

# Academic Institution Name Recognition Based on Representation Learning and Semantic Matching

Jinyu Wang and Zhijie Ban

College of Computer Science, Inner Mongolia University, Hohhot, 010021, China  
banzhijie@imu.edu.cn

**Abstract.** Recognition of scholar’s institution name has been extensively researched for accurately parsing academic papers. Existing rule-based methods are primarily applicable in the cases where the writing styles of the organization name are regular. Other approaches need to pre-establish a knowledge base for mapping organization names, which demands considerable human resources. This paper presents a method based on representation learning and semantic matching, primarily leveraging institution’s textual information and academic network’s structure. We first construct an author-institution heterogeneous graph, on which maximal random walk and Word2Vec are used to obtain representation vectors for institution nodes. Then, we convert institution names into semantic vectors by the SimCSE model and institution candidate sets are generated by employing the locality sensitive hashing algorithm. Finally, in order to avoid setting the cluster number, we propose a connected subgraph partitioning method to divide institution clusters. Experimental results on two real datasets demonstrate that our method significantly outperforms the existing state-of-the-art recognition methods.

**Keywords:** Academic Institution Name Recognition, Representation Learning, Semantic Matching.

## 1 Introduction

Assessing scientific institutions' comprehensive strength is crucial for expert recommendations and resource integration [1]. Accurate paper, author, and institution matching is vital for evaluating outputs and citations. With growing researchers and papers [2], diverse institution names pose challenges. Correctly identifying and distinguishing institution names [3] is essential for accurate data utilization. Institution name recognition aims to identify real-world entities in academic data [4], enabling applications in name disambiguation [5], career trajectories [6], talent mobility [7], paper management [8], collaboration networks [9], and research performance assessments [10].

Researchers have studied institutional name recognition through various methods. Early approaches relied on manually curated lists or manual identification of affiliations. Rule-based methods [1] improved accuracy but struggled with irregular names. Deep learning techniques, especially entity linking [3], utilizing knowledge graphs like

XLore [11], enhance performance but depend on graph richness. Today, the complexity of data poses significant challenges for institutional name recognition.

Our method captures institutional relationships and name attributes, enabling comprehensive and accurate modeling. We construct a heterogeneous network, learn institutional structure via representation learning, capture semantic name info, and propose a subgraph-based division to avoid pre-set clusters. Key contributions: network modeling, representation learning, and subgraph division.

- We construct an author-institution heterogeneous graph using information such as papers, authors and affiliations. The structural representation vectors of the institution nodes are generated by the maximal random walk strategy and the neural network model. This enhances the provision of comprehensive information, facilitating a more precise comprehension of inter institutional relationships.
- Semantic vectors for the institution names are obtained by the SimCSE semantic model. Then, we utilize the locality sensitive hashing algorithm to calculate the semantic matching degree between institution names.
- A method for partitioning institution clusters is proposed, where edges are established between institutions when their similarity is larger than a predefined threshold. Connected subgraphs among institutions are constructed and each connected subgraph represents an institution cluster.

## 2 Literature Review

Currently, the research methods of academic institution name recognition mainly include traditional machine learning-based method, knowledge-based strategy and entity linking-based approach.

At the early stage, the researchers used unsupervised or semi-supervised methods for institution name identification due to the lack of publicly available validation sets. Yang et al. [12] proposed an organization name mapping algorithm to study the one-to-many mapping rules between organization names by the statistical analysis technology. Wang et al. [13] proposed an improved method aided institutional name normalization and attribute enrichment. In summary, these methods mainly target the variant forms of institutional names with regularity and have some enhancement in solving the ambiguity of institutional name identification. However, due to the problems of large data volume and diverse writing forms, the proposed matching rules are difficult for all data types and have poor robustness.

The method based on knowledge base usually creates a mapping knowledge base about institution names. Huang et al. [14] proposed a new identification tool based on the existing institution name disambiguation technique by introducing institution persistence identifiers. But the construction of a knowledge base requires regular manual updates and high maintenance cost due to the existence of name changes, the variations and mergers of institutions.

Third-type approaches leverage deep learning. Shao et al. [4] introduced an automatic framework for institutional name disambiguation, filtering contextual info with XLore, calculating string matching probabilities, and positioning scholars' affiliations

based on achievements and geographic info [15]. However, they overlook potential structural [16] changes among institutions.

### 3 Our Method

#### 3.1 Problem Definition

Given a collection of papers  $P=\{p_1, p_2, \dots, p_n\}$ , each paper contains specific information such as paper id, title, authors, affiliations, and keywords. In the paper, the attribute information is associated with a specific value, where each author's affiliation corresponds to a real-life institutional entity, and the remaining attributes correspond to a phrase [17].

**Definition 1. Scholar Institution Name Recognition.** The goal of scholar institution name recognition is to find a function  $\Phi$  that partitions the set of scholar institution names  $D$  into a set of disjoint clusters  $C$  by incorporating the institution names from the conference proceedings  $P$ . Therefore, each cluster represents a unique institutional entity in the real world, i.e.,  $\varphi(D) \rightarrow C$ , where  $C = \{C_1, C_2, \dots, C_k\}$  and  $C_i \cap C_j = \emptyset (i \neq j)$ .

### 4 Overview of Research Methodology

The whole framework model of our institutional name recognition is shown in

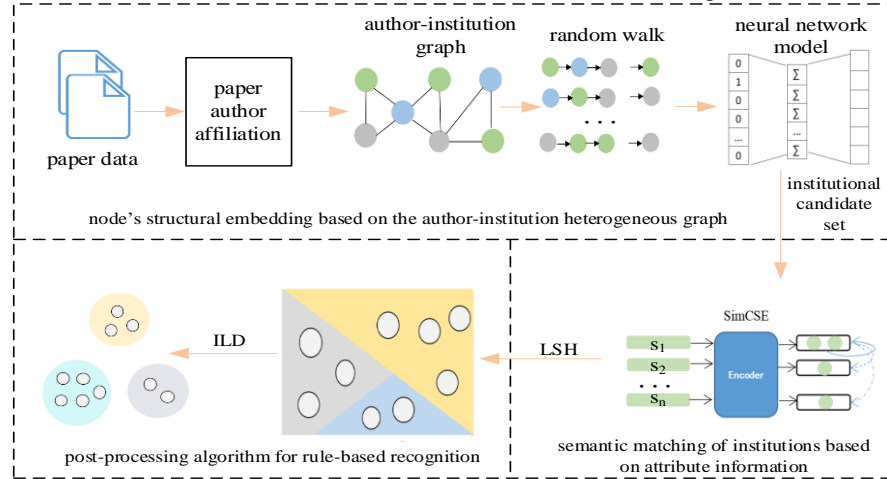


Fig. 1, which mainly consists of three parts.

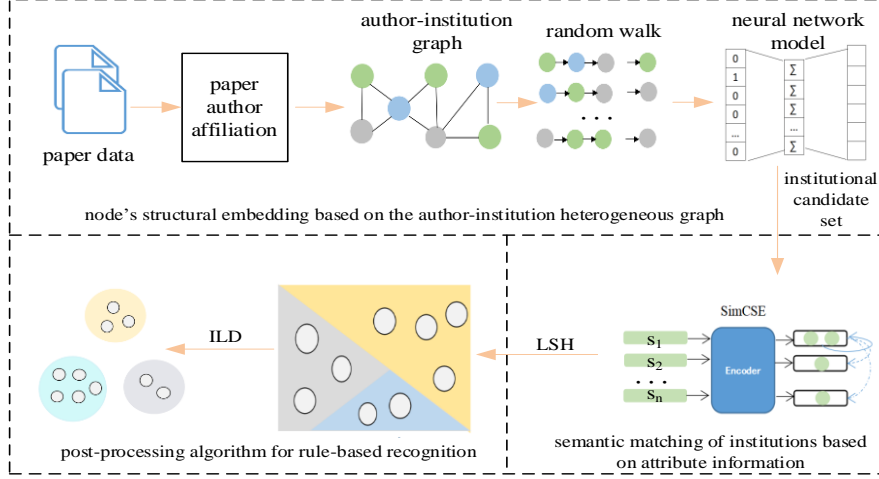


Fig. 1. The whole framework

The proposed approach consists of three main parts. Firstly, node structure embedding is performed based on an author-institution heterogeneous graph. This involves preprocessing affiliation relationships in papers, extracting relevant information, and leveraging member attributes to connect distinct institutions. Random walks are then conducted to generate sequences, which are fed into a Word2Vec model to obtain representation vectors for institutions. Secondly, semantic matching is achieved using the institution's text attributes. The candidate set of institutions from the first part is input into the SimCSE[18] model to generate semantic vectors. Locality Sensitive Hashing is employed to generate candidate sets of institutions that are similar in both structure and semantics. Thirdly, a rule-based post-processing algorithm calculates the string similarity between institution names. If the similarity exceeds a threshold, an edge is established, forming multiple connected subgraphs, each representing an institution cluster.

#### 4.1 Node's Structural Embedding Based on the Author-Institution Heterogeneous Graph

The relationships between institutions are crucial for mapping institution name recognition. By exploring the connections between institutions, institution names can be more accurately identified and classified.

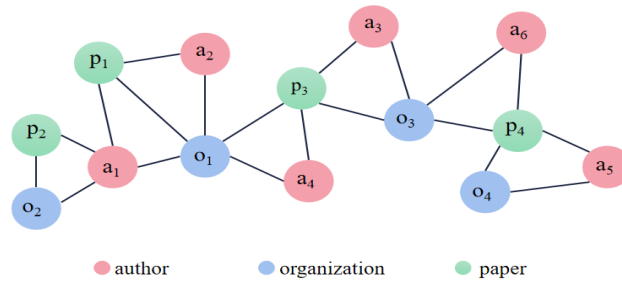
**Construction of the Author-Institution Heterogeneous Graph.** The appearance of diverse institution names within the public dataset can usually be attributed to four main factors.

- The same author may use different forms of institution names in their published papers.
- A paper is completed by a research team of an academic institution and the members may write their institution name in the different forms.

- An academic institution  $A$  collaborates with another one  $B$ . Because the  $B$ 's cooperation members may change, there will be many different name forms of the institution  $B$  in the different academic papers. This situation is generally common in the scientific research organization.
- In the process of continuous development and growth of an institution, there may also be the situations such as renaming and changes.

Based on the relevant information included in the academic literatures, we extract every paper's id, author names, and affiliated institution names. This forms the heterogeneous graph  $G = \{V, \delta\}$ , composed of the node set  $V$  and the edge set  $\delta$ . The node types are respectively paper, author and institution. Edges include the relationship between the paper and the author, the author and the institution, and the paper and the institution.

We establish the edges between the elements within each paper through three connection ways such as paper-author, author-organization, and paper-organization, which forms a subgraph of the paper. If two papers have the same author, this means that the subgraphs of two papers are connected, thereby establishing indirect connections between the associated institutions. Similarly, if two papers share the same institution, the subgraphs of these two papers are connected through their respective institution nodes, and other connected institutions within the subgraphs are also linked. By progressively establishing connections, a network is constructed among all the institutions, as shown in Fig. 2. The paper nodes are represented as  $p_1, p_2, p_3, \dots, p_k$ , author names as  $a_1, a_2, a_3, \dots, a_k$ , and affiliations as  $o_1, o_2, o_3, \dots, o_k$ . By constructing the heterogeneous graph of the author-institution relationships, we can effectively address the issue of institutional name ambiguity.



**Fig. 2.** Author-institution heterogeneous graph

**Node's Representation Learning Based on the Author-Institution Heterogeneous Graph.** The maximal random walk approach learns institutional associations by traversing the graph randomly. Nodes are described by wandering sequences, capturing local and higher-order neighbors. The heterogeneous graph of author institutions handles homonyms, revealing institutional correlations through random walks.

Graph representation learning transforms institutions into vectors, overcoming one-hot encoding's semantic limitations. Word2vec's Skip-Gram model maps words to a

low-dimensional space, clustering similar words. Here, it captures institutions' local structure, exploring highly connected regions via random walks, generating sequences for training and structured institution representations.

As nodes on a graph share more common neighbors, their contextual structures become more similar. This proximity in the graph's actual space translates to a shorter distance between their feature space vectors, indicating a higher likelihood of representing the same institution. Cosine similarity is often used to measure these vector distances, and institutions with the top-k highest similarities are selected to form a set of candidate institutions.

#### 4.2 Semantic Matching of Institutions Based on the Attribute Information

For the identification of organization name, the semantic information of institution name itself is also a good identification factor. Although the same organization will have different writing forms, the probability of the organization names with similar semantics being the same institution will be greater. We utilize a semantic generation model called SimCSE to map institution names to a vector space and train it accordingly. Additionally, we construct a matching query library. Then, the Locality Sensitive Hashing (LSH) algorithm is used for fast matching. As a result, we obtain a candidate set of institutions that exhibits both structural and semantic similarities. Using this approach, we can more accurately filter out similar institutions and improve the matching results.

**SimCSE Model.** Traditional methods sum word vectors for sentences, ignoring word interactions. The BERT model, though using Transformers to capture bidirectional context, has limitations in semantic similarity. The SimCSE model, via contrastive learning, generates superior sentence vectors by clustering similar instances and separating dissimilars. We adopt SimCSE for institution representations, utilizing unsupervised training due to lack of validation data.

The SimCSE model obtains a more discriminative vector representation by inputting the same sentence twice into the encoder to obtain different representation vectors as positive examples, while using the representation vectors of other sentences as negative examples.

**Locality Sensitive Hashing Algorithm.** After obtaining the semantic vectors of institutions, we need to utilize their semantic information for filtering. The higher the semantic similarity, the greater the probability of the same institution is. The time complexity of linear search is very high when semantic matching is carried out in massive high-dimensional data. To address this, we employ Locality Sensitive Hashing (LSH) algorithm. Firstly, the data is formed into a matrix form, and a signature matrix is generated through a Hash function. Then, LSH algorithm deals with the signature matrix and maps the data to different buckets. This ensures that similar data are mapped to similar hash values, while dissimilar data are mapped to different hash values. This algorithm achieves dimensionality reduction and local matching, significantly reducing

the time complexity for querying similar data and avoiding pair-wise comparisons of all data points, thus improving matching efficiency.

### 4.3 Post-Processing Algorithm for Rule-Based Recognition

By the above two sections, we have obtained a set of possible institutions that may belong to the same organization. The ultimate goal of institution name recognition task is to partition scholar institution names into multiple disjoint clusters. However, a key challenge of the clustering algorithms is determining the cluster number. To overcome the difficulty, we propose an idea of partitioning institution clusters using connected subgraphs.

To better distinguish different organization strings and improve the accuracy of matching the same institutions, we adopt an improved string edit distance algorithm (IDL) [19] for calculating the string similarity of organizations.

Given two organization strings  $X$  and  $Y$ , we tokenize them into words, resulting as  $X = \{x_1, x_2, \dots, x_p\}$  and  $Y = \{y_1, y_2, \dots, y_q\}$ . Then, a pairwise institution word matching matrix  $E$  with  $p$  rows and  $q$  columns is constructed, where its element is  $e_{ij} = sim(x_i, y_j)$ . The specific calculations are as follows.

- To calculate the similarity between two words in the organization, let  $d(x_i, y_j)$  representing the string edit distance between  $x_i$  and  $y_j$ . If  $x_i$  is a substring of  $y_j$ , or  $y_j$  is a substring of  $x_i$ , then  $d(x_i, y_j) = 0$ . The similarity between the two words  $x_i$  and  $y_j$  is

$$sim(x_i, y_j) = 1 - \frac{d(x_i, y_j)}{\max(len(x_i), len(y_j))} \quad (1)$$

where  $max$  is the maximum value function and  $len$  is the string length function.

- If  $sim(x_i, y_j)$  equals to 1, the string  $x_i$  and  $y_j$  exactly match. For the matrix  $E$ , if existing at least one  $sim(x_i, y_j) = 1$  on the NO.  $p$  row or No.  $q$  column, we think that the affiliation  $X$  and the affiliation  $Y$  have one word exactly matched. It is expressed as follows.

$$CMX(i) = \begin{cases} 1, & \exists sim(x_i, y_j) = 1 \text{ for } 1 \leq j \leq q \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$CMY(j) = \begin{cases} 1, & \exists sim(x_i, y_j) = 1 \text{ for } 1 \leq i \leq p \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

The similarity of exact matching words in organizations  $X$  and  $Y$  is as follows.

$$sim(X, Y)_{cm} = \frac{\text{average}(\sum_p CMX(p), \sum_q CMY(q))}{\text{average}(p, q)} \quad (4)$$

where  $average$  is the average function.

- The similarity between non-matching words in institution  $X$  and  $Y$  is as follows.

$$sim(X, Y)_{other} = \frac{\text{average}(\sum_p \max(E_{pq}), \sum_q \max(E_{pq}))}{\text{average}(p, q)} - \frac{\text{average}(\sum_p CMX(p), \sum_q CMY(q))}{\text{average}(p, q)} \quad (5)$$

where  $\sum_p \max(E_{pq})$  denotes the summation of the maximum values in each row of the institution word matching matrix E. Similarly,  $\sum_q \max(E_{pq})$  denotes the summation of the maximum values in each column of the institution word matching matrix E.

- The similarity between institution  $X$  and  $Y$  is as follows.

$$\text{sim}(X, Y) = \text{sim}(X, Y)_{cm} \times W_1 + \text{sim}(X, Y)_{other} \times W_2 \quad (6)$$

where  $W_1$  and  $W_2$  are weight parameters.

The proposed method assesses institution name string similarity to partition institutions into non-overlapping clusters. Our post-processing algorithm: (1) Calculates a string similarity matrix, (2) Establishes connections between organizations exceeding a similarity threshold, and (3) Divides the graph into connected subgraphs, each representing a distinct institution cluster. Each institution belongs to one cluster exclusively.

Our method performs well in solving the problem of difficulty in determining the number of clusters in the process of institutional clustering, fully utilizing the similarity information between institutions in the dataset, thus significantly improving the accuracy and interpretability of the clustering results.

## 5 Experimental Result and Analysis

### 5.1 Datasets

To evaluate the effectiveness of our proposed method, we used the DBLP-V12 dataset, which is publicly available on the Aminer website (<https://www.aminer.cn/>). We selected two subsets from this dataset representing different subfields of computer science, which are Information Security Dataset abbreviated as ISD and Natural Language Processing abbreviated as NLPD. We specifically chose papers published between 1840 and 2020. These subsets contain a significant number of instances where institution names exhibit synonymy or variation, making the task more challenging. After extracting and reduplicating the information from the datasets, Information Security subset contains 11,612 unique institution records and Natural Language Processing subset contains 60,750 unique institution records.

Due to the lack of official data sets for the verification of organization name identification, the existing studies mainly use two methods to verify the results. One approach uses manual group inspection and the other is to annotate the data before the experiments. We adopt an annotation method. For instance, an institution A may have multiple variations in its written form, such as A1, A2, ..., Am. By annotating the institution data, we can provide strong support for the subsequent validation of the experiments.

### 5.2 Baseline Methods

To validate our method, we compare it with two institution name recognition approaches: Huang's rule-based method[10], influenced by rule and knowledge quantity, and Shao et al.'s PAAS[15] algorithm, which uses CRF++ and fuzzy matching on Aminer data to identify institutions.



### 5.3 Experimental Results

**Parameters Sensitivity Analysis of Semantic Matching Mechanism.** For semantic matching of institutions, we set a semantic similarity threshold  $\alpha$  and a ranking parameter rankK=10 to cover most name variations. After five experiments with  $\alpha$  values from 0.65 to 0.85,  $\alpha=0.75$  achieved optimal Precision, Recall, and F1-Score on two datasets, as shown in Fig. 3.

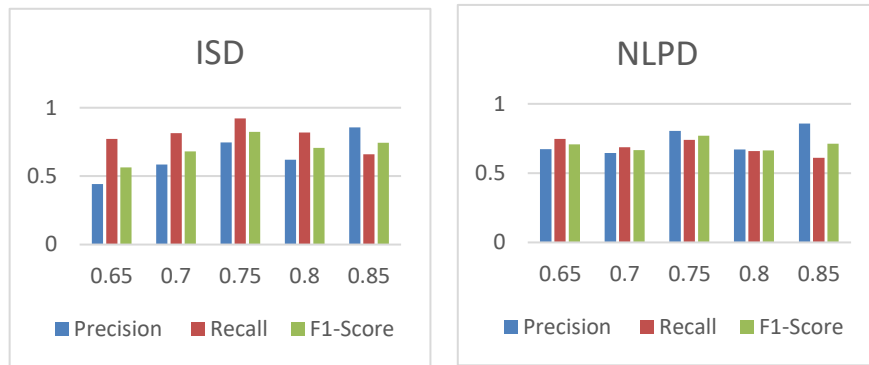


Fig. 3. The analysis of parameter  $\alpha$

**Parameters Sensitivity Analysis of Post-Processing Algorithm Based on Rule.** For the rule-based post-processing algorithm, we selected a threshold  $\beta$  for institution name string similarity. After experiments with thresholds from 0.65 to 0.85,  $\beta=0.75$  or 0.8 achieved the best institution partitioning performance on both datasets, as shown in Fig. 4.

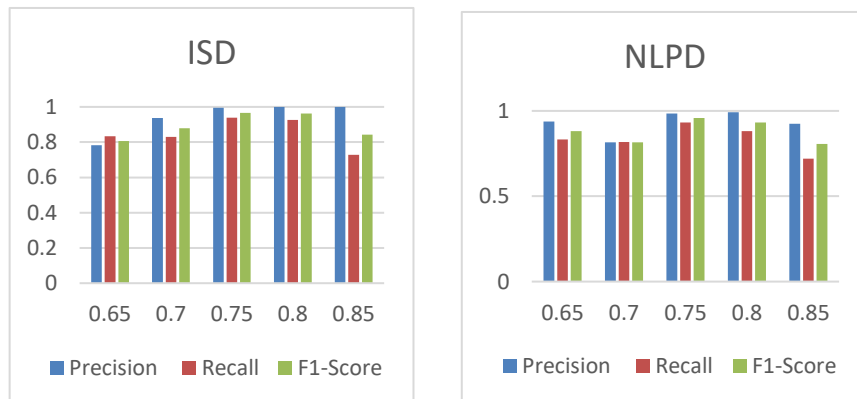


Fig. 4. The analysis of parameter  $\beta$

**Sensitivity Analysis of System Framework Parameters.** The statistical framework has three parts with optimized parameter combinations. Using Top1000 and Top2000 institutional data, the structural embedding was evaluated. For semantic matching, we tested  $\alpha=0.75$  and rank10 ,  $\alpha=0.7$  and rank20. For the rule-based post-processing,  $\beta=0.75$  and  $\beta=0.8$  were compared.

As shown in Table 1 and Table 2, the experimental results are better with the first set of parameters. The subsequent comparison experiments use the parameter set.

**Table 1.** Experimental results of combination parameters on ISD.

Parameter Setting	Precision	Recall	F1
Top1000; $\alpha(0.75)$ -rank10; $\beta(0.75)$	0.9962	0.9386	0.9665
Top1000; $\alpha(0.75)$ -rank10; $\beta(0.8)$	1	0.9273	0.9622
Top1000; $\alpha(0.7)$ -rank20; $\beta(0.75)$	0.9899	0.9268	0.9573
Top1000; $\alpha(0.7)$ -rank20; $\beta(0.8)$	0.9994	0.8985	0.9462
Top2000; $\alpha(0.75)$ -rank10; $\beta(0.8)$	1	0.9101	0.9529
Top2000; $\alpha(0.7)$ -rank20; $\beta(0.8)$	1	0.9138	0.9549

**Table 2.** Experimental results of combination parameters on NLPD.

Parameter Setting	Precision	Recall	F1
Top1000; $\alpha(0.75)$ -rank 10; $\beta(0.75)$	0.9851	0.9325	0.9581
Top1000; $\alpha(0.75)$ -rank 10; $\beta(0.8)$	0.9932	0.8814	0.9339
Top1000; $\alpha(0.7)$ -rank 20; $\beta(0.75)$	0.9446	0.8442	0.8915
Top1000; $\alpha(0.7)$ -rank 20; $\beta(0.8)$	0.9923	0.8728	0.9357
Top2000; $\alpha(0.75)$ -rank 10; $\beta(0.8)$	0.9831	0.8822	0.9299
Top2000; $\alpha(0.7)$ -rank 20; $\beta(0.8)$	0.9717	0.8636	0.9145

**Comparative Experiments.** We compare our method to Huang's, which uses an author institution table and matching rules, and PAAS, which extracts paper data with CRF++ and regex, for evaluating effectiveness on two datasets.

Table 3 and Table 4 show the experimental results. Compared with two Huang's method and PAAS, our method achieves the best performance. This is because we introduce deep learning ideas. We adopt a more comprehensive and accurate feature representation, which effectively compensates for the shortcoming of measuring only from the string perspective. This is also attributed to the use of multiple information in the organization name matching, which improves the accuracy. Our method enables more comprehensive coverage of various organization name variants.

**Table 3.** Experimental Comparisons on ISD.

Algorithm	Precision	Recall	F1
-----------	-----------	--------	----

Huang's Method	0.6357	0.7139	0.6725
PAAS	0.9430	0.6387	0.7616
Our Method	0.9962	0.9386	0.9665

**Table 4.** Experimental Comparisons on NLPD.

Algorithm	Precision	Recall	F1
Huang's Method	0.3227	0.5831	0.4152
PAAS	0.9548	0.5819	0.7231
Our Method	0.9851	0.9325	0.9581

**Ablation Experiments.** To validate our framework, we conducted ablation experiments on the graph-based structure embedding and semantic matching components. Based on the experimental results presented in Table 5 and Table 6, our comprehensive approach, incorporating both the author-institution heterogeneous graph and semantic name information, achieves superior results in precision, recall, and F1 scores, surpassing the performance of individual components on both datasets.

**Table 5.** Ablation experiments on ISD.

Algorithm	Precision	Recall	F1
Graph	0.4104	0.7765	0.5369
Sentence	0.6497	0.7294	0.6872
Our Method	0.9962	0.9386	0.9665

**Table 6.** Ablation experiments on NLPD.

Algorithm	Precision	Recall	F1
Graph	0.5232	0.7584	0.6192
Sentence	0.6381	0.7517	0.6902
Our Method	0.9851	0.9325	0.9581

## 6 Conclusions and Future Work

This paper addresses the issue of ambiguity in recognizing institution names in the academic domain and proposes a method for institution name recognition based on representation learning and semantic matching. The results demonstrate that our method outperforms two basic methods.

In the future, we plan to develop a scholar institution recognition online system which can provide Instant recognition for the researches of institution management and assess.

## ACKNOWLEDGEMENTS

This work was supported by Natural Science Foundation of Inner Mongolia Autonomous Region(2021MS06015). Zhijie Ban is corresponding author.

## References

1. Huang S, Yang B, Yan S, Rousseau R. Institution name disambiguation for research assessment. *Scientometrics*. 2014; 99: 823-838. doi: 10.1007/s11192-013-1214-2.
2. Zhu H, Jiang S, Chen H, Roco M. International perspective on nanotechnology papers, patents, and NSF awards (2000–2016). *Journal of Nanoparticle Research*. 2017; 19: 1-11. doi: 10.1007/s11051-017-4056-7.
3. Basile A, Crupi R, Grasso M. Disambiguation of company names via deep recurrent Networks. arXiv preprint arXiv: 2303.05391, 2023.
4. Shao Z, Cao X, Yuan S, Wang Y. ELAD: An entity linking based affiliation disambiguation framework. *IEEE Access*. 2020; 45: 176-186. doi:10.1109/access.2020.2986826.
5. Zhang Y, Zhang F, Yao P, Tang J. Name disambiguation in AMiner: Clustering, maintenance, and human in the Loop. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2018, p. 1002-1011.
6. Kan W, Tang J, Shao Z, Xu X, Gao B, Zhao S.: Career map: visualizing career trajectory. *Science China Information Sciences*. 2018; 47: 250-263. doi: 10.1007/s11432-018-9469-5.
7. Moed H F, Halevi G. A bibliometric approach to tracking international scientific migration. *Scientometrics*. 2014; 101: 1987-2001. doi:10.1007/s11192-014-1307-6.
8. Jiang Y, Zheng H T, Wang X, Lu B, Wu K. Affiliation disambiguation for constructing semantic digital libraries. *Journal of the American Society for Information Science and Technology*. 2011; 62: 1029-1041. doi:10.1002/asi.21538.
9. Xu H. A regional university-industry cooperation research based on patent data analysis. *Asian Social Science*. 2010; 6: 88-94. doi:10.5539/ass.v6n11p88.
10. Sabah F, Hassan S U, Muazzam A, Lqbal S, Soroya H S, Sarwar R. Scientific collaboration networks in Pakistan and their impact on institutional research performance: A case study based on Scopus publications. *Library Hi Tech*. 2018; 37: 19-29. doi: 10.1108/LHT-03-2018-0036.
11. Wang Z, Li J, Wang Z, Li S, Li M, Zhang D, Shi Y, Liu Y, Zhang P, Tang J. XLOre: A large-scale english-chinese bilingual knowledge graph. *Proceedings of the 12th International Semantic Web Conference*. 2013, p. 121-124..
12. Yang B, Yang J W, Yang S L. Research on Rule-based Normalization of Institution Name. *New Technology of Library and Information Service*. 2015; 31: 57-63. doi:10.11925/INFOTECH.1003-3513.2015.06.09.
13. Wang L, Hu J, Wang Q, Yang Y S, Pei L, An F. Big open data aided institutions' name normalization and attribute enrichment. *Proceedings of 2022 3rd Information Communication Technologies Conference (ICTC)*. 2022. p. 173-177.
14. Huang Y, Li J, Sun T, Xian G J. Institution information specification and correlation based on institutional PIDs and IND tool. *Scientometrics*. 2020; 122: 381-396. doi: 10.1007/s11192-019-03268-9.
15. Shao Z, Yuan S, Xu J, Wang Y I. A statistical feature data mining framework for constructing scholars' career trajectories in academic data. *Applied Soft Computing*. 2022; 118: 1-11. doi: 10.1016/j.asoc.2022.108550.

16. Ratcliff J W, Metzener D E. Pattern matching: the gestalt approach. *Dr Dobbs Journal*. 1979; 43: 313-327.
17. Li Z Z, Zhi J B. Author Name Disambiguation Based on Rule and Graph Model. *Proceedings the 9th CCF International Conference on Natural Language Processing and Chinese Computing*. 2022. p. 617-628.
18. Gao T, Yao X, Chen D. Simcse: Simple contrastive learning of sentence embeddings. *Proceedings the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021. p. 6894-6910.
19. Zhang S, Huang T, Xin H E, Fan Y. ANDMC: An algorithm for author name disambiguation based on molecular cross clustering. *Journal of the Association for Information Science and Technology*. 2019; 70: 42-58. doi:10.1007/978-3-030-18590-9-12.