An Ensemble Model with Multi-Scale Features for Incorrect Assignment Detection

Ming Shen
shenming807@gmail.com
Xiangyuxing Investment
NingBo, China

Abstract

With the number of the publication increasing, the name ambiguity problem is becoming increasingly complex. To improve this research, OGA-Challenge Team published a large-scale dataset and hosted KDD Cup 2024 Challenge for detecting paper assignment errors based on each author and their paper matadata. This paper presents an effective and resource-efficient solution to the aforementioned challenge. Rather than utilising LLM, we have elected to employ an embedding model for the representation of text information. Furthermore, we have implemented multi-scale feature extraction and a graph neural network for the extraction of relationships between papers. Finally, with our solution, our team LoveFishO won 2nd place in task1(WhoIsWho) among 400+ participants.

CCS Concepts

• **Computing methodologies** → *Machine learning*; *Information extraction.*

Keywords

Name Ambiguity, Machine Learning, Text Embedding, Feature Extraction,Tree Model, Graph Neural Network

ACM Reference Format:
Ming Shen. 2024. An Ensemble Model with Multi-Scale Features for Incorrect Assignment Detection. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (KDDCup'24).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The issue of name disambiguation represents a significant challenge for online scholarly systems, particularly in light of the growing volume of published papers. As the number of published papers continues to increase, this challenge is likely to become increasingly complex and demanding to address[1]. While considerable attention has been devoted to name disambiguation, comparatively little attention has been directed towards the study of incorrect assignment detection (IND). Conversely, greater attention is being paid to scratch name disambiguation (SND) and real-time name disambiguation (RND)[5]. In order to stimulate further research in this area, the OAG-Challenge Team has organized the WhoIsWho-IND KDD Cup 2024.

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

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

2 OVERVIEW

As depicted in Figure 1, our solution was composed of three parts: feature extraction, feature combination, and ensemble. In the feature extraction phase, a variety of features were extracted from disparate perspectives. In the feature combination phase, diverse features were integrated as input to the different model. Finally, in the model ensemble phase, multiple models were trained based on different features, and these models were integrated by assigning different weights.

3 PREPROCESS

- Fill null value.
  - Fill null value of the year with 0.
  - Replace None with an empty string.
- Clean the text.
  - Convert uppercase to lowercase.
  - Remove spaces, stop words, and special symbols.

4 FEATURE EXTRACTION

To fully represent the papers, we extract features from multiple dimensions. The first feature dimension is the basic statistical feature of the paper. Secondly, we use pre-trained embedding model to encode the textual information in the paper. These embedding vectors could effectively describe the content and theme of the paper. Finally, we build amount of cross features using basic statistical features and text vector features to describe the relationships between all papers under the same author. Here are few powerful features:

- Basic Statistical Features
  - Keyword Feature: Count of keywords
  - Author Feature: Count of authors
  - Organization Features: Nunique of organization; Nunique of organization divided by total number of organizations; Count of same organizations; Count of same organizations divided by total number of organizations.
- Text Embedding Features
  - Embedding Features: Encode the title, abstract, and venue of the paper using E5-Instruct[1][4] and Voyage[2]
- Author-Paper Features

[1] multilingual-e5-large-instruct
[2] voyage-large-2-instruct

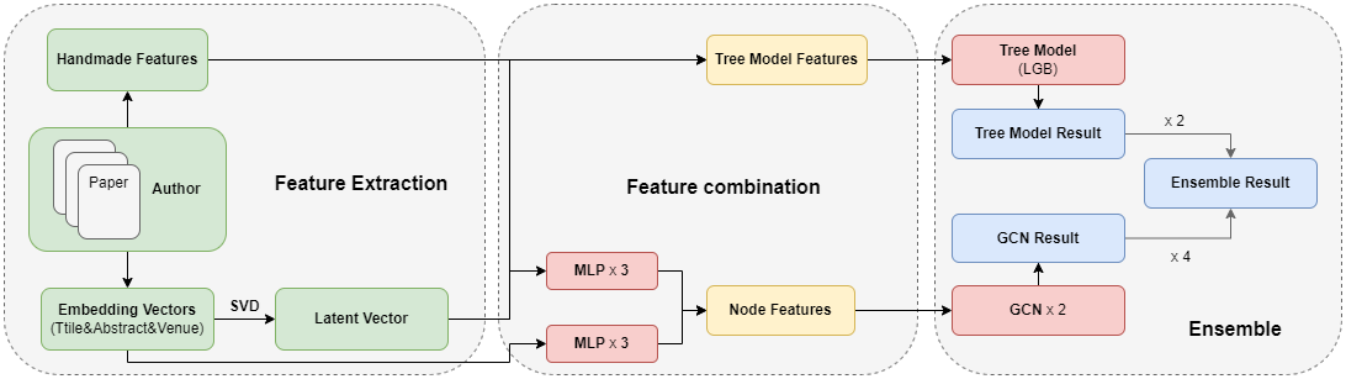

Figure 1: The overview of architecture

- Text Similarity Features: This refers to the similarity between this paper and the author's other papers.

$$CosSim(i, X) = AGG(cosine(v_i, v_j)|j \in X) \quad (1)$$

where AGG is the aggregation function, such as max, mean and std. $cosine$ is the cosine similarity function. v_i and v_j are embedding vectors for title, abstract, and venue.

$$JWSim(i, X) = AGG(sim\_func(s_i, s_j)|j \in X) \quad (2)$$

where AGG is the aggregation function, such as max, mean and std. sim_func is the Jaro–Winkler similarity function. s_i and s_j are the text of title and venue.

$$SumSim(i, X) = JWSim(i, X) + CosSim(i, X) \quad (3)$$

where $JWSim(i, X)$ is the Jaro–Winkler similarity, $CosSim(i, X)$ is cosine similarity.

- Basic Statistical Cross Features:
  - Max count of the same keyword, author, and organization.
  - Max count of the same keyword, author, and organization divided by the count of keywords, authors, and organizations.
  - The absolute gap in publication years between papers with the max count of the same keyword, author, and organization.
  - The gap between the year with the max, median, mean, and min of year
  - The gap between the year with the average of max of year and min of year
  - The gap between the closest and second closest years to the current publication year.
  - Has the author previously published a paper at this venue.
  - Is the publication year within the prescribed range.
  - et al.

5 TREE MODEL

The tree model[3] is employed for the purpose of identifying misclassified papers, with the extracted features mentioned above serving as the basis for this identification. Nevertheless, the dimension of the text embedding vectors is excessive and greatly exceeds the number of other features. This results in the model being unable to effectively capture the other features, leading to sub-optimal modelling outcomes. To address this issue, it is necessary to downscale the text vector features. The SVD[4] method is selected for this purpose as it is capable of identifying nonlinear relationships and latent space information, which are not as readily discernible through PCA[5]. A trial has demonstrated that the optimal reduction of the text embedding vector is to 32 dimensions. The data analysis revealed that the data categories were not balanced. To mitigate the model overfitting to a single category, a 10-fold StratifiedKFold cross-validation was employed to train the tree model.

6 GRAPH NEURAL NETWORK

In the context of graph neural networks, it is necessary to construct the graph in accordance with the specific task, and to define the nodes and edges within the graph. In this study, each author is defined as a graph, with papers representing nodes and connections between papers by the same author other than the original author represented as edges. The node features are comprised of two distinct parts. The first part encompasses the features utilized in the aforementioned tree model, while The second part incorporates the text embedding vector features that have not undergone downscaling. The Edge features are comprised of two key elements: the percentage of identical author names and keywords, and the similarity of organisational and conference names.Although edge features are incorporated into the graph convolution process, they are not utilized directly. Instead, they are employed to filter out some of the less significant connections. For this task, it was decided that the GCN[3] model would be used as the base model.

7 ENSEMBLE

The train data were not employed for the purpose of fine-tuning the embedding models. In order to guarantee the relative reliability of the embedding models within the context of the thesis, two embedding models were therefore utilized for the purpose of representing the paper information. Concurrently, in order to ascertain

[3]Lightgbm[2]

[4]Singular Value Decomposition

[5]Principal Components Analysis

Table 1: Model Weight

Model	Weight
LGB-Voyage	0.385
LGB-E5-Instruct	0.315
GCN-E5-Instruct	0.075
GCN-Voyage	0.075
GCN-E5-Instruct-Voyage	0.075
GCN-Voyage-E5-Instruct	0.075

Table 2: Overall Performance

Model	Score
LGB-Voyage	0.81433
LGB-E5-Instruct	0.81827
GCN-E5-Instruct	0.78082
LGB(E5-Instruct/Voyage)x2+GCN(E5-Instruct/Voyage)x4	0.82486

Table 3: Top 10 score in task. Our team "LoveFishO" won 2nd in this task of KDD Cup 2024

Rank	Team Name	Score
1	BlackPearl	0.83454
2	**LoveFishO**	**0.82487**
3	AGreat	0.81349
4	Kozuki Cats	0.80890
5	M1stic	0.80720
6	DOCOMOLABS	0.80487
7	qianlan	0.80137
8	LGB YYDS	0.79941
9	DeepMayNotLearn	0.79774
10	Leo_Lu	0.79738

the interrelationships between the papers from disparate perspectives, we employed not only the tree model but also the graph neural network model. In total, six models have been integrated, designated LGB-E5-Instruct, LGB-Voyage, GCN-E5-Instruct, GCN-Voyage, GCN-E5-Instruct-Voyage, and GCN-Voyage-E5-Instruct. All models are integrated with different weights1 assigned to each.

8 EXPERIMENT

The performance of our models is listed in Table 2. The results of the analysis indicate that the tree model outperforms the graph neural network. Additionally, the E5-Instruct model demonstrates greater efficacy than the Voyage model in characterising papers. Although the single graph neural network is not particularly effective, its integration with the tree model is beneficial.

9 CONCLUSION

This paper presents our solution to the WhoIsWho task for the KDD Cup 2024. Our approach utilizes tree model and graph neural network to detect paper assignment errors. Instead of using dataset to train a traditional language model to vectorise the text, we use a pre-trained embedding model based on a large corpus to vectorise the text. This approach not only capitalises on the capabilities of large models but also reduces the consumption of resources. It was also a key factor which won the 2nd place in Task of KDD Cup 2024 Challenge.

References

[1] Bo Chen, Jing Zhang, Fanjin Zhang, Tianyi Han, Yuqing Cheng, Xiaoyan Li, Yuxiao Dong, and Jie Tang. 2023. 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* (Long Beach, CA, USA) *(KDD '23)*. Association for Computing Machinery, New York, NY, USA, 3817–3828. https://doi.org/10.1145/3580305.3599930

[2] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[3] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. arXiv:1609.02907 [cs.LG] https://arxiv.org/abs/1609.02907

[4] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. arXiv:2402.05672 [cs.CL] https://arxiv.org/abs/2402.05672

[5] 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).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009