In-Context Language Modelling For Paper Source Tracing

YunGui Zhuang

Algorithm Engineer,Tgnet,GuangZhou

Abstract

With the rapid advancement of technology, the number of academic papers is surging at an alarming rate. Millions of papers are published globally every year, and the number continues to climb. It is becoming increasingly challenging for researchers to trace the pulse of technological development from such a vast ocean of literature. For this reason, we propose a new approach: a comprehensive analysis using the contextual information of a paper as well as other literature it cites. This approach demonstrates significant results on a test set and a rapid inference process, providing researchers with an efficient tool for literature screening and analysis.

1 Introduction

The purpose of the source tracing task is to find a ref-source from a paper, given the full text of the paper p. The ref-source is the most important reference (called the "source paper"), generally the one that most inspired[1] the paper. Each paper may or may not have one or more ref-sources, and for each reference in the paper, the paper's source is given an importance score in the range [0, 1].
The following points define whether a reference is a ref-source paper:

1. whether the main idea of paper p was inspired by that reference
2. or whether the main approach of the paper p is derived from that reference
3, Is the reference essential to the paper p? i.e., the paper p would not have been possible without the work of that reference.

In order to gain a deeper understanding of the nature of scientific progress, how can we simplify citation networks to trace[2] the origins of publications and reveal mutually inspiring relationships between papers? An intuitive idea might be to treat the most cited literature in each paper as its source of inspiration and ignore other citation links. But this approach is not accurate.

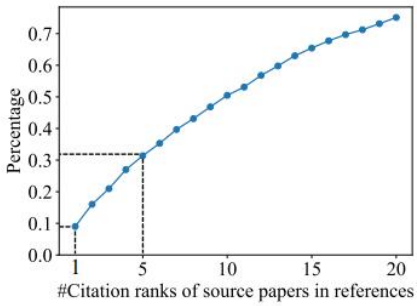


Figure 1: The cumulative distribution function (CDF) of the references' citation ranks for source papers.

By analysing about 1500 papers in computer science and their professionally labelled source literature, we show in Figure 1 the cumulative distribution function (CDF) of the citation rankings of these sources. The data shows that if we consider only the most cited literature as the source, the identification accuracy is well below 10%. Going further, we find that the true source of about 70% of the papers is not among their top 5 cited literatures. This finding reverses our traditional perception of citation counts as the primary basis for determining the source of a publication.

2 Methodology

We When dealing with academic paper data, we adopt an innovative approach to transform it into a binary classification problem to improve the accuracy of identifying interrelationships and inspirations between papers. We first integrate the title of each paper, the number of the cited paper, and the associated contextual information into a unique data structure in the format of "Title [SEP] Citation Number [SEP] Context". This structure is designed to capture the basic information and contextual links of the paper, providing rich information for the model.

In order to balance the ratio of positive and negative samples during model training, we adopt a 1:4 ratio for sample construction, i.e., for every positive sample (i.e., papers that have explicit citation or inspiration relationships with each other), we generate four negative samples (i.e., pairs of papers that do not have such a relationship). Such a ratio helps the model to better learn and distinguish real citation relationships.

For model selection, we used SciBERTa-cs, which is a pre-trained BERT model specifically designed for the computer science domain, and which provides an in-depth understanding of the terminology and concepts in computer science. By fine-tuning the SciBERTa-cs model with five-fold cross-validation, we ensured that the model maintains a high generalisation ability across different datasets.

Ultimately, the output of the model is transformed by a sigmoid function to obtain the predicted probabilities for each sample. We take the average of these probabilities as the final prediction result, and this method can reduce the randomness of single prediction and improve the stability and reliability of prediction.

With this method, we can not only identify more accurately the citation and inspirational relationships between papers, but also provide a powerful tool for academic researchers to help them quickly discover valuable information and connections from the vast amount of literature.

3 Experiments

This study aims to enhance the performance of text classification tasks through natural language processing techniques. The experimental process is divided into the following stages:

Baseline Model Training

In the initial phase, we employed the DeBERTa-base model as the baseline model.The dataset was constructed by filtering positive and negative samples in a 1:4 ratio for binary classification.During the preliminary competition, the model achieved a score of 36.48%.

Model Fine-tuning and Activation Function Optimization

To further improve model performance, we replaced the baseline model with SciDeBERTa-CS and fine-tuned it. During this process, we also changed the activation function from the default ReLU to Sigmoid, expecting a smoother output distribution. After these adjustments, the model's score increased to 40.00%.

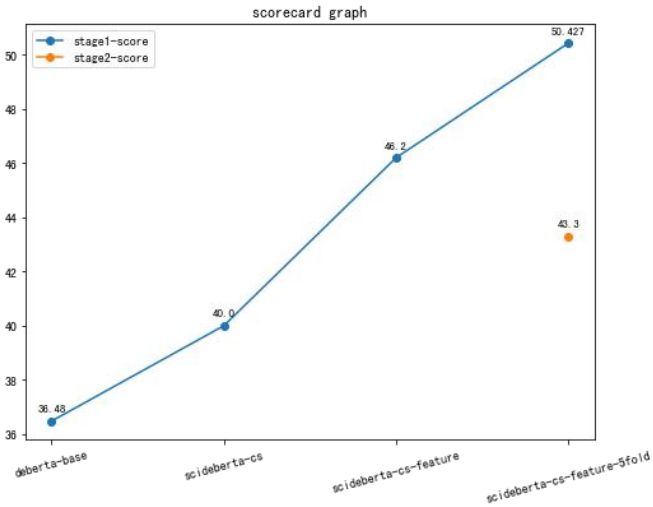
Post-processing Strategy

To further enhance the model's predictive accuracy, we introduced specific post-processing strategies. With these strategies, the model's score was further optimized to 46.20%.

Contextual Information Enhancement

Finally, we enhanced the model's ability to process contextual information. Using a 5-fold cross-validation method, the model's score ultimately increased to 50.427%.

Below is a chart of the final score:



4 Discussion

The rapid growth of academic literature poses significant challenges for researchers seeking to understand the evolution and interconnections within their fields. The sheer volume of published papers, as highlighted by the Scopus database, underscores the necessity for effective literature tracing and citation analysis tools. The task of Paper Source Tracing (PST) is particularly relevant in this context, as it aims to identify the most influential references, or "source papers," that have significantly inspired or are indispensable to a given paper.

Our experiments, starting with the DeBERTa-base model and progressing to SciDeBERTa-CS with a sigmoid activation function, demonstrate a clear trend of performance improvement. The initial baseline score of 36.48% indicates the complexity of the PST task. The subsequent increase to 40.00% with model fine-tuning and the adoption of a sigmoid function suggests that model architecture and activation function selection are critical factors in enhancing citation relevance detection.

The introduction of post-processing strategies, leading to a score of 46.20%, highlights the importance of refining the output to better align with the criteria for identifying source papers. The final enhancement through increased context sensitivity[3], resulting in a score of 50.427%, underscores the significance of understanding the broader research landscape in which a paper is situated.

However, it is important to acknowledge the limitations of our approach. The reliance on automated methods may overlook subtle nuances in citation practices and the subjective judgment of what constitutes a "source paper." Furthermore, the dynamic nature of academic research means that the criteria for identifying source papers may evolve, necessitating continuous updates to our models and methods.

5 Conclusion

In conclusion, the PST task is a critical component of modern academic research, facilitating a deeper understanding of the intellectual lineage and development of ideas within scholarly communities. Our experiments have shown that with careful model selection, fine-tuning, and contextual enhancement, it is possible to significantly improve the identification of source papers. However, this task remains complex and requires a nuanced approach that balances automated analysis with human insight.

As we look to the future, the integration of advanced machine learning techniques, such as deep learning and natural language processing, holds promise for further advancements in PST. Additionally, the development of more sophisticated post-processing algorithms and the incorporation of expert knowledge will be essential to refine the accuracy and reliability of source paper identification.

The PST task not only aids individual researchers in navigating the vast sea of academic literature but also contributes to the broader academic endeavor of knowledge synthesis and innovation. By enhancing our ability to trace the origins and influences of scholarly work, we can better appreciate the interconnectedness of ideas and the cumulative nature of scientific progress.

References

[1] Zhang, F., Cao, K., Cen, Y., Yu, J., Yin, D., & Tang, J. (2024). PST-Bench: Tracing and Benchmarking the Source of Publications. ArXiv, abs/2402.16009.

[2] Yin, D., Tam, W.L., Ding, M., & Tang, J. (2023). MRT: Tracing the Evolution of Scientific Publications. IEEE Transactions on Knowledge and Data Engineering, 35, 711-724.

[3] Hassan, S., Akram, A., & Haddawy, P. (2017). Identifying Important Citations Using Contextual Information from Full Text. 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 1-8.