An end-to-end entity recognition and disambiguation framework for identifying Author Affiliation from literature publications

Lianghong Lin, Wenxiu Xie, Zili Chen, Tianyong Hao
Lianghong Lin, School of Artificial Intelligence, South China Normal University, Guangzhou, China, linlianghong@m.scnu.edu.cn
Wenxiu Xie, Department of Computer Science, City University of Hong Kong, Hong Kong, China, wenxixie-c@my.cityu.edu.hk
Zili Chen, School of Professional Education and Executive Development, CPCE, The Hong Kong Polytechnic University, Hong Kong, China, spczili@speed-polyu.edu.hk
Tianyong Hao, School of Computer Science, South China Normal University, Guangzhou, China, haoty@m.scnu.edu.cn

Abstract

Author affiliation information plays a key role in bibliometric analyses and is essential for evaluating studies. However, as author affiliation information has not been standardized, which leads to difficulties such as synonym ambiguity and incomplete data during automated processing. To address the challenge, this paper proposes an end-to-end entity recognition and disambiguation framework for identifying author affiliation from literature publications. For entity disambiguation, an algorithm combining word embedding and spatial embedding is presented considering that author affiliation texts often contain rich geographic information. The disambiguation algorithm utilizes the semantic information and geographic information, which effectively enhances entity recognition and disambiguation effect. In addition, the proposed framework facilitates the effective utilization of the extensive literature in the PubMed database for comprehensive bibliometric analysis. The experimental results verify the robustness and effectiveness of the algorithm.

1 Introduction

Bibliometrics is becoming increasingly important in academia, revealing disciplinary trends and assessing the impact of research by analyzing literature data, and has become an important tool for promoting academic communication and research decision-making (Ninkov et al., 2022). In particular, assessing the contribution of research elements such as journals, countries/regions, institutions, and authors to a given field plays a key role in bibliometric analyses, helping to reveal the contribution and impact of scholarly research and guiding academic decision-making and resource allocation (Lim and Kumar, 2023). Such assessments provide an important basis for trends in disciplinary development, opportunities for collaboration and science policy development, and promote progress and innovation in academic research (Donthu et al., 2021).

The PubMed database plays a crucial role in providing rich, reliable, timely and standardized literature data for bibliometric analysis in the field of medicine in order to support quantitative analysis of quantitative features and patterns of medical research (Fiorini et al., 2018). It not only helps to improve the quality and efficiency of bibliometric analysis, but also promotes knowledge innovation and dissemination in the medical field (Lu, 2011; Kokol et al., 2021; Thompson et al., 2015). For datasets exported from PubMed databases, statistics on country/region or institution data often rely on author affiliation information existing in every publication. Author affiliation information in PubMed typically includes the name of the university, research institution, hospital, or company where the authors are affiliated, as well as possible department or laboratory names to indicate the institutional affiliation or organizational information of the authors. Additionally, author affiliation data includes information about the address of the institution, such as city, state or province, and country. There are problems of incomplete data and synonym ambiguity in the author affiliation

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 120–129

information. Specifically, the processing of textual affiliation data often encounter incomplete information (e.g., missing country)(Shao et al., 2020), inconsistent representation forms (e.g., University of Cambridge and Cambridge University)(Rimmert et al., 2017), and abbreviations (e.g., UCLA)(Huang et al., 2014), which may affect the results of subsequent analysis that assessed the contribution of the research elements to the field of study.

Author affiliation information disambiguation is essential because it concerns the use of entities such as institution, country, etc. in bibliometric analysis. The accuracy of bibliometric indicators provided by databases themselves is limited. Relying solely on these data for bibliometric analysis can lead to significant inaccuracies in the indicator values. As suggested by Donner et al. (2020), additional data cleaning is necessary to disambiguate affiliation data. While there are numerous studies on entity disambiguation, relatively few have explored author affiliation information, especially for PubMed databases. Moreover, existing research on author affiliation information is mainly confined to institutions mainly confined to institutional entities, ignoring the identification and disambiguation of Location entities (L'Hôte and Jeangirard, 2021). Furthermore, some disambiguation methods focus only on local features and ignore the semantic information embedded in the surrounding context, which may lead to inaccuracies and limitations in the results (Han and Zhao, 2009; D'Angelo et al., 2020).

Building upon this foundation, the paper presents an end-to-end framework designed for entity recognition and disambiguation within author affiliation text. Harnessing sophisticated entity recognition techniques, the framework meticulously extracts crucial details such as institutions, countries, and more from the textual data. To tackle entity ambiguity, an innovative disambiguation algorithm, amalgamating word embedding and spatial embedding, is introduced. This algorithm adeptly resolves ambiguities by capturing contextual semantic and geographic features. Specifically, word embedding endeavors to grasp the semantic relationships among entities, while spatial embedding enriches disambiguation prowess through geographical insights. By enhancing identification accuracy and robustness, this framework not only mitigates information redundancy and misidentification but also furnishes a more dependable groundwork for subsequent bibliometric analyses. Overall, the contributions of this work can be summarized as follows:

1) A novel framework for entity recognition and disambiguation is proposed to address inconsistencies and incompleteness in author affiliation texts.

2) A new disambiguation algorithm combining word embedding and spatial embedding is introduced by capturing contextual semantic and geographic features.

3) Experiment validation is conducted to demonstrate the superiority of the proposed algorithm over publicly available baseline models, showcasing its robustness and efficacy.

2 Related work

Entity disambiguation has been extensively studied in the past. Its methodological studies can be broadly categorized into three types: rule-based approaches, knowledge-based approaches, and combinations of these two approaches (Chandrasekaran and Mago, 2021). The rule-based approach relied mainly on manually crafted rules. These rules could be based on the name of the entity, contextual information, or other linguistic features, such as "IBM" commonly referring to the International Business Machines Corporation. However, these rules were often limited by their lack of generalization and scalability. With the emergence of large-scale corpora and structured knowledge bases, researchers began exploring how to fully leverage these rich resources to address entity disambiguation problems. Common knowledge bases include Wikipedia, DBpedia, and Yago (Sanyal et al., 2021). Knowledge-based approaches utilize the structured information and entity relationships within these knowledge bases, such as entity descriptions, hypernym relationships, and linking relationships, to enhance the accuracy and efficiency of entity disambiguation. Meanwhile, word embedding models such as Word2Vec, GloVe, and FastText were introduced and widely applied. These models map words to high-dimensional vector spaces and capture semantic relationships between words by learning distributed representations of words (Zwicklbauer et al., 2016). Pre-trained word embedding models can generate vector representations for entities and their contexts, encoding the semantic information

of entities into positional relationships in continuous vector space. This approach enables the semantic information of entities to be more effectively utilized for entity disambiguation tasks, thereby improving the performance and effectiveness of entity disambiguation (Basile et al., 2024). In addition, methods such as tf-idf (Term Frequency-Inverse Document Frequency), LSI (Latent Semantic Indexing), and LDA (Latent Dirichlet Allocation) have also been introduced into entity disambiguation tasks to extract semantic information and contextual associations of entities in text (Bouarroudj et al., 2022).

The use of a single method no longer satisfies current research needs, prompting researchers to combine different methods to obtain richer semantic information and more accurate entity representations. Babelfy leverages large-scale multilingual knowledge graphs and word embedding models to achieve entity linking and disambiguation. It adopts a context-based approach by analyzing the contextual information surrounding entities in the text to determine their semantics. Specifically, Babelfy treats each word in the text as a potential entity mention and uses contextual information and semantic relationships in the knowledge graph to infer the most likely entity corresponding to each mention(Li et al., 2020). The WeLink model proposed by Bellatreche et al. achieves entity disambiguation by generating a list of entity candidates from knowledge graph DBpedia, and utilizing context similarity, entity coherence, relation exploitation, entity name distance, and syntactic features to compute scores and rank candidate entities (Nedelchev et al., 2020).

Some research focuses on social media datasets, such as tweets and news headlines, which often possess characteristics of brevity and noise. However, they also contain rich spatiotemporal metadata. Consequently, some researchers have begun exploring the utilization of spatiotemporal signals in the entity disambiguation process (Agarwal et al., 2018; Fetahu et al., 2021; Rafiei, 2016). Fang and Chang (2014) proposed a method for entity disambiguation of Weibo tweets by integrating spatiotemporal signals. Considering that the dataset used in this paper contains abundant spatial signals, we drew inspiration from their work and introduced spatial signals to enhance the results of semantic similarity calculation. In this process, we reference the method proposed by Srinivasan and Rafiei (2021) that utilizes containment relationships between locations. However, unlike them, they emphasize that location heavily depends on the locations mentioned in each document, while this paper directly considers all mentioned locations equally important.

3 The Proposed Framework

This paper proposes a new framework aimed at entity recognition and disambiguation of author affiliation data extracted from PubMed-exported datasets, which can provide a reliable basis for bibliometric analysis. As shown in Figure 1, the framework can be divided into two parts: text preprocessing and entity disambiguation. In general, text preprocessing can be summarized into two main steps. The first is to extract author affiliation information, and the second is to identify entities involved in the text, such as institution, country, and other bibliometric indicators. The purpose of text preprocessing is to prepare text data to ensure usability, providing efficient data processing for subsequent entity disambiguation tasks. Disambiguation is the most critical part of this framework, which addresses the issues of candidate generation and candidate ranking for ensuring that ambiguous entities are accurately disambiguated in order to determine the exact meaning and denotation.

3.1 Preprocessing

Although the dataset exported from PubMed contains a large amount of literature information, this paper mainly focuses on the processing of author affiliation data. Therefore, the author affiliation information is at first extracted using a rule-based approach. Then, standard tools are used to detect mention of named entity, where the named entity type is organization and location. The preprocessing process is important because it affects subsequent entity disambiguation tasks.

- Affiliation extraction

Author affiliation usually appears in the form of a specific identification, prefixed by “AD –”, in datasets exported from PubMed databases. A rule-based approach is consequently used, specifically regular expressions to identify text paragraphs in the dataset that begin with “AD –” and to store these paragraphs. The “AD –” prefix is removed from the stored results to ensure the purity of the

data. Such a process is intended to remove redundant information so that the data can be more efficiently prepared for subsequent analysis and research. Furthermore, an author may have multiple affiliations, with only the first one being taken as the author's affiliation address.

- Entity recognition

Named Entity Recognition (NER) is a natural language processing technique used to automatically identify and extract specific named entities from specified text, such as names of individuals, locations, and organizations (Naseer et al., 2021). Although the focus of this work is not on researching entity recognition methods, we are still committed to selecting NER tools suitable for this type of dataset. This is because the failure to recognize the corresponding entity can adversely impact subsequent processing results. In this study, the GliNER model is employed for the entity recognition task, selected for two main reasons. On one hand, previous studies have demonstrated the excellent performance of the GliNER model, particularly in zero-shot testing, where it outperforms ChatGPT and fine-tuned autoregressive language models. On the other hand, in the dataset used in this paper, a simple comparison with various NER toolkits including spaCy, NLTK, and Stanford NER reveals that the GliNER model exhibits superior performance (Loper and Bird, 2002; Schmitt et al., 2019).

The GLiNER model is a model for named entity recognition that utilizes a bidirectional Transformer structure. The framework of the GLiNER model is depicted in Figure 2. The input to the model is a uniform sequence containing the entity types represented in natural language and the input text to be processed for recognition.

While processing the input, the model computes the interactions between all the tokens through token representations, generating contextualized representations to better understand the relationships between tokens and the contextual information. Next, the Entity and Span Representation module encodes entity types and span embeddings into a unified latent space. In this process, the entity type is refined by a two-layer feedforward network, and the span is obtained by calculating the feedforward network results for the representations at the two locations. The computation of the span representation can be easily parallelized and an upper limit on the span length is set to maintain linear complexity. Finally, the Entity Type and Span Matching module calculates the matching score, which is used to assess whether the type and span correspond (Zaratiana et al., 2023).

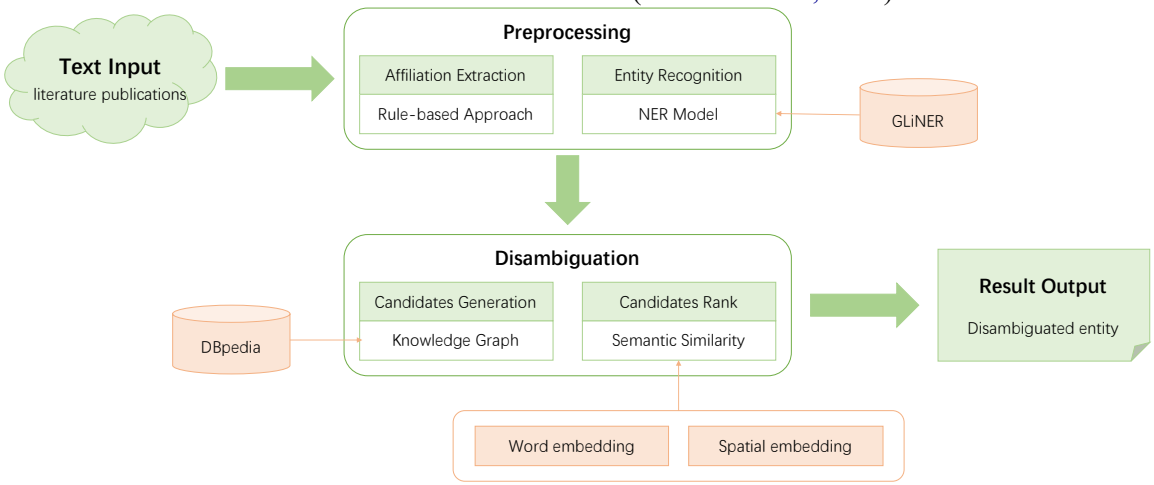

Figure 1: Overall framework of entity recognition and disambiguation.

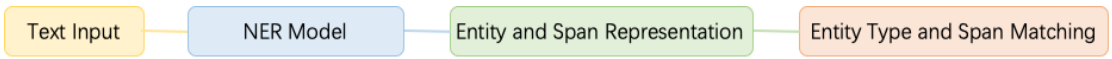

Figure 2: The Architecture of GLiNER.

3.2 Disambiguation

- Candidates generation

DBpedia is a knowledge graph constructed based on Wikipedia content, covering a large number of entities and their related information, including attributes, relationships, abstracts, categories, and

associations with other entities (Auer et al., 2017). By querying the DBpedia knowledge graph, one can obtain a list of candidate entities related to the mentioned entity. These entities have rich attribute descriptions and association information, which can provide important clues and support for subsequent entity ranking (Lehmann et al., 2015).

Candidate generation involves retrieving candidate entities related to entity mentions from the DBpedia knowledge graph, while retaining their characteristics, summaries, and other relevant information. To enhance matching accuracy and coverage, we not only directly obtain entities with names identical to the entity mentions from DBpedia, but also match entities related to the entity mentions, thereby expanding the matching scope of entities and making the query results more comprehensive. Additionally, entity types are utilized during candidate generation to limit the number of candidate words and to some extent eliminate ambiguity, thereby improving the quality of candidate word generation.

- Candidates Rank

Candidate ranking actually assigns a score to each candidate entity and ranks the generated candidate entities based on the score. The semantic similarity score is composed of context similarity and geographic similarity. The context similarity score of the candidate entity is obtained by comparing the semantic similarity between the context of the mentioned entity and each candidate entity. While geographic similarity is calculated by comparing the similarity between the geographic information contained in the input text and in the candidate entity abstracts.

1) Context similarity

Computing the similarity between entities and candidates using word embedding models is the most intuitive way to solve the ambiguity problem. Therefore, the similarity calculation between the entity context and the returned candidate entity summaries is performed using the word embedding model. At first, a pre-trained GloVe word embedding model is used, which maps words into a real vector space with hundreds of dimensions to capture semantic relationships between words. The selection of the GloVe model is based on its utilization of global co-occurrence information and its straightforward and efficient training methodology. The GloVe model learns word embeddings by minimizing the difference between the co-occurrence probabilities of words and their vector inner products (Pennington et al., 2014). For two words i and j, their co-occurrence probabilities can be represented by a co-occurrence matrix X. The element X_{ij} in this matrix represents the co-occurrence frequency of word i and word j. The optimization objective of the GloVe model is to minimize the loss function J in order to learn the best representation of the word vectors which allows the co-occurrence probabilities of the words to be well fitted.

$$J = \sum_{i,j=1}^{V} f(X_{ij})\left(w_i^T \cdot w_j + b_i + b_j - \log(X_{ij})\right)^2 \quad (1)$$

$f(X_{ij})$ is a weight coefficient introduced to regulate the effect of certain word pairs with too high co-occurrence frequency, w_i and w_j are the word vectors for words i and j, and b_i and b_j are bias terms used to regulate the position of the word vectors.

The text to be compared is converted into word embedding vectors, and then the overall representation vector of the text is obtained by simply averaging the vectors of all words in the text. Next, cosine similarity is used to calculate the similarity between the context vector V_e of the entity mention and the context vector V_c of the candidate entity.

$$Context_{similarity}(e,c) = \frac{V_e \cdot V_c}{\|V_e\| \cdot \|V_c\|} \quad (2)$$

2) Geographic similarity

The inheritance hypothesis refers to the assumption that the geographical location of a named entity can be inferred from the documents containing that entity (Kamalloo and Rafiei, 2018), such as news article headlines, tweet contents, etc. In other words, the frequency of occurrence of a certain location in the summary of a candidate entity can reflect the degree of association between the entity and that location. Using the NER tool to tag all locations l in the candidate summary d and obtain the corresponding frequency count $f(l,d)$. With these frequency counts, calculate the inheritance probability $P(l|d)$ of each location in the document. We assume that the geographical location with the highest frequency of occurrence l_c can represent the primary location of the candidate entities. It's worth noting that, unlike retrieving associated locations from the summary

of candidate entities, entity mention l_e directly relies on the original input text. This is because affiliations often already contain relevant geographical location information.

$$P(l|d) = \frac{f(l,d)}{\sum_{l' \in d} f(l',d)} \quad (3)$$

The geographic names associated with entities and candidates are described as vectors and spatial similarity is calculated using cosine similarity.

$$Geographic_{similarity}(e,c) = \frac{V_{l_e} \cdot V_{l_c}}{\|V_{l_e}\| \cdot \|V_{l_c}\|} \quad (4)$$

During this process, for each mention of a possible location, a search is conducted in a geographic database to obtain a list of potential matches. Additionally, to address geographic ambiguity, the inclusion relationships between different mentions within the same text are utilized. Specifically, this involves mapping their names to corresponding geographic location names in a country-level hierarchical structure to specify the exact level of geographic entity and thus determine the name of the geographic entity. GeoNames is selected for the use of geodatabases. The GeoNames database is a global geographical information database that encompasses a wide range of geographical data, including place names, geographical features, and administrative divisions. This database contains geographic coordinates and related attribute information for millions of locations worldwide. Serving as a crucial data source, the GeoNames database provides fundamental support for the semantic information of geographical locations, effectively facilitating geographic relevance analysis and other related research endeavors (Grütter et al., 2017).

3) Objective function

Each mentioned entity e_i has multiple candidates, i.e., $|C(e_i)| \geq 2$, and for each candidate $c_{i,j} \in C(e_i)$, two different similarity scores need to be computed, namely, the similarity score between context-aware embeddings of the mentions and candidate embeddings, and the similarity score between spatial embeddings of the mentions and candidates. The overall disambiguation Score for each candidate is then obtained by combining the contextual similarity and spatial similarity and ranking the candidates according to the total score.

$$Score = \alpha \cdot Context_{similarity}(e_i, c_{i,j}) + \beta \cdot Geographic_{similarity}(e_i, c_{i,j}) \quad (5)$$

4 Experiments & results

4.1 Data

The data employed are author affiliations extracted from PubMed database. More specifically, the data is made up of strings containing information about organizations, regions, countries, etc. The annotation process is time-consuming as it usually requires manual annotation to recognize the raw data. In order to improve the efficiency of data labeling, a two-step strategy is used. Some entity names are first pre-labeled using a rule-based criterion that suggests target entity names, and then these suggested entity names are manually checked. The rule-based approach refers to the method of text processing based on predefined rules or patterns. Considering that the text structure and format of the dataset is relatively fixed, when pre-tagging the country entity data, it is straightforward to divide the text according to the specified punctuation and select the last part as the tagging target on the condition of ensuring that the last part is not an email message. While pre-tagging the data of institutional entities, texts containing terms such as “University”, “Center”, “College”, “Institute”, or “Foundation” are used as tagging targets, taking into account the common features and representative terms in the names of institutions. After the pre-tagging process, these suggested entity names are manually validated. The 9,000 records are annotated eventually. It is noteworthy that this study exclusively considers the highest organizational level of institutions, namely universities or research organizations. Specifically, all affiliations are simplified to their respective universities or primary research institutions, without further subdivision into colleges, departments, or divisions. This approach aids in data processing and analysis, enhancing data consistency and comparability.

The dataset is equally divided into three subsets S_1, S_2, and S_3, each of equal size, using K-fold cross validation (Bhagat and Bakariya, 2022). In each iteration, one of the subsets is used as the test set S_{test} and the concatenated set

of the remaining parts is used as the training set S_{train}^{L}. To further analyze the effect of different sizes of training data on model performance, 1/3 of them are sampled from S_{train}^{L} as the medium-sized training set S_{train}^{M}, and then 1/20 of them are sampled from S_{train}^{M}as the smaller training set S_{train}^{S}. Three training sets with different sizes are obtained in the end: smaller S_{train}^{S}, medium S_{train}^{M}, and the initial size S_{train}^{L}. The inclusion relationship of these three training sets is $S_{train}^{S} \subset S_{train}^{M} \subset S_{train}^{L}$, i.e., S_{train}^{S} is a subset of S_{train}^{M}, which in turn is a subset of S_{train}^{L}.

4.2 The results

To assess the stability of the performance of model, we conducted a series of experiments covering training datasets of different scales, ranging from smaller-scale training data to the initial scale of training data. This design enables a comprehensive examination of the model's performance when faced with different data volumes, allowing for a more accurate evaluation of the performance and robustness of model. Table 1 intuitively demonstrates the performance of the model under different data volumes. Despite the differences in the scale of the training datasets, there is not much variation in the measured metrics of Precision, Recall, and F1-score on the test set. This indicates that our model exhibits a certain degree of stability and robustness when dealing with training data of different scales.

Table 1: the performance of proposed model on different size datasets

Data size	Precision	Recall	F1
Small-scale S_{train}^{S}	0.748	0.761	0.754
Medium-scale S_{train}^{M}	0.752	0.761	0.756
Large-scale S_{train}^{L}	0.745	0.771	0.758

Furthermore, to further evaluate the effectiveness of our algorithm, we conducted comparison experiments with other methods. From Table 2, it is evident that our algorithm demonstrates good performance, particularly in terms of accuracy. In fact, there is further potential for improvement in rule-based entity recognition and disambiguation, but this requires continuous observation of actual text samples and iterative adjustment and addition of rules based on identification results. This not only demands more time and effort but also rules may not comprehensively cover all cases, leading to less accurate or incomplete identification results in certain scenarios.

Table 2: compare with other models

Methodology	Precision	Recall	F1
Rule-based	0.582	0.708	0.639
WeLink	0.711	0.732	0.722
Our algorithm	0.745	0.771	0.758

To validate the importance of geographical information in semantic similarity calculation and to gain a deeper understanding of its impact on disambiguation results for different entity types, we conducted ablation experiments separately for location and organization entity types to explore the influence of geographical embeddings on disambiguation algorithm. The results are illustrated in Table 3. By incorporating geographical information, we can observe a significant improvement in disambiguation effectiveness for Location entity types. Specifically, the Precision metric increased by 0.288, Recall increased by 0.221, and the F1-score also relatively increased by 0.179. This is because utilizing containment relationships between geographical locations in the text can more effectively eliminate geographic ambiguity, thus enhancing the accuracy and overall performance of entity recognition. However, upon examining the analysis results, we found that when using the GeoNames database to process geographical names, especially when mapping them to the country level, the database cannot handle aliases of individual country names such as “UK”, “The Netherlands”. For instance, in the input text “Cancer Services, South Eastern Health & Social Care Trust, UK.”, it fails to correctly identify the mentioned entity “UK” as “United Kingdom, and incorrectly maps it to “[‘Japan’, ‘Japan’, ‘Kazakhstan’, ‘Pakistan’, ‘Russia’, ‘Ukraine’, ‘United States’]”, which somewhat affects the ranking of candidates. Based on this, this study plans to solve the problem by creating a higher priority mapping table designed to store those country aliases that the GeoNames database cannot handle. As for entity type of Organization, their metric values also improved. Specifically, accuracy increased by 13.5%, recall increased by 3.8%, and F1-score increased by 8.1%. Although the improvement is not as significant as for Location types, it still

contributes to enhancing entity disambiguation results. Particularly, when dealing with organizations with the same name, the introduction of geographical information can quickly identify the correct candidates. For example, it can effectively differentiate between the "University of California, Los Angeles" and "University of California, Santa Barbara". Additionally, it is noteworthy that compared to the recall values, the change in precision is more pronounced, indicating that the introduction of geographical information prioritizes improving the precision of entity disambiguation. Actually, some of the author affiliation information that is severely lacking in correct and necessary information does not lead to correct results. For example, "Department of Biology", but the proposed algorithm still significantly improves the precision and recall.

Table 3: proposed model with and without location embedding for different entity type

Entity Type	Methodology	Precision	Recall	F1
Location	Context	0.660	0.740	0.698
	Context + Geographic	0.948	0.961	0.955
Organization	Context	0.703	0.667	0.684
	Context + Geographic	0.838	0.705	0.765

5 Case study

Table 4: The top 10 institutions by count

Institution	Count
University of California	127
Stanford University	108
University of Texas	69
University of Washington	61
Massachusetts General Hospital	58
University of Pennsylvania	56
University of Oxford	54
Mayo Clinic	52
University of Michigan	44
University of Toronto	44

Table 5: The top 10 countries by count

Country	Count
United States	1843
China	1250
United Kingdom	489
India	413
Germany	344
Italy	333
Canada	324
France	265
Japan	263
Australia	259

In this paper, two case studies are defined to demonstrate the practical application of the proposed framework. In these case studies, we aim to identify the most productive organizations and countries on the topic of "artificial intelligence" in the PubMed database. We apply the proposed framework to extract author affiliation data and perform entity recognition and entity disambiguation. The 10 most productive institutions and countries are listed in Tables 4 and 5, respectively.

6 Conclusion

This paper proposes an end-to-end named entity disambiguation framework designed to achieve entity recognition and disambiguation for author affiliation texts, enabling direct application in bibliometric analysis. Particularly, a novel algorithm combining word embeddings and spatial embeddings is introduced for entity disambiguation. This work not only provides methodological references for handling homogeneous types of textual data but also facilitates the automation of dataset processing, entity recognition, and disambiguation, thereby minimizing post-processing steps. These advancements allow for direct application in bibliometric analysis, ultimately enhancing its efficiency. Certainly, the current focus of this paper is primarily on resolving named entity ambiguities at the entity type of LOC and ORG. In the future, we aim to extend this entity disambiguation algorithm to analyze other entity types, such as author name. Additionally, the Research Organization Registry (ROR) is a system for managing and recording information about various types of organizations or institutions, which can be utilized as a knowledge repository to further expand our research.

References

Agarwal, Prabal, et al. "diaNED: Time-aware named entity disambiguation for diachronic

corpora." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2018. https://doi.org/10.18653/v1/P18-2109

Auer, Sören, et al. "Dbpedia: A nucleus for a web of open data." *international semantic web conference*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.https://doi.org/10.1007/978-3-540-76298-0_52

Basile, Alessandro, et al. "Disambiguation of company names via deep recurrent networks." *Expert Systems with Applications* 238 (2024): 122035. https://doi.org/10.1016/j.eswa.2023.122035

Bhagat, Meenu, and Brijesh Bakariya. "Implementation of logistic regression on diabetic dataset using train-test-split, k-fold and stratified k-fold approach." *National Academy Science Letters* 45.5 (2022): 401-404. https://doi.org/10.1007/s40009-022-01131-9

Bouarroudj, Wissem, Zizette Boufaida, and Ladjel Bellatreche. "Named entity disambiguation in short texts over knowledge graphs." *Knowledge and Information Systems* 64.2 (2022): 325-351. https://doi.org/10.1007/s10115-021-01642-9

Chandrasekaran, Dhivya, and Vijay Mago. "Evolution of semantic similarity—a survey." *ACM Computing Surveys (CSUR)* 54.2 (2021): 1-37. https://doi.org/10.1145/3440755

D'Angelo, Ciriaco Andrea, and Nees Jan Van Eck. "Collecting large-scale publication data at the level of individual researchers: A practical proposal for author name disambiguation." *Scientometrics* 123 (2020): 883-907. https://doi.org/10.1007/s11192-020-03410-y

Donner, Paul, Christine Rimmert, and Nees Jan van Eck. "Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems." *Quantitative Science Studies* 1.1 (2020): 150-170. https://doi.org/10.1162/qss_a_00013

Donthu, Naveen, et al. "How to conduct a bibliometric analysis: An overview and guidelines." *Journal of business research* 133 (2021): 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

Fang, Yuan, and Ming-Wei Chang. "Entity linking on microblogs with spatial and temporal signals." *Transactions of the Association for Computational Linguistics* 2 (2014): 259-272. https://aclanthology.org/Q14-1021

Fetahu, Besnik, et al. "Gazetteer enhanced named entity recognition for code-mixed web queries." *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021. https://doi.org/10.1145/3404835.3463102

Fiorini, Nicolas, et al. "PubMed Labs: an experimental system for improving biomedical literature search." *Database* 2018 (2018): bay094. https://doi.org/10.1093/database/bay094

Grütter, Rolf, Ross S. Purves, and Lukas Wotruba. "Evaluating topological queries in linked data using DBpedia and GeoNames in Switzerland and Scotland." *Transactions in GIS* 21.1 (2017): 114-133. https://doi.org/10.1111/tgis.12196

Han, Xianpei, and Jun Zhao. "Named entity disambiguation by leveraging wikipedia semantic knowledge." *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009. https://doi.org/10.1145/1645953.1645983

Huang, Shuiqing, et al. "Institution name disambiguation for research assessment." *Scientometrics* 99 (2014): 823-838. https://doi.org/10.1007/s11192-013-1214-2

Kamalloo, Ehsan, and Davood Rafiei. "A coherent unsupervised model for toponym resolution." *Proceedings of the 2018 world wide web conference*. 2018. https://doi.org/10.1145/3178876.3186027

Kokol, Peter, Helena Blažun Vošner, and Jernej Završnik. "Application of bibliometrics in medicine: a historical bibliometrics analysis." *Health Information & Libraries Journal* 38.2 (2021): 125-138. https://doi.org/10.1111/hir.12295

Lehmann, Jens, et al. "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia." *Semantic web* 6.2 (2015): 167-195. DOI: 10.3233/SW-140134

L'Hôte, Anne, and Eric Jeangirard. "Using Elasticsearch for entity recognition in affiliation disambiguation." *arXiv preprint arXiv:2110.01958* (2021). DOI: 10.48550/arXiv.2110.01958

Li, Fei, et al. "An efficient approach for measuring semantic similarity combining WordNet and Wikipedia." *IEEE Access* 8 (2020): 184318-184338. DOI: 10.1109/ACCESS.2020.3025611.

Lim, Weng Marc, and Satish Kumar. "Guidelines for interpreting the results of bibliometric analysis: A sensemaking approach." *Global Business and Organizational Excellence* 43.2 (2024): 17-26. DOI: 10.1002/joe.22229

Loper, Edward, and Steven Bird. "Nltk: The natural language toolkit." *arXiv preprint cs/0205028* (2002). https://aclanthology.org/P04-3031

Lu, Zhiyong. "PubMed and beyond: a survey of web tools for searching biomedical literature." *Database* 2011 (2011): baq036. DOI: 10.1093/database/baq03

Naseer, Salman, et al. "Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance." *Pakistan Journal of Multidisciplinary Research* 2.2 (2021): 293-308.

Nedelchev, Rostislav, et al. "End-to-end entity linking and disambiguation leveraging word and knowledge graph embeddings." *arXiv preprint arXiv:2002.11143* (2020). https://doi.org/10.48550/arXiv.2002.11143

Ninkov, Anton, Jason R. Frank, and Lauren A. Maggio. "Bibliometrics: methods for studying academic publishing." *Perspectives on medical education* 11.3 (2022): 173-176. DOI: 10.1007/S40037-021-00695-4

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014. https://aclanthology.org/D14-1162

Rafiei, Jiangwei Yu, and Davood Rafiei. "Geotagging named entities in news and online documents." *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 2016. DOI: 0.1145/2983323.2983795

Rimmert, Christine, Holger Schwechheimer, and Matthias Winterhager. "Disambiguation of author addresses in bibliometric databases-technical report." (2017). https://pub.uni-bielefeld.de/record/2914944

Sanyal, Debarshi Kumar, Plaban Kumar Bhowmick, and Partha Pratim Das. "A review of author name disambiguation techniques for the PubMed bibliographic database." *Journal of Information Science* 47.2 (2021): 227-254. https://doi.org/10.1177/0165551519888605

Schmitt, Xavier, et al. "A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate." *2019 sixth international conference on social networks analysis, management and security (SNAMS)*. IEEE, 2019. DOI: 10.1109/SNAMS.2019.8931850

Shao, Zhou, et al. "ELAD: An entity linking based affiliation disambiguation framework." *IEEE Access* 8 (2020): 70519-70526. DOI: 10.1109/ACCESS.2020.2986826

Srinivasan, Maithrreye, and Davood Rafiei. "Location-aware named entity disambiguation." *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021. https://doi.org/10.1145/3459637.3482135

Thompson, Dennis F., and Cheri K. Walker. "A descriptive and historical review of bibliometrics with applications to medical sciences." *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 35.6 (2015): 551-559. https://doi.org/10.1002/phar.1586

Zaratiana, Urchade, et al. "Gliner: Generalist model for named entity recognition using bidirectional transformer." *arXiv preprint arXiv:2311.08526* (2023). https://doi.org/10.48550/arXiv.2311.08526

Zwicklbauer, Stefan, Christin Seifert, and Michael Granitzer. "Robust and collective entity disambiguation through semantic embeddings." *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 2016. https://doi.org/10.1145/2911451.2911535