CS 145 Team 3 Project Proposal

Anvesha Dutta
B.S. Data Theory
`dutta.anvesha06@gmail.com`

Sam Hopkins
B.S. Computer Science
`samthehopkin@gmail.com`

Kehan Li
B.S. Financial Actuarial Mathematics
`kehan1230@gmail.com`

Tsunming Liu
B.S. Mathematics of Computation
`tsunmingliu2024@gmail.com`

Adithi Ramesh
B.S. Computer Science
`adithi.ramesh02@gmail.com`

Andy Wang
B.S. Data Theory
`andywang0321@gmail.com`

Abstract

Develop a model to detect incorrect paper assignments among all one's papers.

1 Problem statement

Testing the performance of different Open-Source LLMs on the Incorrect Name-Assignment Detection problem.

2 Literature review

In the ever-expanding academic landscape, accurate attribution of papers to authors is paramount, yet challenges persist due to errors in assignment. This review surveys the literature on detecting and addressing such errors.

Early efforts relied on matching algorithms based on author names, keywords, and content, but faced limitations like name duplication and spelling disparities. Recent advancements in machine learning and natural language processing have led to more effective methods rooted in deep learning and text mining.

Studies have explored techniques to enhance allocation systems and academic databases, introducing innovations such as author matching algorithms based on entity recognition and automatic paper assignment systems.

The WhoIsWho project has notably introduced a large-scale academic name disambiguation benchmark and toolkit to address the challenge of ambiguous author names. They offer a comprehensive set of tasks and competitions to spur method development.

The complexity of name disambiguation is attributed to non-uniform task designs and errors in noisy data. WhoIsWho's benchmarking process and toolkit aim to address these challenges, emphasizing a multimodal approach integrating semantic and relational features.

Community-driven and open-source, WhoIsWho welcomes contributions to advance name disambiguation methods. Future research should focus on more accurate detection methods, analysis of error mechanisms, and optimization of academic databases and allocation systems.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

A proposed machine learning model aims to automate paper attribution by discerning relevant features from limited information such as abstracts, titles, publication dates, and journal names. This model holds potential for academic search engines, journals, publishers, research institutions, and libraries to enhance resource management and utilization.

In summary, advancements in paper assignment error detection offer promising avenues for enhancing the efficiency and accuracy of academic research attribution and resource management.

3 Tentative schedule

- May 13th: Project proposal due
- May 17th: Individual Exploratory data analysis (everybody)
- May 20th: Midterm Exam
- May 21st: Brainstorm session (Requirement: everyone did enough EDA)
- May 24th: All LLMs will be finetuned for the first time by now
- May 30th: The best LLM will be further finetuned by now
- May 31st: Test data released, All participants have 7 days to submit results. We run the test set on our best model.
- June 7th: KDD Competitions End + Final Report typed
- June 10th: Final Presentation
- June 14th: Winners announced

4 Tentative approach

Out of the three baseline methods, we see that fine tuned ChatGLM has the best performance. Based on that observation, we think that LLMs could be a good approach to this problem. Using ChatGLM as a baseline, with that observed we will compare the performances of other popular Large Language Models. We will use Meta LLama3, Mistral AI Mistral, Google Gemma, and Apple OpenELM to solve the task, and see which one does the best.

5 Division of Workload per Member

- Andy Wang (OpenELM fine tuning https://huggingface.co/apple/OpenELM)
- Tsun Ming Liu (LLama3 fine tuning https://huggingface.co/meta-llama/Meta-Llama-3-8B)
- Kehan Li (Further data analysis and help out everyone else)
- Adithi Ramesh (Mistral fine tuning)
- Anvesha Dutta (Gemma fine tuning)
- Sam Hopkins (Visualizations and help out everyone else)

References

[1] Chen, B., Zhang, J., Zhang, F., Han, T., Cheng, Y., Li, X., … & Tang, J. (2023, August). Web-scale academic name disambiguation: the WhoIsWho benchmark, leaderboard, and toolkit. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 3817-3828).

[2] Zhang, F., Shi, S., Zhu, Y., Chen, B., Cen, Y., Yu, J., … & Tang, J. (2024). OAG-Bench: A Human-Curated Benchmark for Academic Graph Mining. arXiv preprint arXiv:2402.15810.

[3] https://arxiv.org/pdf/2402.15810