Grafting Learning Is All You Need

Haoru Chen
MeiTuan
BeiJing, China
chenhaoru02@meituan.com

Xiaocheng Zhang
MeiTuan
BeiJing, China
zhangxiaocheng@meituan.com

Yang Zhou
MeiTuan
ShangHai, China
zhouyang96@meituan.com

Mengjiao Bao
MeiTuan
BeiJing, China
baomengjiao@meituan.com

Peng Yan*
MeiTuan
BeiJing, China
yanpeng04@meituan.com

Abstract

This paper presents the solution of team BlackPearl in the KDD Cup 2024 OAG Challenge - PST (paper source tracing).

The goal of this competition is to identify "ref-sources" from the full texts of a given paper. A ref-source refers to the most important reference (called the "source paper"), which generally refers to the literature that has provided the greatest inspiration for this paper.

Our solution proposes an LLM (Large Language Models) system based on grafted learning, which fully leverages all noisy and noiseless data, transferring the output confidence of BERT models to the LLM. Additionally, we have developed an automatic feature engineering pipeline based on RAG (Retrieval-Augmented Generation), effectively supplementing the knowledge graph information of the paper. Our method ranks 1st place in the final leaderboard of Task PST. Our solution and code are publicly available at this link: https://github.com/BlackPearl-Lab/KddCup-2024-OAG-Challenge-1st-Solutions/tree/main.

CCS Concepts

• **Computing methodologies** → Information extraction.

Keywords

Grafting Learning, Large Language Models, Knowledge Graph, Bert, Paper Source Tracing, KDDCup 2024

ACM Reference Format:
Haoru Chen, Xiaocheng Zhang, Yang Zhou, Mengjiao Bao, and Peng Yan. 2024. Grafting Learning Is All You Need. In *Proceedings of Proceedings of KDD 2024 OAG-Challenge Cup (KDDCup '24).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Due to the swift advancement of technology, there has been an exponential increase in the volume of research papers. Millions of papers are published globally every year, and the number of publications continues to rise. According to the Scopus database, as of 2021, the number of academic papers reached 220 million, including all academic papers published since the 17th century, covering various fields such as natural sciences, social sciences, humanities, and more. For researchers, it has become increasingly difficult to grasp the ins and outs of technological development from numerous literature sources.

To improve the performance of the PST (paper source tracing) task, Zhipu AI provided the "Tracing and Benchmarking the Source of Publications Dataset," which contains thousands of papers. It hosted the KDD CUP 2024 OAG Challenge[13] PST[12].

1.1 Dataset Description

The dataset is divided into three parts: the manually annotated supervised training dataset, the rule-labeled supervised training dataset, and the unsupervised large-scale knowledge graph dataset (DBLP[8][9][10]). The input for the supervised datasets is divided into two parts: the original text information of the paper and the list of important references (source papers). The rule-labeled supervised training dataset is collected using rule-based methods from the paper dataset. Specifically, references appearing in the context of keywords such as "motivated by" and "inspired by" are extracted; hence, this dataset contains a significant amount of noisy labels. The DBLP dataset contains extensive knowledge graph information with data attributes including paper authors, institutions, citation counts, citation graphs, publication years, journals/conferences, abstracts, keywords, etc. The fundamental statistics of the datasets are summarized in Table 1.

Table 1: Fundamental Statistics of The Datasets

Dataset	Data Volume
Manually Annotated Dataset	788
Rule-Labeled Dataset	4854
Validation Dataset	394
Test Dataset	394
DBLP Dataset	6,404,472

1.2 Task Description

The purpose of the paper source tracing task is to identify "ref-sources" from the full texts of a given paper. A ref-source refers

*Corresponding author of this research.

KDDCup '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

to the most important reference (called the "source paper"), which generally refers to the literature that has provided the greatest inspiration for this paper. The following points define whether a reference is a source paper:

1. Is the main idea of paper p inspired by the reference?
2. Is the core method of paper p derived from the reference?
3. Is the reference essential for paper p? Without the work of this reference, paper p cannot be completed.

Participants in the PST task must output an importance score (between 0 and 1) for each reference paper in the test set. The importance score should be higher for references that are more likely to be the source papers of the given paper. The online evaluation metric employed is MAP (Mean Average Precision).

2 METHODOLOGY

In traditional knowledge graph link construction, the PST task is usually abstracted as a text binary classification task. This method constructs text pairs between the paper and the reference and then performs binary classification. Each reference of the paper gets a quantified importance score, and then the importance scores of all references can be sorted to obtain the final importance ranking. In this competition, we also adopt the method of converting the ranking task into a binary classification task. Our solution includes multiple instances of transfer learning, such as transferring rule-labeled data to the manually annotated dataset using BERT and transferring BERT's ability to process the original paper's XML contextual information to LLM. The final transfer learning pipeline of our solution is shown in Figure 1.

2.1 Data Process

In this competition, the given original paper data is presented in the form of XML files, and participants are required to extract the relevant information from the XML. After experimental validation, there are two effective extraction methods:

Assuming the current task is to perform pair binary classification for the 13th reference.

1. Extract the context of the citation markers for the specified references in the XML format. A specific example is shown below:

> Most of the early attempts of remote sensing target detection<ref type="bibr" target="b1">[2]</ref>-<ref type="bibr" target="b5">[6]</ref>are designed with the help of some specifically designed hand-crafted features and supervised classification algorithms. Recent advances in remote sensing target detection methods<ref type="bibr" target="b6">[7]</ref>-**<ref type="bibr" target="b12">[13]</ref>**have been primarily focusing on deep learning based detection methods, especially the ones based on convolutional neural networks.

Note that here only the text information of one citation context is extracted. In practice, all citation contexts will be searched and the texts will be concatenated together. The method of locating the original text context based on the reference number can easily achieve a good score and results in shorter text length. However, some reference citations do not appear in the original text, leading to 40% of the extracted data being null.

2. Filtering out all formatting symbols in the XML format, retaining only the plain text information and citation markers, while preserving the full text without truncation. A specific example is shown below:

> Most of the early attempts of remote sensing target detection[2][6]are designed with the help of some specifically designed hand-crafted features and supervised classification algorithms. Recent advances in remote sensing target detection methods[7]**[13]**have been primarily focusing on deep learning based detection methods, especially the ones based on convolutional neural networks.

Note that here only a paragraph of the paper is extracted for demonstration purposes; in actual use, the full text of the paper is retained. Although this method results in longer text lengths, it preserves all the effective information.

2.2 Feature Engineering and RAG

In the DBLP large-scale knowledge graph dataset, we can obtain a lot of auxiliary information. In this competition, we use the paper titles to construct text vectors and RAG[6] related auxiliary information from the DBLP dataset. To fully utilize this auxiliary information, we constructed the following features:

- **`Citation Count of the Paper`**: The number of times the paper has been cited.
- **`Year of Publication`**: The year in which the paper was published.
- **`Conference or Journal of the Paper`**: The conference or journal where the paper was published.
- **`Abstract and Keywords of the Paper`**: The summary and keywords associated with the paper.
- **`First Three Authors' Names and Institutions`**: The names and affiliations of the first three authors of the paper.
- **`Total Citation Count of the First Three Authors`**: Represents the level of the authors based on their total citation count.
- **`Total Citation Count of the Authors' Institutions`**: Represents the level of the institutions based on their total citation count.
- **`Total Citation Count of the Conference or Journal`**: Represents the prestige of the conference or journal based on its total citation count.

Finally, we integrated these features with the original text and concatenated them before inputting them into the model.

2.3 Grafting Learning

Grafting learning refers to a new paradigm of transfer learning. If there are two different datasets with inconsistent data distributions but consistent prediction targets, or if the label noise distributions differ, grafting learning can be applied. In typical NLP competitions, training on differently distributed data is usually done on the same model using techniques like warm start or pretraining. However, due to inconsistent output spaces, one model cannot fully utilize the effective information from both datasets. Grafting learning involves first training a model on the first dataset (the one more deviated from the prediction target), then using this model to predict on the second dataset (the one closer to the prediction target). The

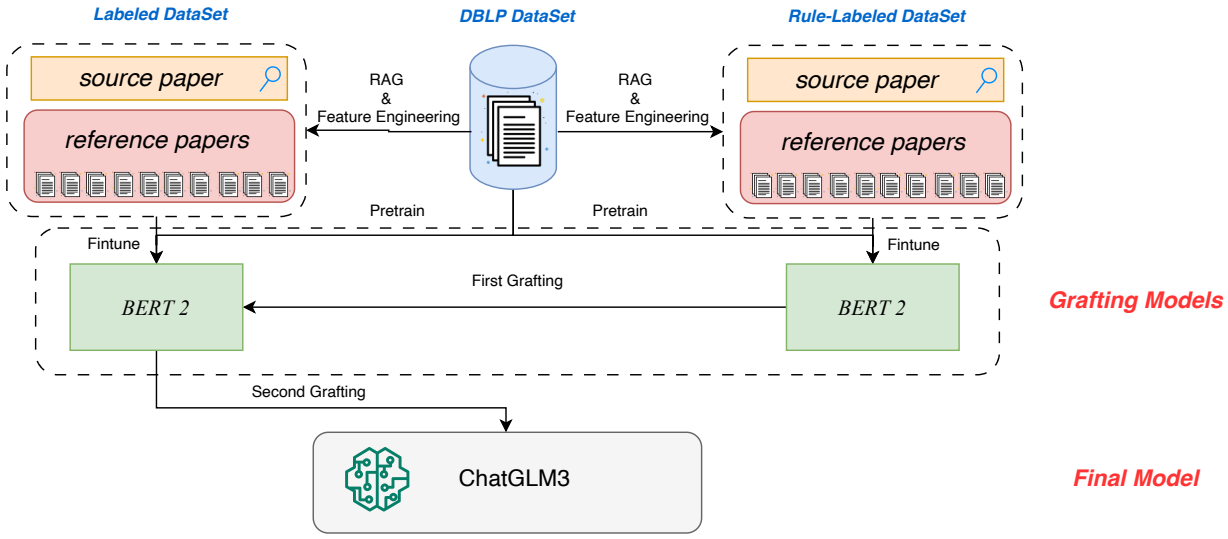

Figure 1: Pipeline of Grafting Learning

prediction results are then used as one of the features for training a second model on the second dataset.

Grafting learning can fully utilize the effective information from differently distributed data while avoiding conflicts between different datasets. Under the operation of grafting learning, the model makes full use of all available key textual information, resulting in more confident output results.

In this competition, we adopted grafting learning twice, as detailed below:

1. First Grafting

We first trained a BERT[1] model on the rule-labeled dataset (using the first extraction method described in section 2.1). The trained model was then used to infer prediction results on the manually labeled dataset (using the first extraction method described in section 2.1).

We then trained a second BERT model using the manually labeled dataset (using the first extraction method described in section 2.1) along with the prediction results from the first model, completing the first grafting.

2. Second Grafting

We performed K-Fold cross-validation using the second BERT model, inferring non-leakage prediction results on the manually labeled dataset that is more closely aligned with the test set results.

Using these results, we concatenated them with the manually labeled dataset (using the second extraction method described in section 2.1) and fed them into an LLM, completing the second grafting.

2.4 Training Strategy

The BERT model often exhibits instability during the training process, for which we employed various training techniques. Additionally, exploring how to use generative LLM models for binary classification modeling is also a worthwhile endeavor.

2.4.1 BERT. In the BERT model, to achieve more stable convergence, we adopted the following training techniques:

- `N-Gram Masked LM Pre-training Task[1]`: We pretrained the model using information from 6 million papers in the DBLP dataset. The titles and abstracts of the papers were concatenated and used as unsupervised pretraining corpus.
- `R-Dropout[11]`: We enhanced the model's stability by using a contrastive learning training strategy.
- `TTA(Test-Time Augmentation)[7]`: We randomly shuffled the text each time and averaged the results from multiple inferences.

2.4.2 LLM. Currently, the mainstream approach in the industry for using CausalLM-structured LLM models for classification tasks is to take the hidden layer state of the last token of each sample and then attach a classification head to output the classification result. However, this method often disrupts some of the model's prior knowledge. Therefore, effectively transferring the capabilities of CausalLM-structured LLM models to the classification domain has become a worthwhile endeavor.

In this competition, we adopted a generative CausalLM model for binary classification tasks. During training, after inputting the prompt and effective information, we instructed the model to output a single token, either *"yes"* or *"no"*. This method effectively transforms the classification task into a generative task.

During inference, after a single forward pass of the model to obtain probability values, we extracted the probability values corresponding to the token IDs for *"yes"* or *"no"*. The final prediction probability values were then calculated using these extracted probabilities.

3 RESULT AND DISCUSSION

In this section, we present our main results and ablation studies for some crucial components.

Table 2 represents the ablation experiments for the BERT[1] model.

Table 2: The Ablation Experiments For The BERT Model

Training Strategy	Val Score	Test Score
Manually Annotated Dataset	0.4056	-
Pretrain With DBLP Dataset	0.4432	-
First Grafting	0.4537	-
R-Dropout	0.4696	-
TTA	0.4701	0.4173

Our BERT model uses DeBerta-v3-large[5][4]. The initial MAP score after training on the manually annotated dataset was only **0.4056**. Subsequently, we performed secondary pretraining on the DBLP dataset and then fine-tuned the model, which resulted in a MAP score of **0.4432**. At this point, we utilized a rule-annotated dataset for the first grafting, making full use of the effective information from its 4854 papers, and the MAP score reached **0.4537**. With the addition of R-Dropout and TTA techniques, our score improved to **0.4701**.

Next, we performed the second grafting, grafting the output results of BERT[1] into the LLM. The LLM we selected is ChatGLM3-32K[3][2].

Table 3 represents the ablation experiments for the LLM model after The Second Grafting Learning.

Table 3: The Ablation Experiments For The LLM Model

Training Strategy	Val Score	Test Score
Second Grafting	0.4812	-
Add ref's abstract	0.4917	-
Add ref's cite	0.5025	0.4572
Add ref's feature engineering	0.5314	0.4724
Add paper's feature engineering	0.5392	0.4813

After the second grafting, the score of the LLM reached **0.4812**. Subsequently, with the addition of a series of auxiliary feature engineering based on references, the final single model score reached **0.5392**, and the test dataset score at this point was **0.4813**. At this stage, we constructed three different input methods, and the final multi-model ensemble validation set score was **0.5452**, with a test set score of **0.4879**.

4 CONCLUSION

In this paper, we introduce our pipeline designed for the KDD Cup 2024 OAG Challenge[13] PST Task[12]. Our solution leverages the concept of grafting learning to integrate the complex text semantic matching capabilities of the BERT[1] model into the LLM, thereby enhancing sample confidence. Additionally, our team has developed an automatic feature engineering pipeline based on RAG[6], which alleviates the common issues of excessive text, noisy information, and dirty data in complex semantic texts. Utilizing a 7B single model, our approach exceeded the second-place performance by 1% in the final evaluation metric MAP, ultimately securing the 1st place on the leaderboard.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 320–335.

[3] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, and .. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793

[4] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv:2111.09543 [cs.CL]

[5] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*. https://openreview.net/forum?id=XPZIaotutsD

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[7] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. 2021. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1214–1223.

[8] Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. 2010. A Combination Approach to Web User Profiling. *ACM TKDD* 5, 1 (2010), 1–44.

[9] Jie Tang, Duo Zhang, and Limin Yao. 2007. Social Network Extraction of Academic Researchers. In *ICDM'07*. 292–301.

[10] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: Extraction and Mining of Academic Social Networks. In *KDD'08*. 990–998.

[11] Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems* 34 (2021), 10890–10905.

[12] Fanjin Zhang, Kun Cao, Yukuo Cen, Jifan Yu, Da Yin, and Jie Tang. 2024. PST-Bench: Tracing and Benchmarking the Source of Publications. *arXiv preprint arXiv:2402.16009* (2024).

[13] Fanjin Zhang, Shijie Shi, Yifan Zhu, Bo Chen, Yukuo Cen, Jifan Yu, Yelin Chen, Lulu Wang, Qingfei Zhao, Yuqing Cheng, et al. 2024. OAG-Bench: A Human-Curated Benchmark for Academic Graph Mining. *arXiv preprint arXiv:2402.15810* (2024).