BSEN: A BiSiamese Entity Normalization Method for Biomedicine

Zirong Peng¹, Qinyong Yu, Hui Yang, Yongli Wang*¹
¹Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China
²CETC Big Data Research Institute Co., Ltd, Guiyang, Guizhou 550022, China
³Big Data Application on improving Government Governance Capabilities National Engineering Laboratory, Guiyang, Guizhou 550022, China
118106021968@njust.edu.cn, yongliwang@njust.edu.cn

Abstract. Normalization of named entities in the field of biomedicine is an important task in biomedical text data mining. Compared with other tasks in biomedical text mining research, there are relatively few researches on entities normalization. In this article, a BiSiamese entity normalization method for biomedicine (BSEN) is proposed. Firstly, the text similarity algorithm is analyzed, and an improved similarity measurement algorithm for biomedical inverse text frequency and cosine (BIC) is proposed. Secondly, the data set is trained in pairs using BiSiamese network and combined with BIC to calculate text similarity. The entity corresponding to the maximum similarity calculated in the normalization knowledge base is the normalized result obtained by the BSEN method. The verification experiments on the verification data set show that the BSEN has achieved better normalization results than the existing methods.

1. Introduction
Named entity recognition (NER) and named entity normalization (NEN) in the field of biomedicine have long been important tasks in biomedical text data mining. The problem to be solved by entity normalization is to collect different expressions of the same entity name, establish the correspondence between the canonical name and the variant name, and achieve the purpose of entity identification by giving the entity a unique identifier. Compared with existing biomedical entity recognition tasks, there are few researches on entities normalization. The normalization of named entities is very challenging. To achieve the entity normalization of biomedical related texts, the names of the same normalized entity and the non-normalized names of its different expressions can be connected to improve the efficiency and accuracy of query retrieval. Assigning a unique identifier to a normalized entity is conducive to establishing operations with other systems and sharing resources. The contribution of this paper is as following, 1) an improved text similarity measurement algorithm based on the combination of inverse biomedical text frequency and cosine is proposed, 2) an entity normalization method-BSEN of biomedical text based on BiSiamese network and BIC text similarity measurement algorithm is proposed.

2. Related Work
At present, the research on the entity standardization of biomedicine mainly includes rule-based, dictionary-based pattern matching methods and machine learning-based methods. Most normalization
research relies on the method based on rules and medical dictionary matching. In a series of tasks in the BioCreative competition, the concept normalization problem has played a huge role in genes and proteins[1]. One of the sub-tasks in BioCreative V is the identification and normalization of named entities about disease names[2]. The OSCAR system[3] standardized a variety of chemical corpora, aiming at data mining of chemical substance texts. The ProMiner system[4] uses a dictionary-based approach to approximate string pattern matching to identify and standardize gene and protein names. The system uses pre-processed dictionaries, including biological entities with known synonyms. The MetaMap system[5] maps biological entities to the concept identifiers in the Unified Medical Language System (UMLS)[6], and aims to improve the retrieval of MEDLINE-related citations. The UWM system[7] uses the CRFs algorithm to extract disorder entities from clinical data, and uses the edit distance to automatically learn the change pattern of clinical terminology from the unified medical language system and task training set and uses the edit distance to automatically extract from the training set and unified medical Learn the changing patterns of clinical terms in the language system, and then normalize. In the research of machine learning based on the standardization of biomedical entities, DNorm[8] proposed by Leaman et al. adopts the method of paired learning training. Using the pipeline serial execution architecture, in the NCBI disease data set[9] has achieved good results with an accuracy rate of 82.8%, but it is easy to cause cascading errors between components. So far, TaggerOne[10] is one of the best performing machine learning-based systems on the NCBI dataset. The model uses a semi-Markov model for biomedical entity identification and standardization during the prediction period. The above two methods are combined with NER to achieve entity normalization.

In the field of named entity normalization, rule-based and dictionary-based pattern matching methods rely more on the accuracy of rules and dictionaries, and it is particularly important to choose an appropriate text similarity algorithm. The calculation of text similarity[11] is a very complicated problem. At present, there are three main methods for calculating text similarity: matching-based methods, word vector-based methods, and neural network-based methods. The method based on word matching[12] is the most classic and intuitive method, which measures the similarity of texts by comparing the frequency of the same words appearing in two texts. The word vector-based method[13] maps each word to a fixed-length vector and treats each word as a point in a high-dimensional space, so the distance between the vectors can be used to characterize the similarity between words. But the above two methods do not take into account the semantic information between the contexts.

3. BSEN Method
This paper proposes a biomedical text entity normalization method called BSEN (A BiSiamese Entity Normalization method for Biomedicine). The execution process of BSEN method is shown in Figure 1. In order to obtain the bidirectional semantic information and vector representation of the input text, the method is based on the Siamese network[14] fused with BiLSTM to train the data set in pairs, and then combined with the BIC algorithm to calculate the text similarity. The entity corresponding to the maximum similarity calculated in the knowledge base is the normalized result of the input entity. Finally, determine whether the normalized result is correct.
3.1. Symbol Definition

In this section, according to the BSEN method, the symbols and meanings used therein are given.

**Definition 1** BiomedicalTextSet: A collection of biomedical texts composed of network text content obtained from the network that conforms to the theme of biomedicine.

\[
\text{BiomedicalTextSet} = \{bts_1, bts_2, ..., bts_i, ..., bts_n\}, 0 \leq i \leq n
\]  

**Definition 2** BE: A sequence of word vectors converted into sentences by the BSEN method.

\[
\text{BE} = (be_1, be_2, ..., be_i, ..., be_n), 0 \leq i \leq n
\]  

**Definition 3** PairTraining: In the training process, the training of the data set uses a one-to-one model. A normalized entity and a normalized entity alias are paired for training.

**Definition 4** NormalizationEntityList: The normalization entities in BiomedicalTextSet.

\[
\text{NormalizationEntityList} = \{ne_1, ne_2, ..., ne_i, ..., ne_n\}, 0 \leq i \leq n
\]  

3.2. Text Similarity Algorithm

Due to the particularity of biomedical texts, the importance of core words in the text needs to be considered. The core words in the text play a greater role in matching, and the frequency of inverse text is to measure the importance of a word. This paper presents an improved similarity measurement algorithm for biomedical inverse text frequency and cosine, called BIC. The word idf value is added as a weight to the calculation of the word idf and sentence vector formula as follows (4) (5):

\[
\text{idf}(w) = \log \left( \frac{D}{D_w + 0.5} \right)
\]  

\[
\text{vector}(x) = \sum_{i=1}^{n} v(w_i) \cdot \text{idf}(w_i)
\]  

Where D is the total number of words in the text, \(D_w\) is the number of words w, \(v(w_i)\) is the vector of the word \(w_i\) of the sentence, and \(\text{idf}(w_i)\) is the idf value after \(w_i\) weighting. After obtaining the vector representation of the two sentences in the above manner, the word vector is accumulated horizontally. The similarity \(E(X_1, X_2)\) of the two sentences is obtained by combining the cosine similarity as formula (6). Assuming that the two n-dimensional vectors obtained from the inverse text frequency are \(X_1 = (x_1, x_2, ..., x_n)\) and \(X_2 = (y_1, y_2, ..., y_n)\), then \(E(X_1, X_2)\) is calculated as follows:

\[
E(X_1, X_2) = \frac{\sum_{i=1}^{n} (x_i \cdot y_i)}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i)^2}}
\]
3.3. **BiSiamese Network**

The Siamese network is essentially a method of similarity measurement. This method is suitable for the case where the number of samples in each category is small but the number of categories is large. It can be used for category identification and classification. The traditional classification method needs to know which category each sample belongs to and is suitable for the case where the number of samples in each category is large but the number of categories is small. The Siamese network makes up for the shortcomings of traditional classification methods and can learn a similarity measure from the data, which use this metric to compare and match with samples of new unknown categories, thereby being able to distinguish untrained categories. In this paper, it is designed to integrate BiLSTM on the basis of the Siamese network to obtain the context information of the text, called BiSiamese network. The network structure is shown in Figure 2.

The BiLSTM in the BiSiamese network consists of two LSTM structures with opposite directions which is composed of two opposite LSTM structures. By concatenating the outputs \( h_{i1} \) and \( h_{i2} \) of the forward and reverse hidden layers to obtain the final output representation, such as formula (7), the semantic information of the current vocabulary can be understood to the greatest extent.

\[
H_i = [h_{i1} \oplus h_{i2}]
\]  

This article applies the BiSiamese network to the entity normalization of biomedical texts. The specific steps for normalization are as follows:

**Step 1** Input BiomedicalTextSet in pairs.

**Step 2** Use character embedding to perform vector characterization to obtain a sequence of context word vectors \( BE=(be_1, be_2, ..., be_i, ..., be_n) \).

**Step 3** Input \( BE \) into BiLSTM, and BiLSTM outputs hidden state sequence \( H=(H_1, H_2, ..., H_i, ..., H_n) \).

**Step 4** The output of \( H \) at each moment is averaged and input into the 128-dimensional fully connected layer.

**Step 5** Using the BIC algorithm to calculate the similarity, the entity corresponding to the maximum similarity calculated in the NormalizationEntityList is the normalized result of the BSEN method.

**Step 6** Determine whether the output result is the correct normalized entity.

3.4. **Loss Function**

The loss function used in this article is Contrastive Loss, which mainly considers the paired features of the data set in the BSEN method, and the contrast loss can be well expressed as the degree of matching of the samples, and can also be applied to the model of feature extraction. The loss functions for similar and dissimilar cases are given by equations (8), (9) and (10):
\[ L_s(x_1, x_2) = \frac{1}{4} (1 - E)^2 \quad (8) \]

\[ L(x_1, x_2) = E^2, \text{ if } (E > m) \quad (9) \]

\[ L_s(x_1, x_2) = 0, \text{ otherwise } \quad (10) \]

Where \( E \) is the similarity \( E(X_1, X_2) \) obtained by the BIC algorithm. For similar cases, the Loss function decreases monotonically with the increase of \( E \). For dissimilar cases and \( E \) is greater than the set threshold \( m \), the Loss function increases monotonically with the positive \( E \).

4. Experiment

4.1. Experimental Environment and Parameters

The following Table 1 give the experimental environment and parameters.

| Table 1. Experimental environment and parameters |
|-----------------------------------------------|
| OS            | Windows 10 |
| RAM           | 16GB       |
| Python        | 3.6        |
| Tensorflow    | 1.14       |
| embedding_dim | 150        |
| drop_out      | 0.8        |
| hidden_unit   | 50         |
| batch_size    | 64         |
| epoch         | 100        |

4.2. Experimental Data Set

The data set of the experiment in this paper is mainly composed of a biomedical Chinese text corpus crawled by the web, and the data set is manually annotated and data cleaned. The data contains a total of 17,000 items, of which 10,000 are similar to the comparison test data and the entity normalization experiment. A total of 7000 data (2000 training set, 2000 test set, 3000 verification set).

4.3. Experiment Procedure

First, the text similarity algorithm is analyzed. The data set used in the experiment is 10,000 sentence texts related to biomedicine in the network. There are three main comparison algorithms, Consine, Jaccard and BIC algorithm. The algorithm gives five sets of the same data, each group of 10 sentences, and calculates the average value as the result. The results are shown in Table 2. The main metrics are the calculation time of each algorithm, and the highest similarity and lowest similarity among the five results degree.

| Table 2. Experimental results of three similarity algorithms |
|------------------------------------------------------------|
| Algorithm | Calculating Time/sec | Max | Min |
|-----------|----------------------|-----|-----|
| Consine   | 24.5                 | 0.998 | 0.870 |
| Jaccard   | 0.5                  | 0.932 | 0.404 |
| BIC       | 25.2                 | 0.995 | 0.905 |

From the experimental results in Table 2, it can be seen that Jaccard has the lowest computational complexity and the simplest method, but it cannot process the semantic information of the sentence. Because the characteristic attributes of individuals are identified by symbol measures or boolean values, which is impossible to measure the size of the difference. It is more suitable for calculating the
similarity between individuals of symbol metric or boolean metric; Consine method is the most commonly used algorithm, but the result is very dependent on the result of word2vec training. Although the computational complexity of BIC method is relatively large. But the effect is more accurate than the results of the other two methods. Because of the special nature of the text to the biomedical corpus, the importance of the core words in the text needs to be considered. Therefore, in the process of normalization, this paper combines the BIC method to calculate the similarity.

In the entity normalization experiment, since the Siamese network contains two branches, the training method uses paired training. The data set contains three columns, the first column is the standardized entity name, the second column is the entity alias, and the third column is a canonical identifier. Since the entity name belongs to short text, the Siamese network uses character-based embedding to identify the similarity of structure and syntax, and uses four layers of BiLSTM with 64 hidden units in the network, where the input text feature vector Represents the average value output at each moment, and finally accesses a 128-dimensional fully connected layer, and judges whether it is correctly normalized by comparing the similarity between the entity and the normalized entity. The comparison method in this article is TF-IDF+Kmeans, and the experimental results are shown in the Figure 3 that the main evaluation index is accuracy.

![Figure 3. Comparison with TF-IDF+Kmeans algorithm accuracy](image)

It can be seen from the experimental results in Figure 3 that the accuracy of the two methods tends to be stable when the number of iterations is close to 100. Finally, the accuracy of the method in this paper has reached 88.7%. Compared with the TF-IDF+Kmeans method, the method in this paper has a certain improvement in the effect. The main reason may be that the cluster-based method is more applicable when the sample size of each category is large, but the number of categories is small. And the method in this paper is more suitable for the case where the number of samples in each category is small but the number of categories is large.

### 4.4. Experimental Effect
For the entity normalization of biomedical texts, the normalization effect of using the BSEN method is shown in Table 3, which gives the normalization results and confidence of some data.

| Input                          | Result                              | Confidence |
|-------------------------------|-------------------------------------|------------|
| Thromboembolic vasculitis     | Vascular occlusive vasculitis       | 0.92       |
| Acute characteristic arthritis| Rheumatoid arthritis                | 0.90       |
| Corticosteroids               | Adrenocorticotropic hormones        | 0.88       |
| Cystic tumor of pancreas      | Cystic carcinoma of pancreas        | 0.95       |
| Pancreatoduodenectomy         | Duodenectomy                        | 0.89       |
5. Conclusion
This paper proposes a BiSiamese entity normalization method BSEN for biomedical text. Based on the BiSiamese network, the data set is trained in pairs, which can well handle the case where the number of samples in each category is small but the number of categories is large, and the bidirectional semantic information of the text is obtained. Then, the text similarity is calculated by the BIC algorithm. The calculated entity corresponding to the maximum similarity is the normalized result of the input entity, and finally determines whether the normalized result is correct. The experiment proves that the method in this paper has achieved better results. In the next step of research, we need to consider expanding the number of data sets, considering parameter optimization from all angles, and trying to further improve accuracy.

Acknowledgments
This article was supported by the National Natural Science Foundation of China (61941113), the special fund project of the basic scientific research business expenses of central universities (30916011328, 30918015103), the Nanjing Science and Technology Planning Project (201805036), the Open Fund Project of the National Engineering Laboratory of Big Data Application Technology to Improve Government Governance.

References
[1] Hirschman L, Colosimo M, Morgan A, et al. Overview of BioCreAtIvE task 1B: normalized gene lists[J]. BMC bioinformatics, 2005, 6 Suppl 1(Suppl 1): S11.
[2] Wang Q, S Abdul S, Almeida L, et al. Overview of the interactive task in BioCreative V[J]. Database, 2016, 2016: baw119.
[3] Jessop D M, Adams S E, Willighagen E L, et al. OSCAR4: a flexible architecture for chemical text-mining[J]. Journal of Cheminformatics, 2011, 3(1):41.
[4] Cho H, Choi W, Lee H. A method for named entity normalization in biomedical articles: application to diseases and plants[J]. Bmc Bioinformatics, 2017, 18(1):451.
[5] Aronson A, Lang F M. An overview of MetaMap: historical perspective and recent advances[J]. Journal of the American Medical Informatics Association, 2010, 17(3): 229. 236.
[6] Olivier B. The Unified Medical Language System (UMLS): integrating biomedical terminology[J]. Nucleic Acids Research(suppl_1): suppl_1.
[7] Ghiasvand O, Kate R. UWM: Disorder Mention Extraction from Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns[C]// SemEval 2014. 2014.
[8] Robert L, Islamaj Dogan Rezarta, Zhiyong L. DNorn: disease name normalization with pairwise learning to rank[J]. Bmc Bioinformatics, 2013, 29(22): 2909-2917.
[9] Leaman R, Lu Z. NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization[J]. Journal of Biomedical Informatics, 2014, 47:1.
[10] Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov Models[J], Bioinformatics, 2016.
[11] Hongni Dong, Jiang Wu, Xiaohui Zhao, etc. Study on the Calculation of Text Similarity Based on Key-sentence[C]// The International Conference on E-Business and E-Government, ICEE 2010, 7-9 May 2010, Guangzhou, China, Proceedings. IEEE, 2010.
[12] Yan Z. Word Segmentation Method Research Based on Chinese Text Classification[J]. computer engineering and applications, 2003.
[13] Yang Z T, Zheng J. Research on Chinese text classification based on Word2vec[C]// IEEE International Conference on Computer & Communications. IEEE, 2016.
[14] Chopra S, Hadsell R, Lecun Y. Learning a similarity metric discriminatively, with application to face verification[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005.