

A Natural Language Technology Attempt for Author Attribution Issues in a Chinese Classical Novel

理学専攻・情報科学コース
2040662
Zhou Xinyi

1 Introduction

The book, *Dream of the Red Chamber*, is regarded as the greatest novel in Chinese literature.

Professor Chen Dakang, wrote a book titled *The Economic Accounts of Rong Mansion* [3].

In Chen's book, he uses traditional statistical and mathematical methods to analyze the economic operations of the Jia family, which serves as the backdrop in *Dream of the Red Chamber*.

It inspired me that I also hope to take up the baton from my predecessors by leveraging the latest deep learning techniques.

Dream of the Red Chamber has numerous versions, with the most widely circulated one being the 120-chapter edition attributed to Cao Xueqin. However, it has been widely observed that this version exhibits significant changes in writing style.

Finally, a consensus was reached: only the first 80 chapters of Cao Xueqin's work have been passed down, and the chapters after the 80th were written and compiled by others at the time [5].

This study will be conducted based on this premise, examining whether the performance of deep learning models aligns with this argument.

2 Model

2.1 BERT

BERT is a bidirectional encoder model based on the Transformer architecture.

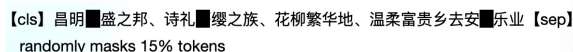
The Transformer introduces the Multi-Head Attention module within its encoder and decoder structures, significantly enhancing its ability to model sequential information. Multi-Head Attention consists of multiple Self-Attention mechanisms, enabling each encoder block to receive and integrate the outputs from previous blocks. This innovative attention mechanism endows the Transformer with a strong capability to model long-range dependencies [4].

Building upon this foundation, BERT retains and extends the Transformer's encoder structure while discarding the decoder. By leveraging bidirectional context modeling and pretraining strategies, BERT achieves powerful language understanding capabilities. This architecture significantly improves the model's performance on specific tasks. In this study, we selected the MLM (Masked Language Model) task for pre-training, with its detailed principles explained in subsequent sections [2].

2.2 BERT-WWM

Chinese has its unique characteristics. Therefore, tokenization in Chinese poses higher demands

on a model's comprehension capabilities. We selected BERT-WWM, the Whole Word Masking model, which randomly masks words in the text. This enables the model to train its understanding of the dataset by continuously predicting the masked words. [1]. MLM task process sentences like fig 1



```
【cls】 昌明■盛之邦、诗礼■纓之族、花柳繁华地、温柔富贵乡去安■乐业 【sep】
randomly masks 15% tokens
```

Figure 1: Sentence Processing with WWM Task

2.3 Pretrained BERT-WWM

The previously discussed BERT-WWM model was pretrained using a corpus of modern Chinese encompassing multiple domains. However, the linguistic characteristics of *Dream of the Red Chamber* differ in some aspects.

Firstly, *Dream of the Red Chamber* was written during the transitional period from classical Chinese to modern Chinese. While its expression largely aligns with modern Chinese, it still retains some classical Chinese habits in terms of vocabulary and stylistic details. Additionally, as a novel, its narrative structure demands coherence between plot and character interactions.

Considering these factors, several novels with linguistic styles and publication periods similar to *Dream of the Red Chamber* were specifically selected to form a separate training corpus. We also use WWM task to train the model.

The resulting model is referred to as Pretrained Bert-WWM.

3 Experiment

3.1 Procedure

The book contains a total of 120 chapters, with the first 80 chapters labeled as the Y and the last 40 chapters labeled as the N. To maintain the original dataset ratio of Y:N = 2:1, 16 chapters from the Y dataset and 8 chapters from the N dataset (24 chapters in total) are selected as the test set for each round of experiments. Consequently, all 120 chapters can be tested across five experimental rounds.

In each round of experiments, the chapters removed for the test set are used as the training set with Y/N labels for fine-tuning the model. Additionally, one-eighth of the chapters in the training set are used as a validation set.

3.2 Performance Standards

In each test, we first evaluate the model's average performance using conventional metrics.

Then, we select the two better-performing mod-

els. Subsequently, we segment each chapter into sentences, label each sentence, and calculate the proportion of Y/N sentences within each chapter.

For the evaluation of sentence proportions, we expect the model to correctly label a higher number of sentences in each chapter. Correct labeling is not merely defined as exceeding 50%; we further expect the model to correctly label more than 70% of the sentences. This threshold ensures that the model can clearly determine whether a chapter was written by the original author. For chapters where the proportion of correctly labeled sentences remains between 50% and 70%, we regard the model’s performance as ”correct but ambiguous.”

4 Results and Analysis

The average performance data of each model tested sequentially on all 120 chapters is shown in Table 1:

Table 1: Comparison of Accuracy by Model

Metric	BERT	BERT-WWM	Pretrained BERT-WWM	Best Model
acc	0.4516	0.5671	0.5777	Pretrained Bert-wwm
pre	0.4484	0.5412	0.5326	Bert-wwm
rec	0.4206	0.4596	0.4697	Pretrained Bert-wwm
f1	0.4259	0.4790	0.4818	Pretrained Bert-wwm

It can be observed that BERT performs the worst, while BERT-WWM, which is pre-trained specifically for Chinese, improves the performance. Ultimately, the Pretrained BERT-WWM model, which underwent pretraining on datasets similar to the Downstream Task, achieves the best performance, ranking first in three out of the four metrics.

The two better-performing models, BERT-WWM and Pretrained BERT-WWM, were selected. Each chapter was segmented into sentences, and each sentence was classified as Y/N. The Y/N classification ratios are as follows:

fig 2 is the sentences labeled percentage in each chapter by BERT-WWM model

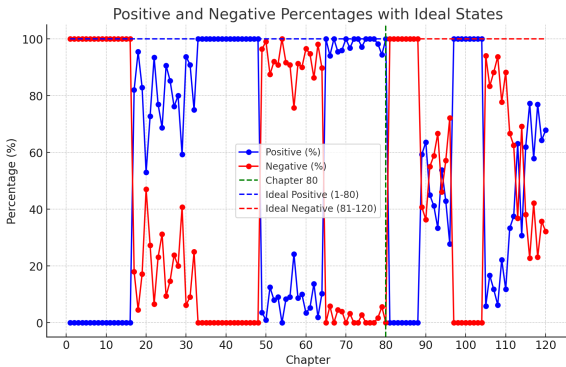


Figure 2: Sentences labeled percentage by BERT-WWM

It can be observed that Pretrained BERT-WWM outperforms in both the number of chapters where more than 50% of the sentences are correctly classified and the number of chapters where more than 70% of the sentences are clearly classified.

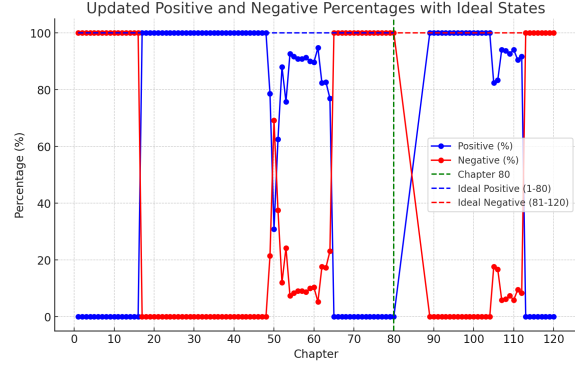


Figure 3: Sentences labeled percentage by pretrained BERT-WWM

5 Conclusion

The comparison among the models supports the conclusion pretraining on datasets similar to the Downstream Task can significantly improve the model’s performance.

Meanwhile, the best-performing model, Pretrained BERT-WWM, assigned a single classification to over 90% of the sentences in the vast majority of chapters. This demonstrates that the model supports the conclusion that Dream of the Red Chamber exhibits two distinctly different writing styles, further validating the findings of literary research.

References

- [1] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.
- [3] 陈大康. 荣国府的经济账. 人民文学出版社, 北京, 2019. 平装.
- [4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017.
- [5] 胡适. 红楼梦考证. 世界书局, 上海, 1947. 再版修订版.