PKU-IAI Technical Report: PKU-IAI-

CFACL : Style Transfer from Causal to Formal Text

Changye Li
Department of Yuanpei
Peking University
2200017853@stu.pku.edu.cn

Abstract

In this work, we constructed a pretraining dataset and a preference dataset from ACL papers to train the transformation from informal text to formal ACL-style text. We trained this transformation using the T5 [7] and GPT-2 models [6]. In addition, we created an evaluation set to assess the quality of style transfer through AI feedback. The trained GPT-2 models [6], a similar size model, Qwen2.5-1.5B-Instruct [9] and a general large model, Mistral-7B-v0.2 [3], were evaluated for their performance.

1 Introduction

Large language models (LLMs) have shown prominent capabilities [1, 10]. , particularly in Natural Language Processing (NLP). The current state-of-the-art GPT-o1 [5] serves as a foundation model, showing general proficiency in various NLP tasks. However, contemporary LLMs require substantial computational resources and memory for both training and inference. When focusing on specialized tasks, this resource consumption may be excessive. Text Style Transfer (TST) represents one such specialized task that warrants attention.

Academic writing poses significant challenges for students, frequently resulting in drafts that lack polish or stylistic consistency. The automation of writing refinement through Text Style Transfer (TST) presents a viable approach to enhance clarity, coherence, and adherence to academic conventions. Our objective is to achieve comparable performance on this specific task using a smaller model trained on a moderate dataset.

Furthermore, evaluating Text Style Transfer results remains a relatively subjective process, lacking well-established universal benchmarks. Mukherjee and Dušek [4] provides a comprehensive review of TST models, datasets, and evaluation metrics, proposing fundamental guidelines for AI-based TST assessment. We have implemented practical AI evaluation algorithms based on these principles.

Specifically, our research addresses the following question:

Can a compact model, through straightforward training and fine-tuning procedures, achieve superior performance in the Style Transfer task from Casual to Formal Text?

The primary contributions of our work are as follows:

1. **Novel Casual text and Formal text pairs dataset**: We have developed a dataset comprising casual text and formal text pairs (30K, 60K) based on The ACL OCL Corpus [8]. To our knowledge, this represents the first TST dataset with Casual-Formal pairs derived from academic paper data, designed for model pretraining and fine-tuning.

2. **Style Transfer Models**: We have trained style transfer models based on the GPT2-Large model [6], capable of generating formal text and performing Casual to Formal Text Style Transfer.
3. **Comprehensive Evaluation Framework**: We have established a complete evaluation pipeline for Style Transfer Models, incorporating automated metrics such as BLEU and perplexity, alongside a quantitative assessment system built upon AI feedback mechanisms.

2 Method

In this section, we present our methods for the creation of datasets, training models, and model evaluation procedures.

2.1 Dataset Construction

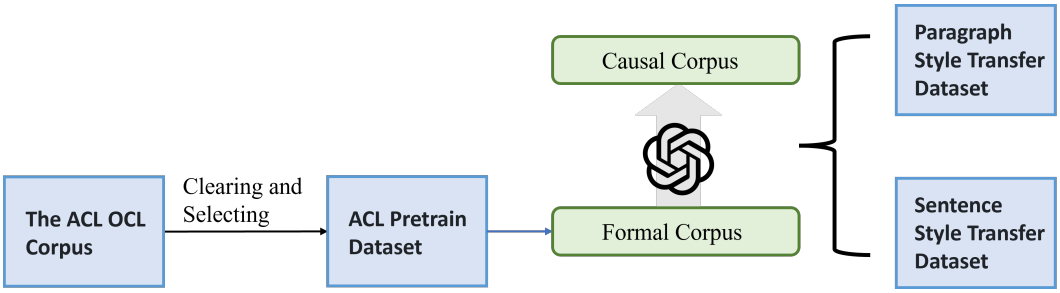

Figure 1: Pipeline for constructing style transfer datasets from the ACL OCL corpus [8]. The process involves clearing and selecting the ACL pretrain dataset, which is further refined into a formal corpus using Deepseek model. This formal corpus is then augmented into a causal corpus. The resulting datasets include paragraph-level and sentence-level style transfer datasets.

We extracted and cleaned the main text content from 20K papers in The ACL OCL Corpus [8] to create the ACL Pretrain Dataset, excluding formulas and citations to retain pure textual content. From this corpus, we extracted the formal corpus and generated the corresponding casual corpus entries through AI annotation using the Deepseek-V2 model [2]. After organizing these annotated text pairs, we constructed two fine-tuning datasets at different granularities: a paragraph-level style transfer dataset containing 30K pairs and a sentence-level style transfer dataset comprising 60K pairs. Our dataset construction pipeline is illustrated in Figure 1.

2.2 Model Training

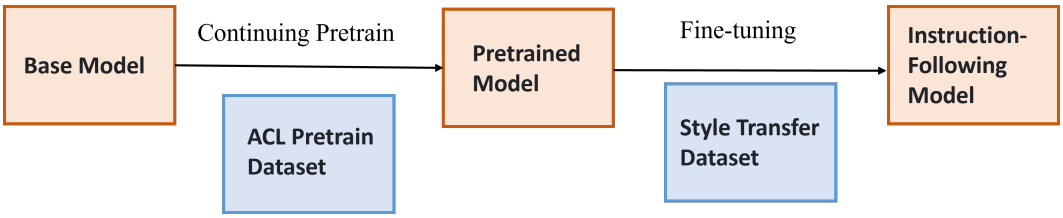

Figure 2: Framework for training an instruction-following model. The process starts with a base model, which is pretrained further using the ACL pretrain dataset to create a pretrained model. This model is then fine-tuned using a style transfer dataset, resulting in the final instruction-following model.

We implemented a hybrid training approach that combines continued pre-training and fine-tuning strategies. Initially, we conducted autoregressive training of the base model on the ACL Pretrain Dataset to adapt the model's inherent text style. Subsequently, we fine-tuned the model using our style transfer dataset to enable instruction-following capabilities for text style transformation. We experimented with multiple model architectures, including GPT2-Large [6] and T5-Large [7] for this training procedure. Our model training pipeline is illustrated in Figure 2.

2.3 Model Evaluation

We developed a comprehensive evaluation methodology that integrates automated metrics with AI feedback mechanisms. For automated assessment, we used BLEU scores and perplexity metrics to evaluate the model's generative capabilities. The AI feedback component encompasses three key dimensions: Style Transfer Strength, Content Preservation, and Fluency. Through carefully designed prompts, we leveraged state-of-the-art LLMs to generate standardized evaluation results. For comparative analysis, we utilized Mistral-7b-v0.2 [3] and Qwen2.5-1.5B-Instruct [9] as baseline models.

3 Experiment

In this section, we present our experiment on the dataset and model, comparing our training result with larger model and state of the art model in similar size.

3.1 Experiment Design

Following the pipeline described in Section 2.2, we trained multiple model variants based on the GPT2-Large [6] and T5-Large [7] architectures. Due to model size limitations and potential context window constraints, we restricted our fine-tuning process to sentence-level data. The resulting models include GPT2-pretrained, GPT2-finetuned, T5-pretrained, and T5-finetuned. We evaluated these models through analysis of training curves, automated evaluation metrics, and AI feedback evaluation metrics.

3.2 Experiment Results

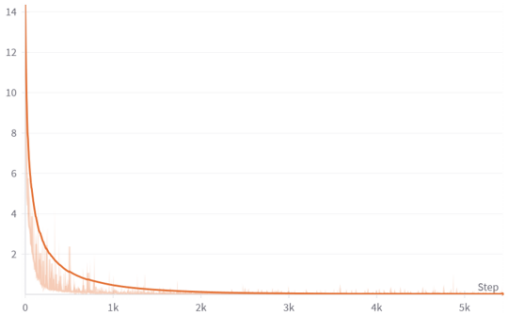

Figure 3: Continuing Pretrain Loss Curve of T5-Large Model

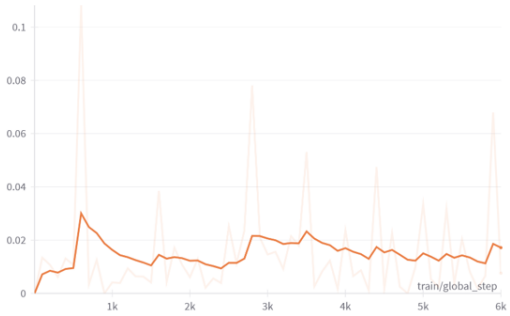

Figure 4: Finetuning Loss Curve of T5-Large Model

Figure 3 presents our pre-training results for the T5 model. We observed training instability during autoregressive training, which can be attributed to the model's original pre-training methodology of Text Infilling, making it incompatible with autoregressive training approaches. Further fine-tuning experiments, shown in Figure 4, demonstrated that the T5 model [7] struggled to adapt to new tasks with limited training data. In conclusion, the T5 model failed to achieve the desired performance on the TST task. Consequently, our subsequent evaluations focused exclusively on the GPT2-Large architecture.

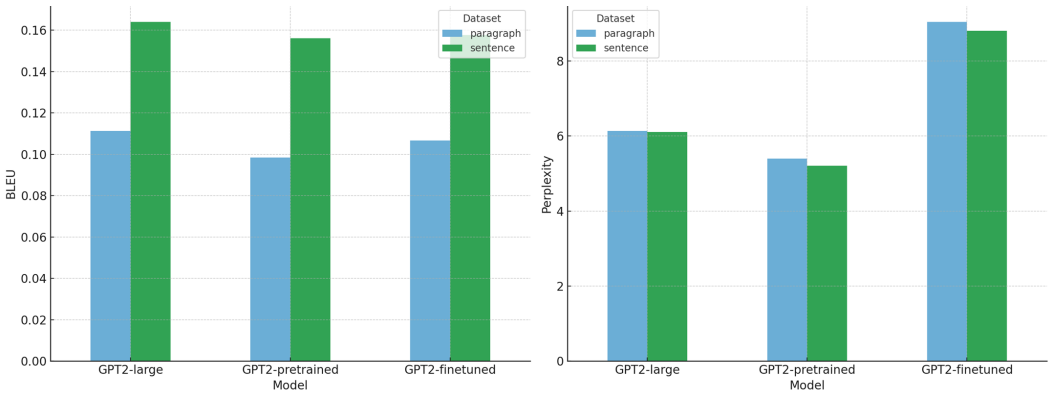

Figure 5: Performance comparison of GPT2 models on BLEU (left) and Perplexity (right) metrics for paragraph-level and sentence-level datasets. The evaluation highlights improvements in BLEU and reductions in Perplexity as the model progresses from GPT2-Large [6] to GPT2-pretrained and GPT2-finetuned stages.

In Figure 5, we present the experimental results of BLEU and perplexity for our models from the GPT2 series. We compared the performance of three models: the original GPT2-Large model, GPT2-pretrain, and GPT2-finetuned. The results indicate comparable BLEU scores across the three models, while the GPT2-pretrain demonstrated superior performance in terms of perplexity metrics.

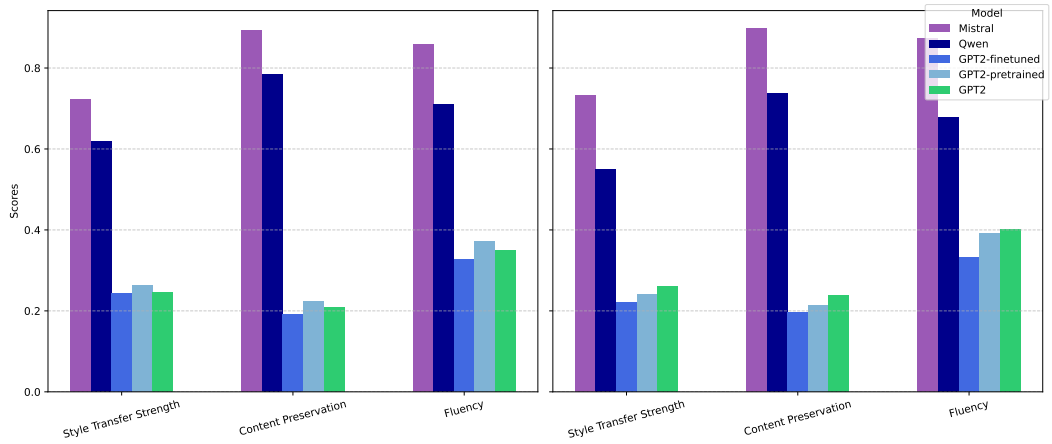

Figure 6: Evaluation of different models on style transfer tasks through AI Feedback on sentence test set (left) and paragraph test set (right). The charts compare Style Transfer Strength, Content Preservation, and Fluency scores across Mistral (Mistral-7B-v0.2), Qwen (Qwen2.5-1.5B-Instruct), GPT2-finetuned, GPT2-pretrained and GPT2 (GPT2-Large).

The results of the AI feedback evaluation are presented in Figure 6, with the left panel showing the results for the sentences and the right panel displaying the results for the paragraphs. The analysis reveals that the pre-trained model achieved optimal performance in sentence-level transformations, while models trained on paragraph-level data demonstrated moderate performance.

4 Conclusion

In this work, we have developed a comprehensive framework for text style transfer that includes data set construction, model training, and evaluation methodologies. Our approach includes a novel dataset for continued pretraining and supervised fine-tuning, along with robust evaluation metrics combining automated measures and AI feedback. Through extensive experimentation, we demonstrated that a relatively compact GPT2-Large model (1.5B parameters) can achieve promising style transfer quality after pretraining, although it does not yet match the performance of state-of-the-art models of similar size or larger architectures.

Our investigations yielded several important insights regarding model architectures and training approaches. The T5 model [7] proved unsuitable for this task due to fundamental architectural

limitations, failing both in the pre-training and fine-tuning scenarios. Furthermore, the suboptimal fine-tuning results observed with GPT2 [6]can probably be attributed to its limited instruction follow-up capabilities. These findings contribute valuable insights to the ongoing development of efficient and effective text style transfer systems, while highlighting important considerations for model selection and training strategies in specialized NLP tasks.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024.

[3] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[4] Sourabrata Mukherjee and Ondrej Dušek. Text style transfer: An introductory overview, 2024. URL https://arxiv.org/abs/2407.14822.

[5] OpenAI. Introducing openai o1. https://openai.com, 2025. Accessed: January 8, 2025.

[6] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[8] Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. The acl ocl corpus: advancing open science in computational linguistics. *arXiv preprint arXiv:2305.14996*, 2023.

[9] Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

[10] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.