

Fine-Tuned Reasoning for Writing Style Analysis

Notebook for the PAN Lab at CLEF 2024

Xiaofeng Liang*, Fanzhi Zeng, Yan Zhou, Xiangyu Liu and Yuexia Zhou

¹Foshan University, Foshan, China

Abstract

Multi-author writing style analysis aims to identify whether different paragraphs are written by the same author. This paper presents a method based on the Fine-tune-CoT approach, incorporating the idea of few-shot prompts. Leveraging GPT-3.5, the method generates thought chains and corresponding answer datasets for given questions, followed by fine-tuning on a small-scale model T5-small to accomplish the task. Experimental results on the PAN 2024 multi-author writing style analysis test dataset demonstrate the effectiveness of the proposed method, yet indicating significant room for improvement.

Keywords

Writing Style Analysis, Chain-of-Thought, Few-shot Prompt

1. Introduction

The task of multi-author writing style analysis involves determining whether consecutive paragraphs in a text are authored by the same individual [1, 2]. Effective detection and analysis of writing styles enable the identification of potential authorship changes within a text and evaluation of stylistic consistency across paragraphs. This task finds broad applications in academic integrity investigations for plagiarism detection or ghostwriting identification, as well as in literary studies for attributing anonymous works or verifying collaborative authorship [3]. The primary approaches to this task typically involve extracting features from text samples and employing a trained discriminator model to assess the similarity of extracted features across different segments [4]. However, the task presents various challenges such as text style diversity and cross-domain generalization, necessitating solutions to these complexities [5].

This paper presents our methodology for the multi-author writing style analysis task at PAN 2024. We utilize a teacher model, GPT-3.5 [6], to generate a dataset of thought chains based on the original dataset. Subsequently, we fine-tune this dataset on a student model, T5 small. Finally, we submit our operational results on TIRA.io to evaluate the performance of our approach in practical applications [7].

2. Background

Multi-author writing style analysis stands as a pivotal research area within natural language processing. Traditional methods predominantly rely on manual feature extraction and conventional machine learning models such as support vector machines, naïve Bayes, and decision trees. These approaches entail manual extraction of features like sentence length, specific word frequencies, and syntactic structures from text, followed by employing classification algorithms to differentiate between different authors. However, these methods often underperform when dealing with complex and diverse human language, especially with large-scale and diverse datasets. With the evolution of deep learning, neural network-based approaches have significantly propelled advancements in the field of writing style analysis. Particularly, Transformer models, excelling in handling long-range dependencies and

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[†] Authors contributed equally

✉ 2794918956@qq.com (X. Liang*); coolhead@126.com (F. Zeng); zhouyan791266@fosu.edu.cn (Y. Zhou); xyliu1805@fosu.edu.cn (X. Liu); fs_zyx@fosu.edu.cn (Y. Zhou)

ORCID 0009-0002-6408-0639 (X. Liang*)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

understanding contextual nuances, have demonstrated outstanding performance in writing style analysis tasks. Pre-trained language models based on Transformers, such as BERT [8], RoBERTa [9], and DeBERTa [10], have found extensive usage in this task.

In tasks such as multi-author writing style analysis, a common approach involves concatenating two sentences into sentence pairs, and then fine-tuning pre-trained models (such as BERT) or directly applying classification learning. This method leverages the pre-trained model’s understanding of text semantics to differentiate authorial style features by assigning different labels to sentence pairs. Such approaches are widely applied in natural language processing for various text classification and text pair tasks.

Innovative techniques like few-shot learning and thought chain prompting further enhance the performance of neural network models in natural language processing tasks. Few-shot learning aims to enable models to perform well even with minimal training samples, which is particularly advantageous for tasks with limited annotated data. On the other hand, thought chain prompting enhances the model’s reasoning abilities by guiding it through logical steps of inference. The Fine-Tune-CoT [11, 12] method amalgamates the advantages of few-shot learning and thought chain prompting to enhance performance through model fine-tuning. This method leverages pre-trained large-scale language models to generate intermediate inference steps and answers, serving as auxiliary datasets for fine-tuning smaller and more efficient models. This fine-tuning process enables smaller models to learn logical reasoning and patterns captured by larger models, thereby exhibiting better performance on the target task.

3. System Overview

3.1. Datasets

The training dataset provided by the PAN@CLEF 2024 organization is sourced from user posts across various subreddits on the Reddit platform. It comprises English text solutions to different problem instances, with each text corresponding to a solution for a specific problem. For each problem instance in the training set, two files are provided. One file contains the solution to the problem, which consists of English text organized into paragraphs. The other file contains information about the number of authors associated with the solution and records whether there are style changes between paragraphs. The dataset includes three difficulty levels: easy, medium, and hard, categorized based on the diversity of topics in the paragraphs. The hard level comprises solutions with only one topic. Table 1 outlines the distribution of the datasets across different difficulty levels.

Table 1
Distribution of the number of datasets with different difficulty levels

Type	Train	Validation
Easy	4200	900
Medium	4200	900
Hard	4200	900

3.2. Methods

In this paper, a methodology is proposed wherein a large-scale model generates thought chains and corresponding answers for given questions, followed by fine-tuning on a smaller pre-trained model. Initially, the training data undergoes simple preprocessing to remove redundant empty lines and special characters, ensuring data consistency. Subsequently, based on the Fine-Tune CoT procedure, the remaining tasks are divided into three steps to be completed.

3.2.1. Generate Datasets

In the first step of generating our inference data, we initially divide each text into paragraphs. Each time, we input adjacent pairs of paragraphs into GPT-3.5, as shown in Figure 1. In the input phase, concrete examples guide the model to reason according to the provided instances, while in the output phase, the model generates the corresponding thought chain for the given question, along with the answer.

The input-output process of the specific examples in Figure 1 can be abstracted as Equation (1).

(1) Input and Output template

- 1) Input: Instruction + CoT guidance + Example question + Example question reasoning + Example question answer + Instruction + CoT guidance + Question,
- 2) Output: Answer

The output obtained through Equation (1) consists of the answer (a_i) and the question reasoning (r_i). We integrate and assemble the original question (q_i) of this instance with the output using the following Equation (2) to form an instance S_i .

(2) S_i Template

Q:< q_i >.A:Let's think step by step < r_i > The answer is < a_i >

In this paper, the sampling parameters adopted are as follows: TOP-P: 0.95, Temperature: 0.7.

3.2.2. Organize Datasets

In the first step, all output instances S obtained undergo filtering and format reconstruction. During the filtering stage, examples that are incorrect in the reasoning process are removed based on the real labels in the dataset. Subsequently, the data is formatted and reassembled. We organize the instances S into a new sample set $T=(p_i, c_i)$, representing the text data structure required for fine-tuning the model, where the prompt section consists of the original question $Q:<q_i>$, and the result section, c_i , includes the reasoning process and answer.

3.2.3. Model Fine-tuning

This paper employs the T5 small model as the student model and utilizes Low-Rank Adaptation (LoRA) for fine-tuning. The T5 model obtained in the second step serves as the fine-tuning dataset. The hyperparameters for fine-tuning are set as follows: r (the rank of matrix decomposition): 8, Epochs: 3, Learning Rate: $2e-5$, Weight decay: 0.01.

4. Results

To evaluate the proposed model, we employ the TIRA evaluation tool, which includes the F1 score as a metric, which is the harmonic mean between precision and recall.

Our results on the PAN24 Multi-Author Writing Style Analysis task test set are presented in Table 2, from which it is evident that although our fine-tuned T5 small model performs better than the baseline on tasks of all three difficulty levels, the overall performance still remains relatively poor. Particularly, the performance on tasks of medium and hard difficulty is notably unsatisfactory. Therefore, in future work, we plan to enhance the fine-tuning process by augmenting the fine-tuning data. Specifically, we may guide the fine-tuning process using data where the teacher model's reasoning is incorrect, aiming to derive the correct answers through reasoning rather than simply discarding erroneous inference data. Additionally, we will introduce multiple thought chains instead of a single chain to enrich the dataset.

Task: Determine whether the following two paragraphs are written by the same author. Answer by following the step-by-step thought process provided and conclude with a judgment of 0 (same author) or 1 (different authors)

Example 1:

Paragraph A: "The quick brown fox jumps over the lazy dog. It's a commonly used sentence in typing practice."

Paragraph B: "An interesting fact about pangrams is that they include every letter of the alphabet at least once."

Step-by-step thought process:

1. Topic: Both paragraphs mention a sentence used in typing practice.
2. Style: Paragraph A has short and direct sentences, while Paragraph B provides more background information and details.
3. Vocabulary: Both paragraphs use some overlapping words like "sentence" and "alphabet," but Paragraph B uses more uncommon words like "pangrams."
4. Conclusion: Although both paragraphs mention language practice, the differences in style and vocabulary suggest they might not be written by the same author.

Judgment: 1

Example 2:

Paragraph A: "In recent years, climate change has become a significant global issue. Many countries are now taking steps to reduce their carbon footprint."

Paragraph B: "Countries are implementing various measures to tackle the adverse effects of climate change, including investing in renewable energy and promoting sustainable practices."

Step-by-step thought process:

1. Topic: Both paragraphs discuss climate change and the measures countries are taking.
2. Style: Both paragraphs have a very formal writing style, using similar sentence structures.
3. Vocabulary: Both paragraphs use similar vocabulary such as "climate change," "countries," "measures," "renewable energy."
4. Conclusion: The consistency in topic, style, and vocabulary suggests these paragraphs are likely written by the same author.

Judgment: 0

Now, you determine if the following two paragraphs are written by the same author. If they are written by different authors, output 1; if they are written by the same author, output 0.

Paragraph A: "Yep. I just did a full text search on the Constitution and neither guns nor firearms were in there. Who knew guns were in the Constitution as much as abortions?"

Paragraph B: "/r/politics is currently accepting new moderator applications. If you want to help make this community a better place, consider !"

-
1. Topic: Both paragraphs seem to be discussing different topics. Paragraph A talks about the absence of certain terms in the Constitution, while Paragraph B mentions moderator applications for a subreddit.
 2. Style: Paragraph A has a casual tone with contractions like "Yep" and colloquial language, whereas Paragraph B is more formal and informational.
 3. Vocabulary: Paragraph A uses terms like "Constitution," "guns," and "abortions," while Paragraph B uses terms like "/r/politics," "moderator applications," and "community."
 4. Conclusion: The differences in topic, style, and vocabulary strongly suggest that these paragraphs are written by different authors.

Judgment: 1

Figure 1: Example of Judging Sentence Style Changes

Table 2

Overview of the F1 accuracy for the multi-author writing style task in detecting at which positions the author changes for task 1, tas 2, and task 3.

Approach	Task 1	Task 2	Task 3
freezing-skeleton	—	—	0.484
coplanar-color	0.606	—	—
fundamental-stool	—	0.455	—
Baseline Predict 1	0.466	0.343	0.320
Baseline Predict 0	0.112	0.323	0.346

5. Conclusion

This paper employs a teacher model to obtain thought chains and answers, then fine-tunes a student model for multi-author writing style analysis tasks. Table 2 presents the final test set results, which indicate that the performance on the final test set is not satisfactory. This may be attributed to the relatively small amount of data used for fine-tuning in the experiment. During the data filtering stage, a large number of samples were discarded, resulting in a final fine-tuning dataset that was too small. Additionally, the experiment did not consider multiple thought chains for a single question, leading to a lack of diversified reasoning.

References

- [1] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Multi-Author Writing Style Analysis Task at PAN 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
- [2] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.
- [3] H. A. Maurer, F. Kappe, B. Zaka, Plagiarism-a survey., *J. Univers. Comput. Sci.* 12 (2006) 1050–1084.
- [4] E. Stamatatos, A survey of modern authorship attribution methods, *Journal of the American Society for information Science and Technology* 60 (2009) 538–556.
- [5] M. Mozafari, R. Farahbakhsh, N. Crespi, Hate speech detection and racial bias mitigation in social media based on bert model, *PloS one* 15 (2020) e0237861.
- [6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [7] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.
- [8] J. Lee, K. Toutanova, Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 3 (2018) 8.

- [9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
- [10] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).
- [11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.
- [12] N. Ho, L. Schmid, S.-Y. Yun, Large language models are reasoning teachers, arXiv preprint arXiv:2212.10071 (2022).