


Article

Normalization of Web of Science Institution Names Based on Deep Learning

Zijie Jia , Zhijian Fang and Huaxiong Zhang *

School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China; 202120503037@mails.zstu.edu.cn (Z.J.); hptnt@zstu.edu.cn (Z.F.)

* Correspondence: zhxhz@zstu.edu.cn

Abstract: Academic evaluation is a process of assessing and measuring researchers, institutions, or disciplinary fields. Its goal is to evaluate their contributions and impact in the academic community, as well as to determine their reputation and status within specific disciplinary domains. Web of Science (WOS), being the most renowned global academic citation database, provides crucial data for academic evaluation. However, due to factors such as institutional changes, translation discrepancies, transcription errors in databases, and authors' individual writing habits, there exist ambiguities in the institution names recorded in the WOS literature, which in turn affect the scientific evaluation of researchers and institutions. To address the issue of data reliability in academic evaluation, this paper proposes a WOS institution name synonym recognition framework that integrates multi-granular embeddings and multi-contextual information.

Keywords: Web of Science; institution name normalization; deep learning

1. Introduction

Academic evaluation is a process of assessing and measuring researchers, institutions, or disciplinary fields. Its purpose is to evaluate their contributions and impact in the academic community, as well as determine their reputation and status within specific disciplinary domains. It is an important factor in government decision-making and resource allocation. As the quantity and quality of publications serve as significant indicators for academic evaluation, Web of Science (WOS), one of the world's most renowned academic citation databases, is commonly used for academic research evaluation [1], ranking [2], and comparison [3] by researchers and academic institutions. In addition, the results of an academic evaluation affect the analysis of university education. For example, Laura [4] analyzed 17 communication and journalism courses from eight of Europe's highest-ranked universities in the field of communication based on the QS World University Rankings to assess the university's educational program.

However, according to a large-scale analysis conducted by Huang [5], the lists provided in WOS's Essential Science Indicators (ESIs) are not as reliable and accurate as one might expect. Approximately 25% of author names (consisting of the initials of their first name and last name) are shared by at least two different individuals. When explicit data are not provided by authors or publishers, data aggregators such as WOS or Scopus find it challenging to provide accurate or statistically reliable data. This issue is commonly referred to as the name ambiguity problem and can be divided into two parts: the one person, multiple names problem (where one author entity is associated with multiple name variants in different publications) and the one name, multiple persons problem (where one author name corresponds to multiple different author entities). Institutional information serves as an identity marker for authors in the literature. Research has shown that the probability of homonyms in secondary institutions is very low [6,7]. One approach to identifying homonymous author entities is by extracting primary and secondary institution names



Citation: Jia, Z.; Fang, Z.; Zhang, H. Normalization of Web of Science Institution Names Based on Deep Learning. *Algorithms* **2024**, *17*, 312. <https://doi.org/10.3390/a17070312>

Received: 27 May 2024

Revised: 4 July 2024

Accepted: 12 July 2024

Published: 14 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

from addresses using patterns, such as comma separators and “university, department, laboratory”.

However, research by Falahati [8] revealed that out of 84 universities in Iran, there are 1668 name variants in WOS, primarily stemming from abbreviations, spelling errors, spatial variations, syntactic arrangements, and vowel/consonant and vowel/consonant combinations, with spelling errors accounting for 34.57% of the variants. Confronted with a vast number of non-standardized institution entities, there are cases of the mislabeling and underlabeling of institution data in the institution lists of the ESIs and InCites (mislabeling refers to indexing an address belonging to institution A as institution B, while underlabeling occurs when an address belonging to an institution is not indexed under that institution). Due to reasons such as author spelling errors, transcription errors in systems, translation issues, variations in institution and department names, and the use of informal names or abbreviations, the same institution may have multiple different representations (Table 1), or an institution entity may be transcribed as another institution entity.

Table 1. Different expressions of the name of the institution.

Official Name	Variant Name	Description
Soochow Univ	Suzhou Univ	alias
	Soochow Univers	word abbreviation
	SUDA	acronym
	SoChow Univ	spelling errors
	SooChow	missing agency identifiers
	SooChow Univ	uppercase/lowercase variation
	University Soochow	syntactic arrangements
	Soochow Univ affiliated Hosp	nested entities

Scholars have conducted extensive research on the institution name synonym recognition task (Table 2). In the early stages of research, scholars extensively explored the similarity of institution names from both character and word perspectives using methods, such as edit distance [9] and Jaccard similarity [10]. However, institutions with low literal similarity may refer to the same entity, such as “Chinese Academy of Science” and “CAS” (full name and abbreviation) or “Chinese Academy of Science” and “Chinese Acad Sci” (full name and keyword abbreviation). Conversely, institutions with high literal similarity may be distinct entities, for example, “Fukushima Univ” and “Fukushima med Univ”. To address the limitations of literal similarity, some researchers have combined author name features, address features (city/state/country names) [11], and institution name features (organizational keywords) [12,13] with string similarity algorithms to achieve better results [14,15].

Another group of researchers [16] introduced statistical approaches by applying the principles of TF-IDF and analyzing a large number of institution names. They found that high-frequency words had limited discriminative power for distinguishing institution entities. To overcome this, they assigned different weights to words in institution names based on their frequencies and used the weighted average of different words in addresses to determine whether they referred to the same institution.

Some scholars have adopted entity linking methods in an attempt to link institution names to external knowledge bases. Initially, researchers used proprietary databases from governments or institutions [17]. However, these private databases were often small in scale and not publicly accessible, limited to disambiguating institutions within specific regions or fields. With the development of publicly available institution knowledge bases and big data technologies, recent studies have focused on constructing standardized models for institution names. These models link institution entities in bibliographic records to multiple-source institution identifiers [18–20], such as Wikidata, GRID, ISNI, Ringgold, ROR, etc.

With the advancement of deep learning, the advantages of automatically learning features from limited annotated data have been widely applied in the field of entity dis-

ambiguation. Currently, deep learning methods are less commonly used in institution disambiguation research on bibliographic data, with the predominant use of word vector-based approaches. These methods utilize word vector models such as Word2Vec, GloVe, and BERT to learn the semantic relationships of institution names and combine clustering, rules, or string-based methods to identify the form similarity, variants, and abbreviations of institution names [21].

Table 2. Methods for institution name synonym recognition.

Method	Advantages	Disadvantages
String Similarity-based	Simple and easy to implement; effective in handling spelling errors or minor variations.	Less effective in handling semantically similar but structurally different names or high literal similarity of distinct institution names.
Statistical-based	Can adapt to various fields and languages of bibliographic data because they rely on contextual information and statistical features.	Depend on the selection and construction of statistical features, requiring substantial data support, and may be influenced by data quality.
Rule-based	Simple and intuitive; effective in handling obvious cases.	Rule formulation can be complex and require human involvement; less effective in handling complex cases.
Entity Linking-based	Can leverage rich information in knowledge bases; effective in handling complex cases; enables automated construction of institution standard files.	Require high-quality knowledge base support; less effective in handling new institutions not present in the knowledge base.
Deep Learning-based	Can automatically learn and extract features and perform feature combinations; effective in handling complex cases.	Require a significant amount of annotated data; training and fine-tuning the models can be complex.

Currently, there are two main issues in institution name standardization using deep learning methods:

1. **Feature Extraction and Fusion:** Institution data features can be categorized into two main types: text features and semantic relationship features. Text features primarily measure the literal similarity of institution names, which are effective in identifying names that are similar in their literal form. However, they may perform poorly in handling institution aliases and abbreviations. On the other hand, semantic relationship features focus on analyzing co-occurrence relationships and hierarchical similarity between institutions, which can better identify aliases and abbreviations. However, they may sometimes incorrectly merge structurally similar but distinct institutions. The current research often employs techniques such as term frequency, TF-IDF, string similarity, and the longest common substring to extract text features and deep learning models such as Word2Vec to extract semantic features. These features are then combined through rules or weighted fusion. This approach separates the association between text and semantic features and introduces uncertainty and subjectivity in feature combination and fusion weight allocation.
2. **Utilizing Multiple Contextual Information of Institution Entities:** The current research often relies on single-context matching, where only the most similar context containing the institution entity is considered during the institution matching process. This

approach fails to fully leverage the multiple contextual information that an institution may appear in, thereby limiting the recognition accuracy.

To address these issues, this study proposes a synonym relationship recognition model that integrates multi-granularity features and multiple contextual information. The model combines Char-CNN and Word2Vec techniques to extract text and semantic features of institution entities and efficiently fuses different features using a Highway network. The model also utilizes BiLSTM combined with a multi-context matching layer to integrate the performance of institution entities in different texts, resulting in a comprehensive entity representation. Finally, the model uses cosine similarity to calculate the similarity between institutions, enabling accurate synonym relationship recognition. This multidimensional feature fusion approach effectively improves recognition accuracy and is suitable for handling complex institution name variants and structures.

This paper's contributions can be summarized as follows:

- Addressing the deficiencies in feature extraction and fusion for institution name standardization: This paper proposes the construction of an embedding layer that extracts and fuses two types of features. There may exist correlations and dependencies between different feature categories. By extracting features from different categories within a unified model, the model can share learned knowledge and representations, thereby improving generalization and effectiveness.
- Solving the issue of underutilizing multiple contextual information of institution entities: This paper introduces a method based on bidirectional matching and multi-context fusion. This approach effectively leverages the multiple contexts in which institution entities may appear. By considering and integrating information from different contexts, the model achieves a more comprehensive understanding of institution entities, leading to improved accuracy in recognition.

These contributions aim to enhance the performance and robustness of the institution name standardization task by improving feature extraction, fusion, and the utilization of contextual information.

This paper is structured as follows: Section 2 summarizes the relevant work. Section 3 describes our approach, including the individual modules of the institutional synonymous recognition model. Next, Section 4 will report on the experiments and results. Section 5 summarizes and discusses future work.

2. Related Works

Based on the existing literature, scholars from both domestic and international contexts have conducted extensive theoretical and practical research on institution name synonymous recognition and standardization.

In the field of synonym recognition for institution names, various methods have been employed. The representative methods include the following:

1. String similarity-based methods: Common algorithms, such as the edit distance, Jaccard coefficient, and TF-IDF, are used to measure the similarity between institution names. The edit distance represents the minimum number of edit operations (insertion, deletion, or substitution) required to transform one string into another. French [9] proposed the relative edit distance, which uses the edit distance divided by the minimum length of the two institution names to measure the similarity. To address syntactic variations in institution names, French also introduced the word-based edit distance, which splits institution names into words and calculates the edit distance based on approximate word matching.
2. Statistical-based methods: These methods leverage the statistical characteristics of institution name occurrences, such as word frequency, co-occurrence relationships, and contextual features, to differentiate between different institutions. Onodera [22] assigned different weights to words based on their frequency and measured the similarity between two institution names by summing the weights of matching words.

Jiang [16] proposed a clustering method using the Normalized Compression Distance (NCD) to match institution documents. The NCD utilizes data compression techniques to measure the similarity between two texts, assuming that if two texts are semantically similar, their compressed representations should exhibit high redundancy and similarity. Cuxac [23] addressed naming ambiguities, spelling errors, OCR errors, abbreviations, and omissions by employing two strategies: one utilizing a Naive Bayes model when training data are available and the other employing a semi-supervised approach combining soft clustering and Bayesian learning when no learning resources are present.

3. Rule-based methods: These methods involve constructing rule libraries based on features derived from institution names (e.g., string similarity, substrings, word length, word order, and institution type) and additional features from the literature data (e.g., country, city, postal code, and author names) to merge institution name matches using feature-based rules. Huang [5] proposed a rule-based and edit distance-based approach for institution name standardization. They first constructed an institution–author table and used the author, country, postal code, and other features for potential institution name matching. Then, they calculated similarity by combining the Jaccard word similarity, substring matching, and the edit distance to identify institution name variants. Researchers from Bielefeld University developed over 50,000 pattern matching rules utilizing features such as institution the name, start and end dates, URL, postal code, sectors (name, URL, and sub-classification), and relationships between institutions to disambiguate the author addresses in WOS and Scopus.
4. Entity linking-based methods: These methods resolve ambiguity by linking institution names in the literature to corresponding institutions in knowledge bases. Shao [20] proposed the ELAD framework, which utilizes knowledge graphs for entity linking, generating a candidate set of institution entities, and then selecting the most probable institution entity based on string similarity. Wang [19] introduced a framework that utilizes open data resources to assist institution name standardization and attribute enrichment. It involves normalizing institution names and enriching attributes using open data resources, constructing a data linking model for multidimensional attribute alignment, and proposing a dynamic management approach for open data.
5. Deep learning-based methods: These methods utilize word embedding models to obtain distributed vectors containing rich semantic information from raw data. These vectors are then used in subsequent deep learning models or for vector similarity comparison. Sun [24] applied the Word2Vec word embedding model to semantically learn the SCI address field and disambiguate institution names based on the similarity of institution word vectors. Chen et al. [21] utilized the GloVe model to learn institution vector representations and applied DBSCAN clustering to institution names based on vector similarity and matching rules.

In WOS, the characteristics of institutional data are divided into two broad categories: textual features and semantic relational features. The text feature method mainly compares the literal similarity of institution names and uses techniques such as word frequency, TF-IDF, string similarity, and the longest common substring to judge the similarity between institutions. This method is effective in identifying literally similar organization names, but it does not perform well when dealing with aliases and abbreviations of institutions. In contrast, the semantic relationship feature focuses on the analysis of co-occurrence and hierarchical similarity between institutions, and can better identify aliases and abbreviations, but sometimes mistakenly groups together structurally similar but substantially different institutions.

In order to improve the recognition effect of synonymous relations, the method of combining these two features is particularly important. Through the manual observation and weighted fusion of these features, the key information that is conducive to distinguishing institutions can be extracted in a targeted manner. However, this method has some uncertainty and subjectivity in constructing feature combinations and assigning fusion

weights. In addition, the current research often relies on single-context matching, that is, only the most similar affiliation strings containing institutional entities are considered in the institution matching process, and the multiple contextual information that may occur in institutions is not fully utilized. This limits the recognition accuracy. Therefore, this chapter proposes a synonymous relationship recognition model that integrates multi-granularity features and multi-context information.

3. Institutional Synonym Recognition Model

3.1. Overview of the Proposed Model

Kim [25] conducted research indicating that the use of subword features, such as stems and affixes, can effectively identify the abbreviated forms of words. This approach reveals the potential of subword features in capturing the microstructure of language, particularly in the recognition of abbreviations and contractions, where it demonstrates high performance. Based on this finding, we have chosen to employ Char-CNN to extract character-level features from institution names. Char-CNN allows for the in-depth analysis of the internal structure of words, enabling the identification and learning of specific character sequences or combinations. This capability proves particularly effective in handling spelling errors, abbreviations, and domain-specific language. By incorporating Char-CNN in our approach, we not only enhance the model's ability to perceive subtle textual differences but also improve its robustness when dealing with anomalous text.

Word2Vec is capable of capturing the semantic similarity between words, but it does not capture the importance and distribution of words within a document collection. Therefore, we utilize Word2Vec in combination with TF-IDF to obtain word-level features. In this paper, we employ Highway networks [26] to integrate character-level and word-level features. This network structure effectively controls the flow of information between different features through its gating mechanism.

Specifically, the sigmoid function in Highway networks determines the proportion of information flow between character-level and word-level embeddings, while the fully connected layer appropriately transforms and adjusts the passed information. This approach allows for the model to flexibly integrate text features at both the character and word levels, fully leveraging the fine-grained information from character-level features and the semantic richness of word-level features. As a result, it enhances the accuracy and robustness of institution name recognition and matching.

In the task of institution synonym recognition, understanding and utilizing the hierarchical relationships of institutions are crucial for accurately determining whether two institutions refer to the same entity. To address this, we employ BiLSTM (Bidirectional Long Short-Term Memory) to aggregate contextual semantic information. BiLSTM is effective in capturing both preceding and succeeding contextual details, including semantic and syntactic information, thus providing a comprehensive semantic understanding.

Furthermore, considering the ambiguity and fuzziness inherent in natural language processing, our model incorporates multi-context matching techniques. By analyzing and comparing the relationships between different contexts, the model enhances its ability to capture semantic information. Multi-context matching allows for the model to automatically determine which contexts are more critical for interpreting the semantics in a sentence by learning the matching relationships and corresponding weights between different contexts. This approach not only improves the model's expressive power but also enhances its robustness and accuracy when dealing with semantic complexity.

Ultimately, by calculating the cosine similarity of the fused feature vectors, the model is able to determine whether two institution entities are synonymous. The architecture of the model is illustrated in Figure 1.

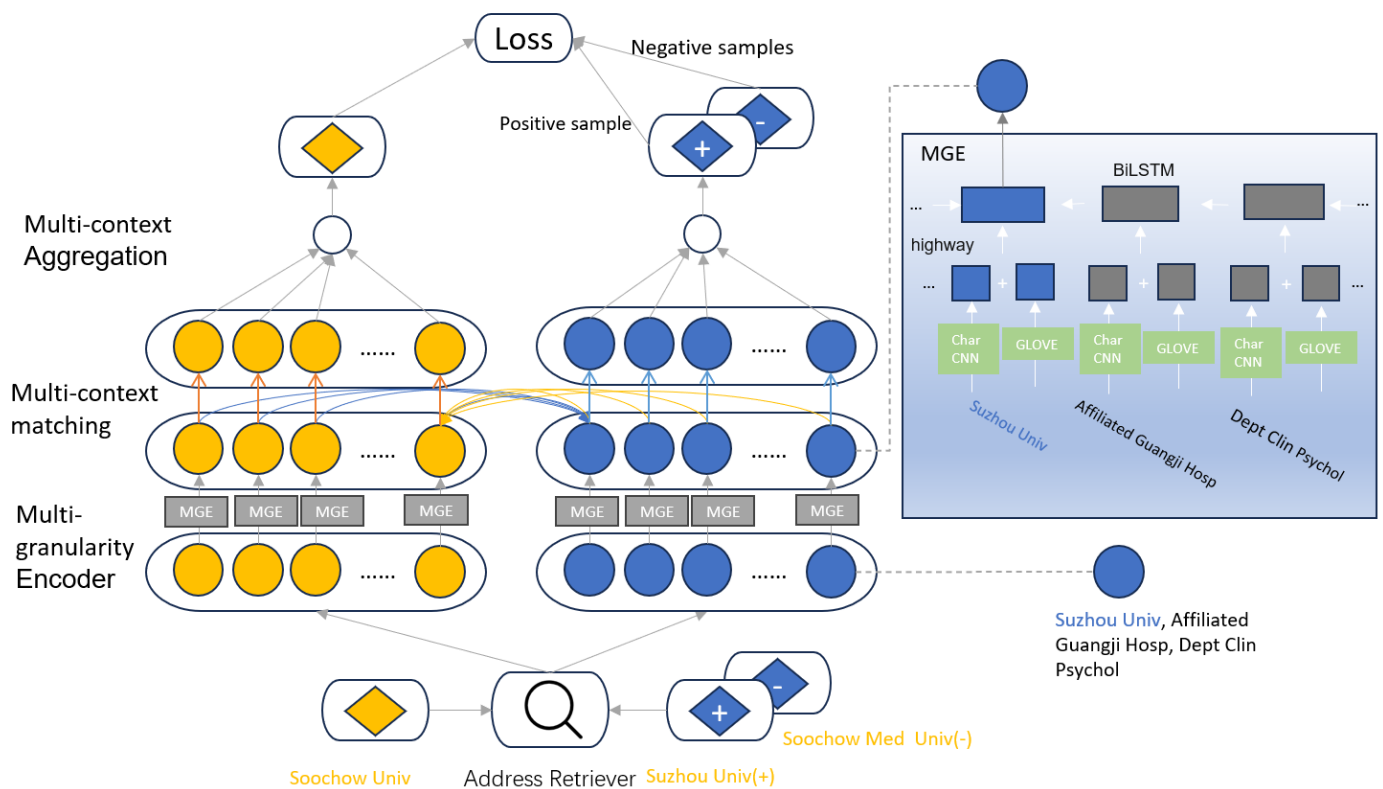


Figure 1. Model architecture.

3.2. Address Retriever

For a candidate entity e_l in the entity set E and its formal name O , the address retriever retrieves the most similar address segments from the corpus D where the entity appears. The retrieved addresses of e are represented as a set $A = \{a_1, a_2, \dots, a_p\}$, where p is the number of address segments.

3.3. Multi-Granularity Feature Embedding Layer

Character-level Feature Extraction: Let c be the vocabulary of characters and d be the dimensionality of character embeddings. For each word a , its character sequence is denoted as (c_1, c_2, \dots, c_l) , where l is the length of word a . The vector matrix representation of word a is denoted as $C^a \in R^{d \times l}$. We use a convolution between C^a and multiple filters (or kernels) $H \in R^{d \times w}$ of width w . Char-CNN does the following:

$$x^a = \max(\sum_0^{l-w+1} \tanh(\langle C^a[i : i + w - 1], H \rangle + b)) \quad (1)$$

where $C^a[i : i + w - 1]$ represents the i to $(i + w - 1)$ th column of matrix C^a , and $\langle A, B \rangle$ is the Frobenius inner product. Filters essentially extract n-gram character sequences from words, where the size of the n-gram corresponds to the width of the filter. This represents taking the maximum value, which is used to capture the most important features for a filter. For a word, this study employs a total of m convolutional filters. The structure of Char-CNN is illustrated in Figure 2 below.

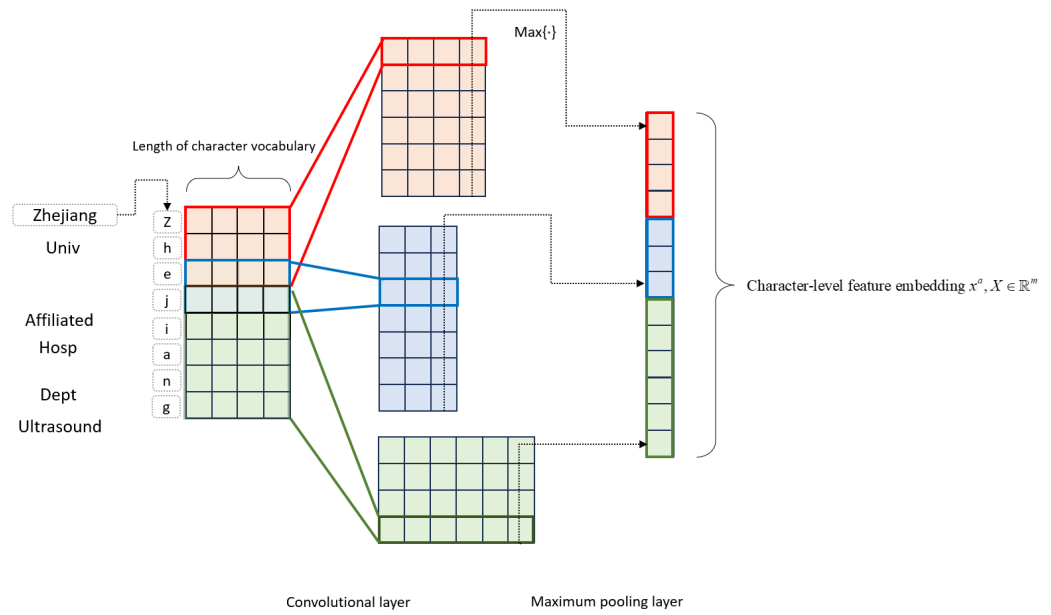


Figure 2. The structure of Char-CNN.

Word Embedding: In this study, the pre-trained Word2Vec and TF-IDF are utilized to obtain semantic embeddings for each word, with a length of n . We concatenate the character feature embeddings with the word semantic embeddings and denote the resulting representation as $y^a = [y_1^a, \dots, y_{n+m}^a]$, $y \in R^{n+m}$. For y^a , this study utilizes a Highway network to adjust the relative contributions of word semantic embeddings and character feature embeddings, thereby obtaining a more effective word representation. The Highway network employs a gating mechanism to control the flow and transformation of information, which is represented by the following equation:

$$z = t \cdot g(W_H y + b_H) + (1 - t) \cdot y \tag{2}$$

$$t = \sigma(W_T y + b_T) \tag{3}$$

Let W represent the weight matrix and b denote the bias. The function g is a non-linear activation function, which can be either ReLU or Tanh. $g(W_H y + b_H)$ is responsible for modifying the input data, allowing for the network to adaptively choose the extent of the transformation applied to the input data, thereby enhancing the network’s expressive power and adaptability. t represents the transformation gate, which determines the amount of information to be transmitted to the next step or bypassed entirely. It serves as a control mechanism for regulating the flow of information and adjusting the relative contribution of the input data. The structure of it is shown in Figure 3 below.

Contextual Embedding: LSTM (Long Short-Term Memory) is a variant of recurrent neural networks (RNNs) that plays a crucial role in contextual encoding. It is capable of modeling sequential dependencies, storing and transmitting contextual information, handling variable-length sequences, and providing rich representational capacity. Therefore, we employ LSTM to encode the contextual information of entity mentions in organization names. To take into account the position of entities in the context, we utilize two LSTMs to encode both the forward and backward directions and halt after encountering the entity word beyond the context: $h_e = [LSTM_{\vec{t}_e}, LSTM_{\overleftarrow{t}_e}]$, where t_e represents the positional index of entity e in the context and $h_e \in R^{1 \times d_{HE}}$, and $d_{HE} = forward_hidden_size + backward_hidden_size$.

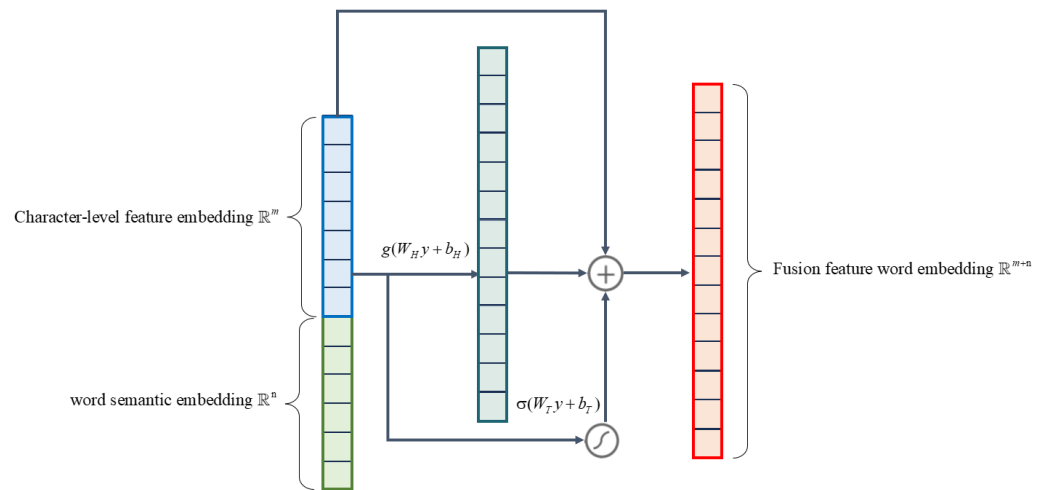


Figure 3. The structure of the Highway network.

3.4. The Multi-Context Fusion Layer Based on Bidirectional Matching

For two institutional entities to be confirmed as the same entity H and G , the context can be expressed as $H = \{h_1, h_2, h_3, \dots, h_p\}$ and $G = \{g_1, g_2, g_3, \dots, g_q\}$; p and q are the number of contexts. To determine whether Entity H and Entity G refer to the same entity, we go beyond considering a single context and instead consider multiple contexts. We evaluate the similarity between Entity H and Entity G by comprehensively considering the information from multiple contexts, H and G .

The influence of different contexts in determining whether two entities refer to the same entity may vary [26]. For a given context h_p , we calculate the influence weight score of $a_p = \max(\text{sim}(h_p, G))$, where $\text{sim}(h_p, G)$ represents the similarity between h_p and the q contexts of Entity G , and \max selects the highest similarity value. The underlying idea is that for institution entities, there might be multiple address information associated with the same entity. However, the matching between two institution entities is often dominated by the most matching addresses between them. An influential context, represented by h_p , is likely to be highly similar to one of the addresses and less similar to the rest of the addresses. Therefore, the influence of a context on the similarity weight between the two entities should be determined by the context that is most similar to the corresponding address in the other entity.

For each h_p in H and g_q in G , the matching score a_p and a_q is calculated from:

$$S_{HG} = HG^T \tag{4}$$

The matching score matrix S can be obtained by taking softmax on the S_{HG} over a certain axis (over 0-axis for $S_{H \rightarrow G}$ and 1-axis for $S_{H \leftarrow G}$). For each piece of encoded context, say h_p for the entity H , we use the highest matched score with its counterpart as the relative informativeness score of h_p to H :

$$a_p = \max(S_{p \rightarrow q} \mid q \in Q) \tag{5}$$

We further aggregate multiple pieces of encoded contexts for each entity to a global context based on the relative informativeness scores:

$$\bar{h} = \sum_{p \in P} a_p h_p, \bar{g} = \sum_{q \in Q} a_q g_q \tag{6}$$

\bar{h} and \bar{g} are the final context embeddings for entities H and G .

3.5. Training Objectives

Our training objective is to enable the model to identify whether two given entity names belonging to different institutions refer to the same entity. To accomplish this objective, we utilize the Siamese loss function:

$$L_{Siamese} = yL_+(e, k) + (1 - y)L_-(e, k) \quad (7)$$

Y represents the label value, which includes two cases for the loss function: $L_+(e, k)$ when entities H and G are synonymous institution entities, and $L_-(e, k)$ when entities H and G are not synonymous institution entities.

$$L_+(e, k) = (1 - s(\bar{h}, \bar{g}))^2 \quad (8)$$

$$L_-(e, k) = \max(s(\bar{h}, \bar{g}) - m, 0)^2 \quad (9)$$

$s(\cdot)$ is a similarity function, such as the cosine similarity, and m is the margin value that represents the desired minimum distance between dissimilar input pairs. $L_+(e, k)$ is within the range $[0, 1]$, where higher similarity scores correspond to lower values. For the loss $L_-(e, k)$, when $s(\bar{h}, \bar{g})$ is less than the margin value m , it remains zero; otherwise, it increases with the increase of $s(\bar{h}, \bar{g})$.

4. Experiments

4.1. Evaluation Metrics

To evaluate the performance of our method, we adopted the precision, recall, and F1 score as the evaluation metrics. The accuracy, precision, recall, and F_1 - score are calculated using the formulas as shown below:

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

$$F_1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (12)$$

TP (True Positive) refers to the number of positive instances correctly predicted as positive. FN (False Negative) refers to the number of positive instances incorrectly predicted as negative. FP (False Positive) refers to the number of negative instances incorrectly predicted as positive.

4.2. Datasets

We employed entity linking techniques to link entities in Wikidata with entities in the Web of Science (WOS) dataset for institutions with a publication count greater than 1000 [27]. The official name of an institution was used as the anchor sample, while the most frequent alias appearing in WOS was considered as the positive sample. Additionally, we selected institutions with the most similar names but representing different entities as the negative sample. Among the 5902 institutions in WOS with a publication count greater than 1000, we successfully linked 3572 institution entities to the institutional knowledge base. Out of these, 1494 institution entities had aliases linked to the knowledge base. Eventually, we obtained 1494 positive and negative pairs, from which we selected 2800 pairs as the final dataset for institution synonym relationships, as shown in Table 3.

Table 3. Datasets.

Count	Anchor Sample	Positive Sample	Negative Sample
Entity	1400	1400	1400
Context	7,586,714	861,632	3,577,596
Vocab	217,436	58,129	108,021

4.3. Baselines

To compare the performance of our proposed method, we selected four other methods as benchmark approaches. The first two methods are classical approaches for institution synonym recognition, while the latter two are classical models for synonym recognition in general.

1. Huang’s Method [5]: This method is considered representative in rule-based institution synonym recognition due to its emphasis on knowledge and rule completeness and generality. In the following sections, we refer to this method as “Huang’s method” for simplicity.
2. Word2vec [28]: This method is commonly used in deep learning-based institution synonym recognition and serves as a baseline model in our comparison.
3. SRN [29]: SRN is a character-level model that encodes entities as a sequence of characters using BiLSTM. The hidden states are averaged to obtain an entity representation, and cosine similarity is used in the training objective.
4. MaLSTM [30]: MaLSTM is a word-level model that takes word sequences as input. Unlike SRN, which uses BiLSTM, MaLSTM employs unidirectional LSTM and utilizes the Euclidean norm to measure the distance between two entities.

4.4. Results

In order to demonstrate the superiority and effectiveness of the proposed model in institution synonym recognition, we conducted comparative experiments and ablation experiments with the four aforementioned models on our custom dataset. The results of these experiments are presented in Table 4, as shown below.

Table 4. Experiment result.

Methods	Precision	Recall	F1
Huang’s method	69.53	77.20	73.17
Word2vec	56.34	92.20	69.95
SRN	46.07	60.29	52.23
MaLSTM	57.72	52.20	54.82
Ours	77.55	88.37	82.61
No Highway	73.88	89.92	81.12
No Char-CNN	74.65	82.17	78.23
No Word2vec	55.67	88.60	68.07
No bidirectional matching	64.60	80.62	71.72

From the upper part of Table 4, we can see that our models consistently outperform the baseline in accuracy and recall and are lower than Word2vec in terms of recall. SRN had the worst overall performance. Our model is 9.44% better in F1 than the best baseline model. To study the contribution of different modules of our model for synonym discovery, we also report the ablation test results in the lower part of Table 4. The Highway contributes 1.49% improvement in F1, Char-CNN contributes 4.38% improvement in F1, Word2vec contributes 14.54% improvement in F1, and bidirectional matching contributes 10.89% improvement in F1.

4.4.1. Results Analysis

From the experimental results, it can be observed that deep learning models based on single-context matching performed poorly in this data environment. This can be attributed to two main issues:

1. Dependency on a single context: These models rely solely on single-context information, which makes them susceptible to absorbing excessive noise during the learning process and limits their ability to fully utilize additional information provided by other relevant contexts. This approach struggles to effectively differentiate between complex scenarios with multiple similar institution names.
2. Emphasis on sentence encoding: These models tend to use the encoding of the entire sentence as the final embedding output, without specifically highlighting the importance of the institution entity itself. For institution synonym recognition, the focus should be on the specific encoding of the institution entity rather than generic information from the entire sentence.

Furthermore, while the Word2vec model demonstrates high recall in such tasks, its precision is limited. This may be because it tends to generalize semantically similar institutions (such as similar departments in different universities) into the same category, leading to a lack of precision.

Comparing the results of the models without Char-CNN and without Word2vec, using word-level features alone outperforms using character-level features alone. However, combining both types of features yields even better results, as it can identify some character-level misspellings. Compared to directly concatenating Char-CNN and Word2vec, using a Highway network has a better effect, indicating that this component positively influences the model's performance. The Highway network better integrates the two types of features, providing the model with better feature representation capabilities. Comparing the model without the bidirectional matching layer and the complete model, the bidirectional matching method effectively utilizes the importance of different contexts in institution name matching. It can leverage information from multiple contexts to enhance the overall performance of the model, significantly improving institution synonym recognition.

4.4.2. Error Analysis

An error analysis is critical for understanding the model shortcomings, thereby contributing to the in-depth analysis of and improvement in the model. We analyzed the data and then observed the error types and causes of errors. The error analysis results are shown in Table 5.

As mentioned in Section 3, deep learning-based models perform well overall, but there are still some problems, and the follow-up work will focus on the above three aspects. In addition to that, because we choose institutions with over 1000 publications, the number of institutional contexts in the dataset is higher than in the database. This is shown in Figure 4; the higher the number of contexts, the better the model performance, so the actual performance of the model may be slightly lower than estimated.

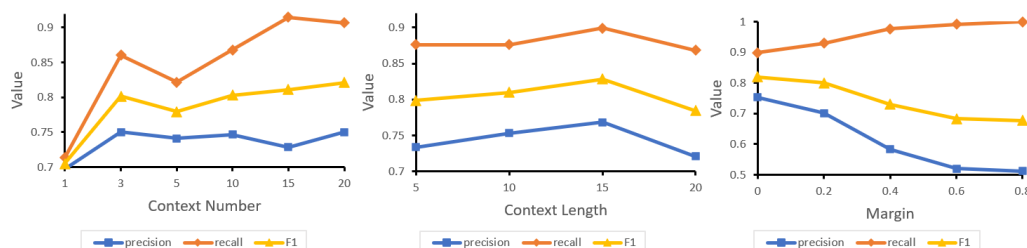


Figure 4. Sensitivity analysis.

Table 5. Error analysis results.

Error Type	Reason	Improvement Direction
Different institutions that are similar both literally and semantically.	The two institutions Universtil Ben Turin and Universtil Ben Turku have similar names and have a large number of the same sub-institutions, such as Compters, Czes, Depatment, and Deputklinxczens.	Introduce better features, such as author name features and geographic attribute features.
Ambiguous institution aliases in authority files.	The alias of “Perbright Institute” is “Institute Animal Health”, which is the same as the aliases of several secondary institutions, resulting in a mismatch of models.	Construct hierarchical relationships and identify synonymous institutions from top to bottom to avoid identical or similar aliases.
The choice of institutional context is not reasonable.	Contextual selection methods are not perfect enough, which affects model matching.	Further refine the contextual selection method.

4.4.3. Hyperparameters

To investigate the impact of different hyperparameters on the experimental results, we trained the proposed model using the following parameter configurations, as shown in Table 6. We varied the number of randomly sampled contexts per entity from 1 to 20 and the maximum context length from 5 to 20. For Char-CNN, we changed the number and size of the convolutional filters. The margin value (m) in the loss function was varied from 0 to 0.8. We also experimented with different optimizers during training.

Table 6. Experiment result.

Hyperparameter	VALUE
epoch	10
lstm_size	512
word_size	200
context_number	20, 15, 10, 5, 3, 1
context length	20, 15, 10, 5
filter_size	1, 2, 3, 4, 5, 6
filter_amount	100, 200, 300
m (margin)	0, 0.2, 0.4, 0.6, 0.8
optimizer	<i>Adam, RMSProp, Adadelta, Adagrad</i>
loss function	Siamese loss
batch size	2
learning rate	2×10^{-5} , 1×10^{-5} , 1×10^{-4} , 5×10^{-5}

Figure 4 depicts the overall trend of the F1 score increasing as the number of contexts increases, indicating that the model generally performs better with more context information. This aligns with expectations, as having more context information allows for better differentiation between two institution names as the same or different entities. However, it can be observed from the graph that all metrics decrease when the number of contexts is five. Upon analyzing the data, this is mainly attributed to the imbalance in the number of contexts for institution names. Some institutions have insufficient contexts to meet the specified number, causing the model to overly rely on features from institution names with an adequate number of contexts and neglect those with fewer contexts. When the maximum context length is set to 15, the model achieves the best F1 score. This is because longer contexts may introduce noise, while shorter contexts may provide less information.

As the margin value (m) increases, the F1 score, precision, and recall all show a decreasing trend. This indicates that learning from negative examples is more important for institution synonym recognition compared to positive examples.

5. Conclusions

To address the limitations of existing matching models in the domain of institution synonym recognition, a novel institution synonym recognition model was proposed, incorporating multiple feature dimensions. However, it is worth noting that the institution synonym recognition model has certain limitations. For instance, its performance may be influenced by context length and context number, and further investigation is needed to assess its effectiveness in such scenarios. Despite these limitations, our model has shown promising results in significantly improving institution synonym recognition performance and addressing the shortcomings in the related research in the field of deep learning. In the future, our work will focus on exploring the interpretability of deep learning models, the construction of datasets from different databases, and the practical impact of improvements in the field of academic evaluation.

Author Contributions: Methodology, Z.F.; Software, Z.J.; Funding acquisition, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the “Pioneer” and “Lingyan” R&D Projects of Department of Science and Technology of Zhejiang Province (2022C01220).

Data Availability Statement: Data available on request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zeini, N.T.; Okasha, A.E.; Soliman, A.S. A review on social segregation research: Insights from bibliometric analysis. *Kybernetes* **2023**, *1*, 1–8. [\[CrossRef\]](#)
2. Kaur, A.; Bhatia, M. Scientometric Analysis of Smart Learning. *IEEE Trans. Eng. Manag.* **2021**, *71*, 400–413. [\[CrossRef\]](#)
3. Dodevska, Z. An Expanded Bibliometric Study of Articles on Emerging Markets. *Management* **2022**, *29*, 11–20. [\[CrossRef\]](#)
4. Laura Cervi, N.S.; Calvo, S.T. Analysis of Journalism and Communication Studies in Europe’s Top Ranked Universities: Competencies, Aims and Courses. *J. Pract.* **2021**, *15*, 1033–1053. [\[CrossRef\]](#)
5. Huang, S.; Yang, B.; Yan, S.; Rousseau, R. Institution name disambiguation for research assessment. *Scientometrics* **2014**, *99*, 823–838. [\[CrossRef\]](#)
6. Zhang, S.; Xinhua, E.; Pan, T. A multi-level author name disambiguation algorithm. *IEEE Access* **2019**, *7*, 104250–104257. [\[CrossRef\]](#)
7. Ding, X.; Zhang, H.; Guo, X. An unsupervised framework for author-paper linking in bibliographic retrieval system. In Proceedings of the 2018 14th International Conference on Semantics, Knowledge and Grids (SKG), Guangzhou, China, 12–14 September 2018; pp. 152–159.
8. Falahati Qadimi Fumani, M.R.; Goltaji, M.; Parto, P. Inconsistent transliteration of Iranian university names: A hazard to Iran’s ranking in ISI Web of Science. *Scientometrics* **2013**, *95*, 371–384. [\[CrossRef\]](#)
9. French, J.C.; Powell, A.L.; Schulman, E. Using clustering strategies for creating authority files. *J. Am. Soc. Inf. Sci.* **2000**, *51*, 774–786. [\[CrossRef\]](#)
10. French, J.C.; Powell, A.L.; Schulman, E.; Pfaltz, J.L. Automating the construction of authority files in digital libraries: A case study. In Proceedings of the Research and Advanced Technology for Digital Libraries: First European Conference, ECDL’97, Pisa, Italy, 1–3 September 1997; 1997 Proceedings 1; Springer: Berlin/Heidelberg, Germany, 1997; pp. 55–71.
11. Jonnalagadda, S.; Topham, P. NEMO: Extraction and normalization of organization names from PubMed affiliation strings. *J. Biomed. Discov. Collab.* **2010**, *5*, 50. [\[CrossRef\]](#)
12. Backes, T.; Hienert, D.; Dietze, S. Towards hierarchical affiliation resolution: Framework, baselines, dataset. *Int. J. Digit. Libr.* **2022**, *23*, 267–288. [\[CrossRef\]](#)
13. Backes, T.; Dietze, S. Connected Components for Scaling Partial-order Blocking to Billion Entities. *ACM J. Data Inf. Qual.* **2024**, *16*, 1–29. [\[CrossRef\]](#)
14. Jacob, F.; Javed, F.; Zhao, M.; Mcnair, M. sCool: A system for academic institution name normalization. In Proceedings of the 2014 International Conference on Collaboration Technologies and Systems (CTS), Minneapolis, MN, USA, 19–23 May 2014; pp. 86–93.
15. Kronman, U.; Gunnarsson, M.; Karlsson, S. *The Bibliometric Database at the Swedish Research Council—Contents, Methods and Indicators*; Swedish Research Council: Stockholm, Sweden, 2010.

16. Jiang, Y.; Zheng, H.T.; Wang, X.; Lu, B.; Wu, K. Affiliation disambiguation for constructing semantic digital libraries. *J. Am. Soc. Inf. Sci. Technol.* **2011**, *62*, 1029–1041. [[CrossRef](#)]
17. Abramo, G.; Cicero, T.; D'Angelo, C.A. A field-standardized application of DEA to national-scale research assessment of universities. *J. Inf.* **2011**, *5*, 618–628. [[CrossRef](#)]
18. Huang, Y.; Li, J.; Sun, T.; Xian, G. Institution information specification and correlation based on institutional PIDs and IND tool. *Scientometrics* **2020**, *122*, 381–396. [[CrossRef](#)]
19. Wang, L.; Hu, J.; Wang, Q.; Yang, Y.; Lou, P.; Fang, A. Big Open Data Aided Institutions' Name Normalization and Attribute Enrichment. In Proceedings of the 2022 3rd Information Communication Technologies Conference (ICTC), Nanjing, China, 6–8 May 2022; pp. 173–177.
20. Shao, Z.; Cao, X.; Yuan, S.; Wang, Y. ELAD: An entity linking based affiliation disambiguation framework. *IEEE Access* **2020**, *8*, 70519–70526. [[CrossRef](#)]
21. Chen, Y.; Li, X.; Li, A.; Li, Y.; Yang, X.; Lin, Z.; Yu, S.; Tang, X. A Deep Learning Model for the Normalization of Institution Names by Multisource Literature Feature Fusion: Algorithm Development Study. *JMIR Form. Res.* **2023**, *7*, e47434. [[CrossRef](#)] [[PubMed](#)]
22. Onodera, N.; Iwasawa, M.; Midorikawa, N.; Yoshikane, F.; Amano, K.; Ootani, Y.; Kodama, T.; Kiyama, Y.; Tsunoda, H.; Yamazaki, S. A method for eliminating articles by homonymous authors from the large number of articles retrieved by author search. *J. Am. Soc. Inf. Sci. Technol.* **2011**, *62*, 677–690. [[CrossRef](#)]
23. Cuxac, P.; Lamirel, J.C.; Bonvallot, V. Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics* **2013**, *97*, 47–58. [[CrossRef](#)]
24. Sun, Y. Research on SCI Address Field Data Cleaning Method Based on Word2Vec. *J. Intell.* **2019**, *38*, 195–200.
25. Kim, Y.; Jernite, Y.; Sontag, D.; Rush, A. Character-aware neural language models. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
26. Zhang, C.; Li, Y.; Du, N.; Fan, W.; Yu, P.S. Entity synonym discovery via multipiece bilateral context matching. *arXiv* **2018**, arXiv:1901.00056.
27. Zhang, J.; Cao, Y.; Hou, L.; Li, J.Z.; Zheng, H. XLink: An Unsupervised Bilingual Entity Linking System. In Proceedings of the China National Conference on Chinese Computational Linguistics, Nanjing, China, 13–15 October 2017.
28. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
29. Neculoiu, P.; Versteegh, M.; Rotaru, M. Learning text similarity with siamese recurrent networks. In Proceedings of the 1st Workshop on Representation Learning for NLP, Berlin, Germany, 11 August 2016; pp. 148–157.
30. Mueller, J.; Thyagarajan, A. Siamese recurrent architectures for learning sentence similarity. In Proceedings of the AAAI conference on artificial intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.