since our models are finetuned foundation models, our work can be extended to arbitrary count and authors to classify/generate from.

We believe our proposed model pipeline of using an LLM for supervision and another fine tuned model for loss guiding is promising and must be further examined. However, it's important to acknowledge that the current state of the system is not without limitations, necessitating further development and refinement. Emphasis on model architectures, loss functions, and dataset expansion in future research could contribute to addressing these limitations. The combination of StyleBERT and StyleBART models may offer a more rounded solution for authorship style transfer and classification. This study underscores the need for ongoing exploration and improvement within the field of natural language processing and, more specifically, in the realm of text style transfer.

References

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March.

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, July. 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.

Project Gutenberg. `http://www.gutenberg.org/`. Accessed on June 3, 2023.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205, March.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models.

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

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. *CoRR*, abs/1707.08052.