applied sciences

MDPI

Article

An Efficient Method for Biomedical Entity Linking Based on Inter- and Intra-Entity Attention

Mamatjan Abdurxit, Turdi Tohti * and Askar Hamdulla

College of Information Science and Engineering (School of Cyber Science and Engineering), Xinjiang University, Urumqi 830017, China; mamatjan@stu.xju.edu.cn (M.A.); askar@xju.edu.cn (A.H.)
* Correspondence: turdy@xju.edu.cn; Tel.: +86-139-9999-4696

Abstract: Biomedical entity linking is an important research problem for many downstream tasks, such as biomedical intelligent question answering, information retrieval, and information extraction. Biomedical entity linking is the task of mapping mentions in medical texts to standard entities in a given knowledge base. Recently, BERT-based models have achieved state-of-the-art results on the biomedical entity linking task. Although this type of method is effective, it brings challenges for fine-tuning and online services in practical industries due to a large number of model parameters and long inference time. In addition, due to the numerous surface variants of biomedical mentions, it is difficult for a single matching module to achieve good results. To address the challenge, we propose an efficient biomedical entity linking method that integrates inter- and intra-entity attention to better capture the information between medical entity mentions and candidate entities themselves and each other, and the model in this paper is more lightweight. Experimental results show that our method achieves competitive performance on two biomedical benchmark datasets, NCBI and ADR, with an accuracy rate of 91.28% and 93.13%, respectively. Moreover, it also achieves comparable or even better results compared to the BERT-based entity linking method while having far fewer model parameters and very high inference speed.

Keywords: biomedical entity linking; candidate generation; candidate ranking; self-attention; cross-attention

Citation: Abdurxit, M.; Tohti, T.; Hamdulla, A. An Efficient Method for Biomedical Entity Linking Based on Inter- and Intra-Entity Attention. *Appl. Sci.* **2022**, *12*, 3191. https://doi.org/10.3390/app12063191

Academic Editor: Valentino Santucci

Received: 9 March 2022
Accepted: 17 March 2022
Published: 21 March 2022

1. Introduction

Entity linking is the task of mapping mentions in a text to standard entities in a given knowledge base [1]. Entity linking is one of the most important parts of information extraction [2,3], especially in biomedical research and clinical applications, and it is also the bridge between mentions and knowledge graphs in the knowledge base intelligent question answering process [4]. The entity linking task remains challenging in that the same word or phrase can be used to refer to different entities, as well as the same entity can be referred to by different words or phrases. In the field of biomedical text processing, this task is more commonly referred to as biomedical entity normalization. Biomedical entity linking [5] maps biomedical mentions such as disease, drug, and procedure terms that appear in a document to standard terminology words in the knowledge base.

Mentions extracted from biomedical texts suffer from a number of problems, such as colloquial, diverse, and erroneous representations, and if these conceptual entities are utilized or stored without processing, they may have adverse effects on subsequent tasks. The particular challenge of biomedical entity linking is not ambiguity, i.e., a word usually refers to only one entity, but the challenge actually lies in the fact that the surface forms differ significantly due to abbreviations, morphological variations, synonyms, and different word orders [6]. For example, "alkaline phosphatase increased" is also written as "ALP increased", "cerebral ischaemia" is also referred to as "ischemic cerebrovascular conditions". The set of standard terms in the knowledge base is large, and the terms are still

similar in form or semantics, making them difficult to distinguish. Moreover, unlike the traditional knowledge bases DBpedia [7] and YAGO [8], which contain information such as entity descriptions and entity attributes, entity information has only one entity name. Furthermore, not all biomedical mentions can be mapped to a specific term. Therefore, determining whether a mention can be mapped to a concept in a given ontology is part of the biomedical entity linking task, and these make medical entity linking very difficult.

In view of the current problems, researchers have deliberately proposed entity linking methods [9,10] for biomedical entity linking. Most of the biomedical entity linking research [11–13] has focused on solving the problem of medical entity diversity. Currently, in the field of entity linking, deep learning shows its powerful advantages and is becoming the mainstream approach to studying biomedical entity linking. Recently, the BERT-based biomedical entity linking method [14] has achieved the best results on different biomedical benchmark datasets. While this type of approach is effective, it poses challenges for fine-tuning in real industry and online services due to a large number of model parameters and long inference times. For deep learning models, the training efficiency of the model is very important, and the efficiency mainly includes training time and model parameters. Despite the fact that there are scientific facilities with a lot of computer capabilities, many people still have limited access to large-scale computational capacity. As a result, it is critical to create a more scalable approach for biological entity linking.

Essentially, biomedical entity linking is a type of semantic matching task. In general, they are mainly divided into representational matching models [15–17] and interactive matching models [18–20]. It is difficult for a representational model to measure the contextual importance between two sentences because the representational model needs to encode the representation of the two sentences separately, which will lose the semantic focus. The disadvantage of the interactive model is that it ignores global information such as syntax, inter-sentence contrast, etc., and thus cannot carve out global matching information from local information. In addition, biomedical entities are mentioned in too many different ways, and it is difficult to obtain good results with a single matching model. Inspired by recent progress, combining the advantages of each of them, we propose an efficient biomedical entity linking method by jointly modeling the intra-entity and inter-entity relationships of mention and candidate in a unified deep model to better capture the information between medical mentions and candidate entities themselves and each other. In summary, our main contributions can be summarized as follows:

1. We propose an efficient network for biomedical entity linking by jointly modeling inter- and intra-entity relationships of biomedical mentions and candidates in a unified model.
2. A novel fusion framework with cross-attention and self-attention is proposed to better exploit not only the relationship within each entity but also the relationship between mentions and candidates.
3. We also designed a biomedical entity linking method based on BERT and pairwise ranking to compare with the lightweight method in this paper.
4. The experimental results demonstrate that the proposed method in this paper achieves fairly competitive performance on two biomedical benchmark datasets. Furthermore, it also achieves comparable or even better results compared to the BERT-based entity linking method while having far fewer model parameters and very high inference speed.

The rest of the article is structured as follows: Biomedical entity linking research will be briefly discussed in Section 2 of this paper. Our methodology for linking biomedical entities will be explained in Section 3, and the general structure and processing flow of each portion will be shown. In order to show the efficacy of our approach, we provide an in-depth analysis of the experimental results in Section 4. Section 5 summarizes the research and discusses possible future directions.

2. Related Work

In the field of biomedical entity linking, earlier studies used rule-based systems to capture string similarity between mentions and entity names. Kang et al. [9] proposed a natural language processing module with five rules to improve the normalization performance of disease terms in biomedical texts. D'Souza and Ng [21] proposed a manual rule-based multichannel filtering system by defining 10 rules with different priorities to measure the morphological similarity between mentions and candidate entities in a given knowledge base for entity linking, and this is the best rule-based system that has worked so far on the NCBI [22] dataset.

To avoid the inefficiency associated with manual rules, machine learning methods automatically learn the appropriate similarity metric between entity mentions and entity names from the training set. DNorm [11] proposed by Leaman et al. uses a vector space model to represent medical entity mentions and a similarity matrix to measure the similarity between a given medical entity mention and a standard entity, with good results on the NCBI disease dataset. Ghiasvand and Kate [23] automatically learned the edit distance pattern of 554 term variations between synonyms of all disease concepts in the Unified Medical Language System (UMLS) [24] and the edit distance between mentions in the training data and the corresponding concepts in the UMLS to perform entity linkage processing. TaggerOne [13] uses a semi-Markov model for biomedical entity identification and linkage and is by far the best machine learning-based system on the NCBI dataset. Xu et al. [10] also defined three features and used linear RankSVM [25] to group each positive ADR mentioned as an entry in MedDRA in the TAC2017 ADR Challenge [26] that achieved the best performance. However, these machine learning methods cannot use semantically relevant information to link entity mentions more accurately.

Deep learning methods are currently showing their strong advantages in the field of entity linking. Since recently, deep learning approaches based on pre-trained embeddings have been effectively applied to many Natural Language Processing tasks, such as word2vec [27] and Glove [28]. In the field of biomedical entity linking, Li et al. [6] proposed a Convolutional Neural Network (CNN)-based entity linking architecture that treats biomedical entity linking as a ranking problem, which exploits the semantic similarity modeling of CNNs between entity mentions and candidate entities, and this approach outperforms the traditional rule-based approach. However, this method only takes the final semantic vectors of mentions and candidate entities, which makes it hard to figure out how much information has been lost and how the information between mentions and candidate entities is not interacting.

In 2019, Wright et al. [29] came up with a deep learning model called NormCo that takes into account the semantics of entity mentions and the consistency of entity mention topics in a single text. The biomedical entity normalization task is accomplished by combining the morphological similarity between entity mentions and candidate entities and the semantic similarity between mentions and entities computed using the GRU model. Phane et al. [30] proposed a new framework for BNE that considers and encodes the similarity between contextual meaning, conceptual meaning, and synonyms during representation learning to learn biomedical names and robust representations of terms. In 2019, Ishan et al. [31] proposed a framework for medical entity linking based on Triplet Networks, which uses three samples to form a training group, and useful features are learned by comparing distances. [32] proposes a new paradigm for learning robust representations of biological names and phrases that takes contextual meaning, conceptual meaning, and synonym similarity into consideration throughout the representation learning process.

Traditional word embedding methods have a context-independent representation for each word. BERT (Bidirectional Encoder Representations from Transformers) pre-trained language model [33] addresses this problem by training deep bidirectional representations from unlabeled texts. Based on the BERT pre-training model architecture, the domain-specific language representation model BioBERT (BERT for Biomedical Text Mining) [34], which is pre-trained on large-scale biomedical texts and clinical notes, was introduced to

improve the performance of many biomedical and clinical natural language processing tasks. Recently, Ji, Wei, and Xu [14] considered biomedical entity linking as a sentence pair classification task and proposed an entity linking architecture by fine-tuning the BERT pre-training model and achieved the best results so far on different types of datasets in the field of biomedical entity linking. A problem with pre-trained models is that they are usually computationally expensive and inefficient in practice. To deal with the above issues, we propose an efficient method for biomedical entity linking based on Inter- and Intra-entity attention. Different from existing methods, the proposed biomedical entity linking model is able to exploit not only the intra-entity relationship within each entity, but also the inter-entity relationship between mention and candidate to enhance each other for mentions and candidates matching.

3. Method

Given the biomedical mentions recognized in the document and the knowledge base consisting of a set of concepts, the goal of the biomedical entity linking task is to link each mention to the correct medical entity in the knowledge base. If a mapping concept is not present in the knowledge base, then it is denoted by NIL as unlinkable. To solve this problem, given a training set that is already linked to the correct entity in the knowledge base, the biomedical entity linking approach in this paper consists of three steps:

1. **Preprocessing:** All entity mentions in the corpus and entity names in the knowledge base are preprocessed to unify the format.
2. **Candidate entity generation:** For each biomedical mention, a set of candidate entities is generated from the knowledge base.
3. **Candidate entity ranking:** For each mention, a candidate ranking model is used to score each pair of mention and candidate entity, and the result with the highest score is output.

3.1. Preprocessing

Abbreviation Resolution: As in previous work on biomedical entity linking, in this paper, we use the Ab3p (Biomedical Text Abbreviation Recognition Tool) toolkit [35] to extend biomedical abbreviations. The Ab3p tool identifies abbreviations in documents and returns a list of replacement terms with probability, and we use the replacement term with the highest probability. For example, Ab3p identifies that "pws" is an abbreviation for "Prader Willi syndrome," and we replace each entity abbreviation with its corresponding expanded term.

Numeric Synonyms Replacement: Biomedical entity names may contain different forms of numbers. Therefore, in this paper, a numeric dictionary was manually created and different forms of numbers in biomedical mentions and concepts were replaced with their corresponding Arabic numerals.

Other Preprocessing: In addition, all punctuation was removed, and all words were converted to lowercase letters.

3.2. Candidate Generation

In the candidate generation phase, for each mention M, the goal of the biomedical entity linking system is to filter out irrelevant entities from the standard knowledge base and generate the candidate entity set C_m, which contains the mention M with all possible standard entities e linked to it. The aim is to narrow down the scope of subsequent reordering links and thus improve overall efficiency.

The candidate entity generation method used in this paper calculates similarity scores for each pair of biomedical mentions in the corpus and entities in the knowledge base and returns the top entity with the highest score as the candidate set. In order to take advantage of the hidden features in biomedical mentions, two retrieval methods are designed. The first approach is to search directly for the closest standard terms to the biomedical entity mentions to be linked. The second way is to find the most similar entity mentions on

the annotated data to be linked. The goal is to find the most similar data to the "original surgical term" to be normalized on the annotated data and take the corresponding standard term as the candidate entity. In this paper, the top 20 standard entities were selected as the candidates by combining the above two search methods.

In this paper, we use an unsupervised alignment method [36], which calculates the cosine similarity between each word in an entity mention and the word embedding of each word in a given knowledge base entity to obtain a cosine similarity score matrix. For the words in each mention, the algorithm selects the most similar words in the text by maximum pooling. Each word is represented by a 200-dimensional word embedding trained from PubMed and the MIMIC-III corpus [37]. A given word $m_i \in M$ is mapped to the most similar word $c_j \in C$ by alignment cosine similarity and returns the cosine similarity score for that word. We calculate the similarity from two directions.

$$\text{alignSim}(m_i, C) = \max_{c_j \in C} \cos(m_i, c_j) \quad (1)$$

$$\text{alignSim}(c_j, M) = \max_{m_i \in M} \cos(c_j, m_i) \quad (2)$$

Then, the similarity scores of mentions and candidate entities are calculated as the sum of the alignment cosine similarity.

$$\text{Sim}(M, C) = \frac{1}{|M|+|C|} \left(\sum_{i=1}^{|M|} \text{alignSim}(m_i, C) + \sum_{j=1}^{|C|} \text{alignSim}(c_j, M) \right) \quad (3)$$

Finally, a candidate entity set is constructed, which contains the previous candidate entities mentioned by each entity and the similarity score of each candidate entity. We find that there are entities with a score equal to 1 in the set of candidate entities, and if there are candidate entities with a score equal to 1 in this set, we can filter the other candidate entities with a score less than 1. Then we use the candidate entity ranking model for the set of entities to output the final result.

3.3. Candidate Ranking

Given a mention *M* and its candidate entity set, the biomedical candidate ranking model calculates scores for each pair of mention and candidate *C*. In this section, we describe the details of this ranking model, which mainly consists of a representation layer, a BiGRU encoding layer, an intra-entity attention module, an inter-entity attention module, and a CNN aggregation layer. The overall architecture of the biomedical candidate ranking model proposed in this paper is shown in Figure 1. As shown in Figure 1, given a biomedical mention-candidate pair, the mention and candidate entity are first converted into a corresponding word vector by querying the word vector table. Then it is concatenated with the character-level features of each word obtained using CNN. The vectors are sent into the BiGRU layer for encoding. Based on these extracted fine-grained representations for mentions and candidates, we model the intra-entity relationship with the Self-Attention Module, and adopt the Cross-Attention Module to model the inter-entity relationships for mentions and candidates. Then the 1d-CNN and pool operation are employed to aggregate the two sequences of matching vectors into fix-length vectors. Finally, we use a two-layer fully connected neural network to compute the final score.

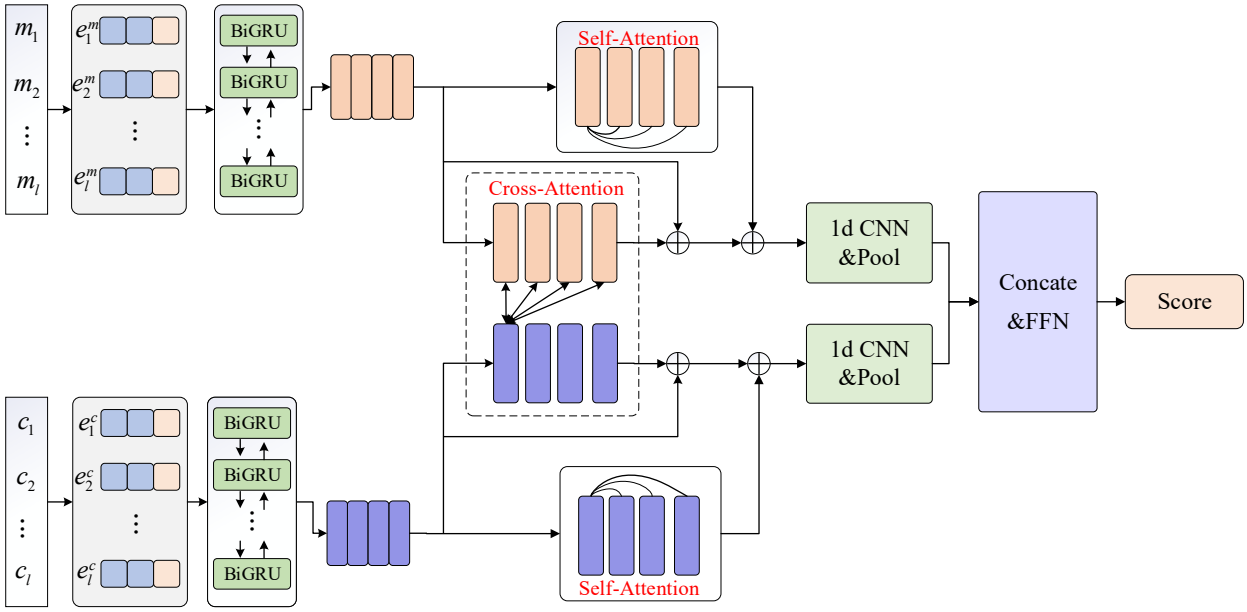

Figure 1. The architecture of our candidate ranking model.

3.3.1. Perfect Match

In the candidate generation stage, it is found that some candidate entities can completely match the standard entities; that is, the alignment cosine similarity is equal to 1. These entities are then linked directly into the knowledge base without being fed into the candidate ranking model. Then, the candidate ranking model is used to output the final result for entities with scores of less than 1 in the candidate set.

3.3.2. Representation Layer

The purpose of the representation layer is to vectorize the mentions and candidates expressed in natural language form. In this paper, the mentions and candidate entities are represented by the set of embeddings in the vocabulary V. Each word is represented by a 200-dimensional word vector trained by PubMed and the MIMIC-III corpus [37]. However, not all words in the dataset are present in the vocabulary V. To deal with the problem of being out of vocabulary, a convolutional neural network (CNN) [38] is used in this paper to capture the character-level features of each word to obtain a character vector. To make full use of both word-level and character-level information, the word vector is finally concatenated with the character vector to represent biomedical entity mentions with candidate entities.

3.3.3. Encoding Layer

We use Bi-directional Gated Recurrent Unit (BiGRU) [39] to encode mentions and candidates separately because GRU is computationally more efficient than LSTM [40], and its performance is comparable to LSTM. BiGRU learns to represent a word (or character) and its context. The output state of the BiGRU at time i over the mention m is denoted by the symbol $\mathbf{m}_\mathrm{i}$. $\mathbf{c}_\mathrm{i}$ is the same way:

$$\mathbf{m}_\mathrm{i} = \mathrm{BiGRU}(m, i), \forall \mathrm{i} \in [1, \ldots, \mathrm{l_M}] \tag{4}$$

$$\mathbf{c}_\mathrm{i} = \mathrm{BiGRU}(c, i), \forall \mathrm{i} \in [1, \ldots, \mathrm{l_C}] \tag{5}$$

The gated recurrent unit (GRU) is a special type of recurrent neural network that captures the contextual order information of sequences. GRU can only encode historical information, while ignoring future contextual information. In this paper, a bidirectional GRU network consisting of both forward GRU and reverse GRU is used. BiGRU obtains

the final hidden layer representation by splicing two different hidden layer representations obtained by sequential and inverse order calculations.

3.3.4. Self-Attention Module

The text sequence features extracted by the BiGRU encoder ignore the different contributions of different words to the semantic representation of the whole entity. Therefore, this paper further improves the ability to extract global features by using a self-attention mechanism [41] in the intra-entity attention layer, which learns different semantic importance in the entity by the attention operations performed between each word. Each word is able to pay attention to the features of other segments in the same entity. The feature weights are dynamically adjusted by the self-attention mechanism to emphasize interdependent word features automatically. It can be used to find relationships within sequences, selectively focus on some important information, and give higher weights to the important ones. Thus, the problem of the equal contribution of each character is effectively solved.

Basically, it can be described as a mapping relationship between a query and a series of key-value pairs. The output is a weighted sum of these values, where the weight assigned to each value is calculated from the compatibility function of the query with the corresponding key-value. The self-attentive mechanism is defined as shown in the following Formula:

$$\text{Attention} = \text{softmax}\left(\frac{\mathrm{HW^{Q}W^{K^{T}}H^{T}}}{\sqrt{\mathrm{d}}}\right)\mathrm{HW^{V}} = \text{softmax}\left(\frac{\mathrm{QK^{T}}}{\sqrt{\mathrm{d}}}\right)\mathrm{V} \tag{6}$$

where $\mathrm{H} \in \mathrm{R^{n\times 2l}}$ denotes the output of the BiGRU layer, l is the hidden layer dimension of the GRU unit. $\mathrm{W^{Q}}, \mathrm{W^{K}}, \mathrm{W^{V}} \in \mathrm{R^{2l\times d}}$ is the trainable weight and d is the output dimension. In this paper, we take h_i as an example to further explore the execution process of the self-attentive mechanism, as shown in Figure 2.

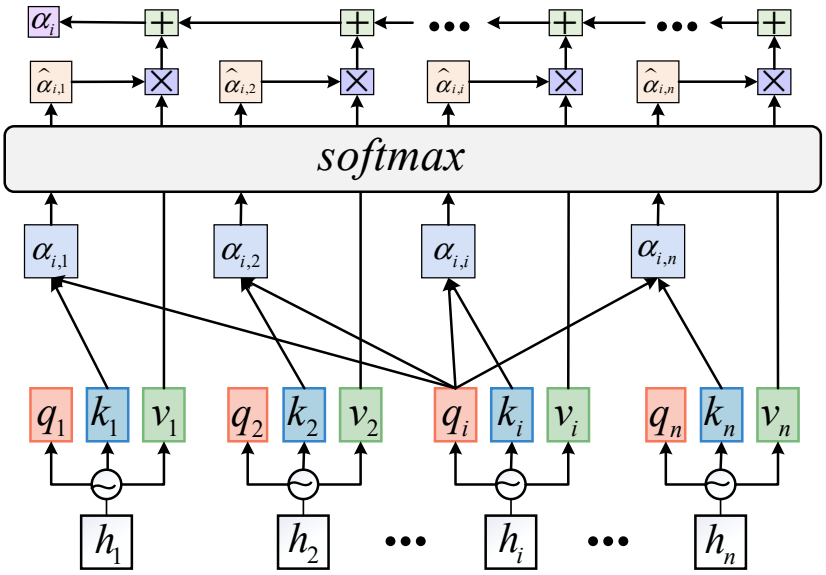

Figure 2. Calculation process of self-attention for h_i.

To be specific, given query vector $q_i \in Q = [q_1, q_2, \ldots, q_n]$ and key vector $k_i \in K = [k_1, k_2, \ldots, k_n]$. Firstly, the similarity $\alpha_{i,j} = q_i k_j / \sqrt{d}$ of q_i for each key k_j is computed by the scaled dot product function. The similarity score is divided by $\sqrt{d}$ to have stable gradients, and the weight coefficient $\alpha_{i.j}$ is computed by the softmax function. After that, a weighted summation operation is performed on each value $v_i \in V = [v_1, v_2, \ldots, v_n]$ to obtain the final attention $a_i = \sum_j \alpha_{i,j} v_j$ according to weighted coefficients $\alpha_{i.j}$.

3.3.5. Cross-Attention Module

Although the self-attention module described above may efficiently utilize the intra-entity relationship, the relationship between mention and candidate is not explored. In this section, we model the inter-entity relationship through the cross-attention module, where the attention weights of mentions and candidates can be mutually learned to learn the close association between text features. This allows us to achieve more accurate matching results by learning the close association between text features.

The inter-entity attention module first learns to capture the importance between the features of each pair of mention and candidate. Then an information flow is passed between the two models to update each mention feature and candidate feature based on the learned importance weights and aggregated features. Such an information flow is able to identify the relationship between mention and candidate entity. The implementation of Cross-Attention Module is illustrated at Figure 3.

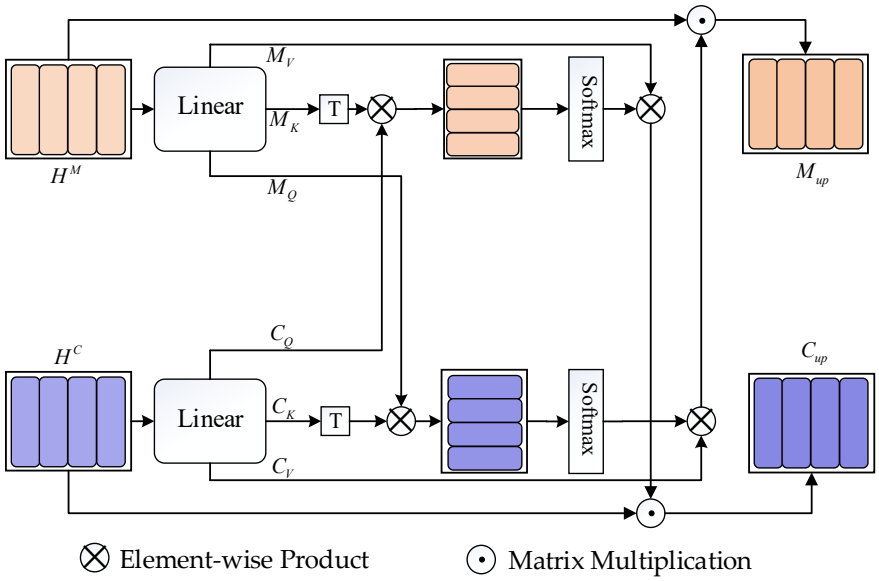

Figure 3. Cross-Attention module.

Given mention and candidate features, we first compute the association weights between each pair of word in mention and candidate. Each word feature is transformed into query, key, and value features by utilizing linear projection, where the transformed mention features are denoted as $\mathrm{M_Q}, \mathrm{M_K}, \mathrm{M_V} \in \mathrm{R}^{20\times \mathrm{dim}}$, and the candidate features denoted as $\mathrm{C_Q}, \mathrm{C_K}, \mathrm{C_V} \in \mathrm{R}^{20\times \mathrm{dim}}$.

By calculating the inner product between the mention features $\mathrm{M_Q}$ and the candidate features $\mathrm{C_K}$, we can obtain the initial attention weights, and then apply the softmax function to normalize them row-wisely.

$$\mathrm{InterA_{M\leftarrow C}} = \mathrm{softmax}\left(\mathrm{M_Q C_K^T}\right) \tag{7}$$

$$\mathrm{InterA_{M\rightarrow C}} = \mathrm{softmax}\left(C_\mathrm{Q} M_\mathrm{K}^\mathrm{T}\right) \tag{8}$$

These two bidirectional InterA matrices capture the importance between each mention and candidate word pair. Taking $\mathrm{InterA_{M\leftarrow C}}$ as an example, each row represents the attention weight between a word in a mention and all word embeddings of the candidate. The final attention vector $\mathrm{M_{attn}}$ is then aggregated as the weighted summation of the candidate word value features C_V.

$$\mathrm{M_{attn}} = \mathrm{InterA_{M\leftarrow C}} \times \mathrm{C_V} \tag{9}$$

$$C_{attn} = InterA_{M \to C} \times M_V \tag{10}$$

Finally, we do a similarity calculation between the features and the attention vector for each moment. We denote the information flow of the updated mention features and candidate features as $M_{up} \in R^{20 \times dim}$ and $C_{up} \in R^{20 \times dim}$ respectively.

$$M_{up} = M_{attn} \odot M_Q \tag{11}$$

$$C_{up} = C_{attn} \odot C_Q \tag{12}$$

where the operator $\odot$ refers to matrix multiplication.

3.3.6. Aggregation Layer

Rich mention and candidate representation are available by stitching together all interaction results, and this layer is used to aggregate two matched vector sequences into a fixed-length vector. CNN has been shown to excel in learning sentence vector expressions from both syntactic and semantic levels simultaneously in a variety of natural language processing tasks, and the unique convolutional operation allows them to learn the features of long sequences of text with stability. The unique convolution operation allows it to learn features of short sequences with stable expressions in long sequences of text, independent of their position of occurrence. The CNN is more suitable for this model since its sequential nature is not strong when the interaction model is used. Therefore, we use the CNN model to apply it to the sequence of entity mentions and candidate matching vectors, respectively, and then stitch the CNN output features together to construct fixed-length matching vectors. The core of CNN is the convolutional layer, which can encode important information in the input data with fewer parameters. The convolutional layer is equivalent to a sliding window, which allows contextual features to be obtained within a local window of the current word. In the biomedical entity linking task, we found that good performance can be achieved with only one convolutional layer. In general, multiple convolutional kernel sizes perform better than a single size. The convolution operation is as follows:

$$h_i = \delta\left(W_h^T \cdot x_{i:i+m-1} + b_h\right) \tag{13}$$

where, $W_h \in R^{m \times d}$, d is the word embedding dimension and m is the convolution kernel size. $X_{i:i+m-1}$ represents a window starting from the i th contextual embedding to the $i + m-1$th contextual embedding, b_h is a bias vector. $\delta(\cdot)$ denotes a nonlinear activation function. The output of the feature map is $H = [h_1, h_2, \cdots, h_n]$, where n is the number of convolution windows and h_i is the result of each convolution.

After that, by selecting the maximum value of each feature map, its most important features can be captured. Using maximum pooling for all convolution kernels and then cascading them together gives the final feature vector f_h.

Finally, we use a two-layer fully connected neural network to compute the final result.

$$O_h = ReLU(W_1 f_h + b_1) \tag{14}$$

$$Score(M, C) = sigmoid(W_2 O_h + b_2) \tag{15}$$

where O_h is the first layer of output features, W_1 and W_2 are trainable weight matrices, and b_1 and b_2 are biases.

3.3.7. Objective Function

In this paper, a triplet loss function [42] is used to train the model. The neural network model based on the triplet loss function can distinguish the details well, especially in the entity linking task. When the mentions are very similar to the candidate entities, the triplet loss function can learn more subtle features for these two input vectors with fewer differences. The purpose of the triplet loss function is to separate positive and negative

sample pairs at a certain distance (margin) by optimizing the embedding space to ensure that the positive sample pair is close enough to each other and the negative sample pair is far enough away from each other. The idea of Triplet loss can be formally expressed as follows.

$$\text{Loss} = \max\{\text{Score}(M, C^+)\text{-Score}(M, C^-) + \eta, 0\} \quad (16)$$

To prevent uneven data selection from leading to unstable performance of the model training process, in this paper, positive examples are randomly obtained from the training set and synonym entities in the knowledge base, and negative examples are drawn from the candidate entities generated in the candidate entity generation phase (excluding the correct entities). This selection makes the negative examples very similar to the positive ones and forces the model to learn more subtle differences between the positive candidate entities and other candidate entities.

4. Experimental Results

4.1. Dataset

To demonstrate the effectiveness of our proposed method, we carried out extensive experiments on two publicly available biomedical entity linking benchmark datasets: the NCBI-NCBI disease corpus and the ADR-TAC 2017 Adverse Reaction Extraction (ADR) dataset. The statistics for both datasets are shown in Table 1.

Table 1. Statistics of the two types of datasets.

	NCBI		**ADR**	
	Train	**Test**	**Train**	**Test**
document	692	100	101	99
mentions	5921	960	7038	6343
NIL	0	0	47	18
concepts	9664		23,668	

NCBI: This is one of the most popular datasets for biomedical entity linking tasks. It contains 792 PubMed abstracts, of which 692 abstracts were used for training and development, and 100 abstracts were used for testing. The 6 July 2012 version of MEDIC, which contains 7827 MeSH identifiers and 4004 OMIM identifiers, was used in this paper, and it contains 9664 disease concepts. All annotated disease mentions have their corresponding concept identifiers.

ADR: This dataset consists of 200 drug labels, divided into 101 labels for training and development and 99 labels for testing. The ADRs in each drug label were manually mapped to the MedDRA 18.1 knowledge base, which contains 23,668 concepts. In this dataset, only 0.7% of the training mentions and 0.3% of the test mentions were unlinkable.

4.2. Evaluation Metrics

The biomedical candidate entity generation task uses recall (Recall) as an evaluation metric, which is calculated as shown in Equation (17).

$$\text{Recall} = \frac{|P \cap Q|}{|P|} \quad (17)$$

where P and Q denote the mentions to be linked and the set of candidate entities, respectively.

In the candidate ranking stage, following previous work, accuracy is used in this paper to evaluate the performance of the entity linking algorithm, i.e., the percentage of mentions that are correctly linked.

$$\text{Accuracy} = \frac{T}{N} \quad (18)$$

where, T is the predicted correct biomedical entities and N is the total number of entities to be linked.

4.3. Experiment Settings

The experiments in this chapter are based on Python 3.7, and the proposed network is implemented using the Tensorflow deep learning framework with an Intel(R) Xeon(R) E5-2678 v3 @ 2.50 GHz CPU, a GeForce RTX 3090 GPU graphics card, and 24 G of running memory.

The parameters of the deep learning model in this experiment are shown in Table 2. In this paper, Adam was chosen as the optimizer for the experiments. We use dropouts the in BiGRU encoding layer, CNN aggregation layer, and the fully connected layer.

Table 2. Hyperparameters Setting.

Hyperparameters	Value
Dimension of word embeddings	200
Dimension of char embeddings	64
Learning rate	0.001
dropout	0.1
GRU hidden size	32
Kernel sizes	1,2,3
Filters	64
Batch size	64
Epochs	30

4.4. Benchmarks

In order to verify the validity of the method proposed in this paper, several recent state-of-the-art methods on the NCBI and ADR will be selected for comparison.

1. Sieve-based Model [21]: A manual rule-based multi-channel sieving system, which is by far the best rule-based system on the NCBI dataset.
2. Dnorm [11]: A pairwise ranking learning approach using similarity matrix to measure the degree of similarity between biomedical mentions and standard entities, and it is a machine learning based approach.
3. TaggerOne [13]: It is the best machine learning based approach on the NCBI dataset using a semi-Markov model jointly for named entity recognition and entity linking.
4. Learning to Rank [10]: A method for learning to rank, best performance in the TAC2017 ADR Challenge, a machine learning based system.
5. CNN-based Ranking [6]: This approach treats biomedical entity linking as a ranking problem and uses CNN to model semantic similarity between mentions and candidate entities.
6. BNE [30]: A novel encoding framework that considers all these aspects in representation learning.
7. NormCo [29]: A deep learning model, in which the biomedical entity linking task is accomplished by combining morphological similarity and semantic similarity computed using the GRU model.
8. TripletNet [31]: We make use of the Triplet Network for candidate ranking.
9. BERT-based Ranking [14]: This approach treats biomedical entity linking as a sentence pair classification task and accomplishes entity linking by fine-tuning the BERT pre-training model.

4.5. Results and Analysis

In this section, we demonstrate the effectiveness of our proposed model on two benchmark datasets. Firstly, the model of this paper is compared with the current model of linking biomedical entities. Next, the properties of the proposed model are demonstrated by some ablation experiments. Finally, the lightweight model is compared with the state-of-the-art BERT-based model.

4.5.1. Candidate Generation Experiment

In this paper, in the process of candidate entity generation using the alignment method, the recall rate increases as the recall range increases, and the Top20 candidate entity recall effect reaches a high level. The recall rate of correct entities on the NCBI and ADR test sets is 94.52% and 96.73%, respectively, and the experimental results are shown in Table 3. The candidate generation method based on aligned cosine similarity used in this paper does not miss too many correct candidate entities, indicating that the method is effective for the biomedical candidate generation task. Finally, we generated 20 candidate entities for each mention to ranking.

Table 3. Comparison of candidate entity generation results in different recall ranges.

Recall Range	NCBI	ADR
Top1	85.46	88.62
Top2	88.07	91.57
Top5	91.39	94.04
Top10	92.95	95.56
Top15	93.96	96.45
Top20	94.52	96.73

4.5.2. Comparison Experiment

We compare our model with several recent state-of-the-art non-BERT methods on NCBI and ADR datasets. The results in Table 4 are taken from the original state-of-the-art papers. Since the two experimental datasets used in this paper are public and the training and testing parts have been divided, we think that the results of the original paper are comparable. The performance results show that the model in this paper outperforms the baseline approach with an accuracy of 91.28% and 93.13%, respectively. Compared with the rule-only or traditional machine learning baseline approaches, the deep learning model in this paper achieves a significant improvement in accuracy by 6.63% over the Sieve-based model, 9.08% over Dnorm, and 2.48% over TaggerOne on the NCBI dataset. On the ADR dataset, it improved by 1.08% compared to Learning to Rank. The CNN-based entity linking model with the NormCo model ignores the rich interaction information between candidate entities and mentions, which limits its performance. The framework of this paper outperforms the CNN-based approach by 5.18% and 2.89% on the NCBI and ADR datasets, respectively, which indicates that the attention mechanism is more effective than the single Siamese representations. From the performance results, we can also see that our model works better than TripletNet. The superiority of this model can be attributed to the fact that it utilizes the self-attention module and cross-attention module to form a unified network to better capture the interaction information between biomedical mentions and candidate entities themselves and each other and to better perform the matching task.

Table 4. Comparison with non-BERT methods. The best performance on each dataset is marked in bold and "-" denotes the result not provided.

Model	NCBI	ADR
Sieve-based Model [21]	84.65	-
Dnorm [11]	82.20	-
TaggerOne [13]	88.80	-
Learning to Rank [10]	-	92.05
CNN-based Ranking [6]	86.10	90.24
BNE [30]	87.70	-
NormCo [29]	87.80	-
TripletNet [31]	90.01	-
Our model	**91.28**	**93.13**

4.5.3. Ablation Study

To demonstrate the effectiveness of the various components of the model, some ablation studies are also conducted in this paper. We construct four ablation models (w/o BiGRU, w/o Cross-Attention, w/o Self-Attention, w/o CNN) by eliminating a component at each time. Table 5 shows the accuracy rates on the test set. First, we study the impact of the BiGRU encoder and compare the ablation model with the "Full model". We find that the impairment of performance by removing the BiGRU is about 1.13% and 1.12% on two datasets. After that, we evaluate the effectiveness of the attention mechanism. For this purpose, we construct the ablation model by removing the attention mechanism. From the experimental results, we can see that the attention mechanism has a significant impact on performance. Removing cross-attention has a larger impact on the model performance, with a decrease of 2.90% and 4.65% on the NCBI and ADR datasets, respectively, proving its effectiveness in considering the full alignment between entity mention-candidate entity pairs. In the case of removing self-attention, the accuracy decreases by 1.04% and 0.54%, respectively. Thus, better matching results can be achieved by weighted features computed by self-attention, which we can use to investigate potential alignments more carefully and precisely. Finally, removing the CNN decreases the accuracy by 2.43% and 1.73%. It is clear that adding CNN to the aggregation layer is complementary to extracting more fine-grained features.

Table 5. Ablation studies of our proposed model on NCBI and ADR test dataset.

Model	NCBI	ADR
Full model	**91.28**	**93.13**
w/o BiGRU	90.15	92.01
w/o Cross-Attention	88.38	88.48
w/o Self-Attention	90.24	92.59
w/o CNN	88.85	91.40

4.5.4. Analysis of Margin Value λ

In addition, in order to study the effect of margin value on the model effect in the triplet loss function, different margin values were set for comparison experiments, and the experimental results are shown in Figure 4. For the NCBI dataset and ADR dataset, the model achieved the best results when the margin value was 0.1. When the margin value was set too low, the loss tended to be close to zero, and it was difficult to distinguish similar entities. When the margin value was set to 0, the accuracy was only 88.38% and 89.65% on both datasets. When the margin value is set too large, the loss value keeps a large value, making it difficult to converge. Therefore, it is critical to set a reasonable margin value, which is an important indicator of similarity.

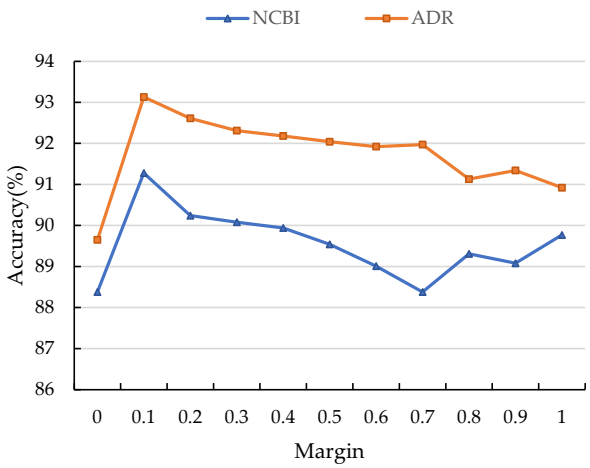

Figure 4. The impact of different margin values.

4.5.5. Comparisons with BERT-Based Methods

The original BERT-based biomedical entity linking method uses the pointwise ranking model to treat the ranking problem as a classification problem. However, this Pointwise ranking model does not abstract the relevant features mentioned by the candidates and mentions well, and it is difficult to distinguish the subtle features.

In response to the above problem, in this paper, we also try to introduce BERT and improve the original model utilizing Pairwise ranking. We propose a biomedical entity linking method (BPR) based on BioBERT and Pairwise ranking to learn better semantic representations. We introduce positive and negative entities and generate the form of triplets with mentions, then obtain semantic relevance representations by BioBERT pre-training model, respectively, and use the triplet loss function for training. Some effect improvement is achieved in the biomedical ranking task.

We also compare the proposed lightweight model with the BERT-Base model, which has 12 layers, 768 hidden dimensions, and 12 attention heads with a total of 107 million. Despite having less than 5 million parameters, our model based on inter-entity and intra-entity attention achieves very competitive or even better results than the BERT-based SOTA model on both datasets, and the experimental results are shown in Table 6.

Table 6. Performance comparison with BERT-based methods.

Model	NCBI	ADR
BERT -based Ranking [14]	89.06	93.22
BPR	90.94	**93.56**
Our model	**91.28**	93.13

To show the efficiency of the model in this paper, we also compared the complexity (parameter size) and inference time of the model with the BERT-base model, and Table 7 shows the comparison results. It is the time in the entire testing set. The comparison results show that the method in this paper has a very high CPU inference speed. The model in this paper is 9 times faster compared to the BERT-Base model and 11.8 times faster compared to the BERT(Pairwise) model, and the complexity of the model is much smaller, with about 23 times fewer model parameters. In summary, the experimental results show that the lightweight biomedical entity linking model proposed in this paper achieves performance comparable to state-of-the-art models on two benchmark datasets, with only a small number of parameters and fast inference. When speed and model size are taken into account, the method in this paper is easier and more practical to use for deployment and application.

Table 7. Comparison of parameter size and inference time with BERT-base model.

Model	Parameters	NCBI	ADR	Average	Speedup
BERT-Base	107 M	317 s	1381 s	849 s	9.0×
BERT (pairwise)	107 M	392 s	1822 s	1107 s	11.8×
Our model	**4.7 M**	**26 s**	**162 s**	**94 s**	**-**

4.5.6. Analysis of Different Dataset Size

By subsampling the dataset, we also investigate the performance of the model on training samples of different sizes, as shown in Figure 5. The performance of the model in this paper grows when the number of training samples is gradually increased. When only 20% of the training samples were used, the accuracy on the NCBI and ADR datasets also reached 89.04% and 91.13%, respectively. More data will bring better performance, and the biomedical entity linking model in this paper can achieve better results despite using a small amount of labeled data.

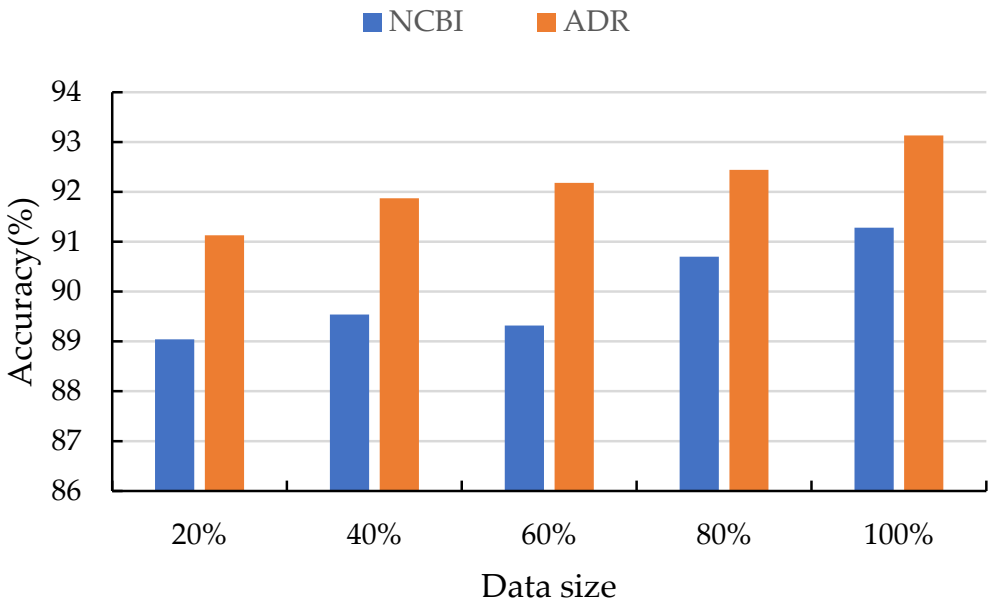

Figure 5. Effects of different data sizes on performance of our model.

4.5.7. Case Study of Removing Inter- and Intra-Entity Attention

In this part, we look more closely at the attention in a typical case. As a comparison, we used the model without self-attention or cross-attention. For the prediction results of the method in this paper, two samples are selected, as shown in Table 8. Through the self-attention mechanism, each token in the mentions and candidates is given a weight to show how important each token is. By adding the cross-attention, the model can capture the keywords in entity mentions, and thus predict the standard words for irregular entity mentions. By integrating the self-attention and cross-attention modules, the model in this paper is highly capable of discovering and distinguishing the matching details and subtle features between candidate entities and entity mentions.

Table 8. Effect of inter- and intra-entity attention on prediction results.

Model	Mention	Prediction	Ground-Truth
with self-attention	bacterial infections opportunistic	bacterial infection	bacterial infection
without self-attention	bacterial infections opportunistic	opportunistic infections	bacterial infection
with cross-attention	difficulty concentration	disturbance in attention	disturbance in attention
without cross-attention	difficulty concentration	liver iron concentration increased	disturbance in attention

4.5.8. Error Analysis

For the prediction results of the method in this paper, three samples of prediction errors were selected, as shown in Table 9. From the prediction error samples in Table 9, the following conclusions can be obtained: for the case that one biomedical mention corresponds to multiple candidate entities, the model in this paper does not predict well. In addition, the more common causes of error cases are those of the same symptom part with different symptom modifiers, which is where this paper can continue to improve and enhance.

Table 9. Sample of error prediction.

Mention	Ground-Truth	Prediction
colorectal breast and other cancers	breast neoplasms, colorectal cancer, cancers	cancer of breast
nail abnormalities	nail disorder	nail pitting
colorectal adenomas and carcinoma	adenomatous polyp, colorectal carcinoma	colorectal adenomas

5. Conclusions and Future Work

Entity linking has received increasing attention as a fundamental task for various types of medical natural language processing tasks. In order to address the challenge of large numbers of parameters in large pre-trained models, the long inference time, and the difficulty of obtaining good results with a single matching model due to the excessive variety of biomedical mention representations, in this paper, we construct an efficient biomedical entity linking method that incorporates inter- and intra-entity attention in a unified model to better capture information between biomedical mentions and candidate entities themselves as well as between each other. The model in this paper is also more lightweight. We have systematically studied the influence of our idea and carried out experiments. Furthermore, we also designed a biomedical entity linking method based on BERT and pairwise ranking to compare with the lightweight method in this paper. Experimental results demonstrate that the proposed method in this paper achieves fairly competitive performance on two biomedical benchmark datasets. Furthermore, it also achieves comparable or even better results compared to the BERT-based entity linking method while having far fewer model parameters and very high inference speed. The results demonstrate the effectiveness of our model by achieving significant performance.

The biomedical entity linking method proposed in this paper can solve the problems in entity linking, but there are still some limitations, which will be addressed in future work. The specific shortcomings and improvement measures are as follows:

Firstly, the recall rate of the candidate phase directly determines the accuracy of the candidate ranking phase. It is worthwhile to further improve the upper bound of the ranking system. In addition, the task in this paper can also include features such as prior information, contextual information, and coherence information, and it is expected that the additional inclusion of this part of information can further improve the effect, which is to be further studied subsequently. Finally, the analysis of the incorrectly predicted entity mentions reveals that the ranking model is inaccurate in predicting the presence of a mention corresponding to more than one criterion word, for which a new model can be designed in future work to handle the prediction task of a mention corresponding to multiple criterion candidates.

Author Contributions: Conceptualization, M.A. and T.T.; methodology, M.A. and T.T.; software, M.A.; validation, M.A. and T.T.; formal analysis, M.A., T.T. and A.H.; investigation, M.A. and T.T.; data curation, M.A.; writing—original draft preparation, M.A.; writing—review and editing, M.A., T.T. and A.H.; visualization, M.A.; supervision, T.T. and A.H.; project administration, T.T. and A.H.; funding acquisition, T.T. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the National Natural Science Foundation of China (62166042, U2003207), Natural Science Foundation of Xinjiang, China (2021D01C076), and Strengthening Plan of National Defense Science and Technology Foundation of China (2021-JCJQ-JJ-0059).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset supporting the conclusions of this article is available at https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/ and https://bionlp.nlm.nih.gov/tac2017adversereactions/, accessed on 21 December 2021.

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

References

1. Shen, W.; Wang, J.; Han, J. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Trans. Knowl. Data Eng.* **2014**, *27*, 443–460. [CrossRef]
2. Huang, K.; Yang, M.; Peng, N. Biomedical event extraction with hierarchical knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1277–1285.
3. Zhang, Z.; Parulian, N.; Ji, H.; Elsayed, A.; Myers, S.; Palmer, M. Fine-grained information extraction from biomedical literature based on knowledge-enriched Abstract Meaning Representation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; Volume 1, pp. 6261–6270.
4. Lee, J.; Sean; Yi, S.; Jeong, M.; Sung, M.; Yoon, W.; Choi, Y.; Ko, M.; Kang, J. Answering questions on COVID19 in real-time. In Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, Online, 15 December 2020.
5. Zheng, J.G.; Howsmon, D.; Zhang, B.; Hahn, J.; McGuinness, D.; Hendler, J.; Ji, H. Entity linking for biomedical literature. *BMC Med. Inform. Decis. Mak.* **2014**, *15*, 1–9.
6. Li, H.; Chen, Q.; Tang, B.; Wang, X.; Xu, H.; Wang, B.; Huang, D. CNN-based ranking for biomedical entity normalization. *BMC Bioinform.* **2017**, *18*, 385. [CrossRef] [PubMed]
7. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A Nucleus for a Web of Open Data. In *Semantic Web*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 722–735.
8. Suchanek, F.M.; Kasneci, G.; Weikum, G. Yago: A core of semantic knowledge. In Proceedings of the 16th International Conference on World Wide Web, Banff, AB, Canada, 8–12 May 2007; pp. 697–706.
9. Kang, N.; Singh, B.; Afzal, Z.; van Mulligen, E.M.; Kors, J.A. Using rule-based natural language processing to improve disease normalization in biomedical text. *JAMIA* **2012**, *20*, 876–881. [CrossRef] [PubMed]
10. Xu, J.; Lee, H.-J.; Ji, Z.; Wang, J.; Wei, Q.; Xu, H.; TAC. UTH_CCB System for Adverse Drug Reaction Extraction from Drug Labels at TAC-ADR 2017. 2017. Available online: https://tac.nist.gov/publications/2017/participant.papers/TAC2017.UTH_CCB.proceedings.pdf (accessed on 8 March 2020).
11. Leaman, R.; Doğan, R.I.; Lu, Z. DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics* **2013**, *29*, 2909–2917. [CrossRef] [PubMed]
12. Luo, Y.; Song, G.; Li, P.; Qi, Z. Multi-Task Medical Concept Normalization Using Multi-View Convolutional Neural Network. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 1, pp. 5868–5875.
13. Leaman, R.; Lu, Z. TaggerOne: Joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics* **2016**, *32*, 2839–2846. [CrossRef] [PubMed]
14. Ji, Z.; Wei, Q.; Xu, H. Bert-based ranking for biomedical entity normalization. AMIA Joint Summits on Translational Science proceedings. *AMIA Joint Summits Transl. Sci. Proc.* **2020**, *2020*, 269–277.
15. Huang, P.S.; He, X.; Gao, J.; Deng, L.; Acero, A.; Heck, L. Learning deep structured semantic models for web search using clickthrough data. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, San Francisco, CA, USA, 27 October–1 November 2013; pp. 2333–2338.
16. Shen, Y.; He, X.; Gao, J.; Deng, L.; Mesnil, G. A latent semantic model with convolutional pooling structure for information retrieval. In Proceedings of the 23rd ACM International Conference on Conferenceon Information and Knowledge Management, Shanghai, China, 3–7 November 2014; pp. 101–110.
17. Hu, B.; Lu, Z.; Li, H.; Chen, Q. Convolutional neural network architectures for matc hing natural language sentences. *Adv. Neural-Form. Processing Syst.* **2014**, *27*, 2042–2050.
18. Pang, L.; Lan, Y.; Guo, J.; Xu, J.; Wan, S.; Cheng, X. Text matching as image recognition. *arXiv* **2016**, arXiv:1602.06359.
19. Guo, J.; Fan, Y.; Ai, Q.; Croft, W.B. A deep relevance matching modelfor ad T hoc retrieval. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Online, 24 October 2016; pp. 55–64.
20. Chen, Q.; Zhu, X.; Ling, Z.; Wei, S.; Jiang, H.; Inkpen, D. Enhanced Istm for natural language inference. *arXiv* **2016**, arXiv:l609.06038.
21. D'Souza, J.; Ng, V. Sieve-Based Entity Linking for the Biomedical Domain. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015), Beijing, China, 26–31 July 2015; pp. 297–302.
22. Dogan, R.I.; Leaman, R.; Lu, Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Inform.* **2014**, *47*, 1–10. [CrossRef] [PubMed]
23. Ghiasvand, O.; Kate, R.J. UWM: Disorder mention extraction from clinical text using CRFs and normalization using learned edit distance patterns. In Proceeding of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; pp. 828–832.
24. Bodenreider, O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32* (Suppl. S1), 267–270. [CrossRef]
25. Lee, C.-P.; Lin, C.-J. Large-scale linear ranksvm. *Neural Comput.* **2014**, *26*, 781–817. [CrossRef] [PubMed]

26. Roberts, K.; Demner-Fushman, D.; Tonning, J.M.; TAC. Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track. 2017. Available online: https://tac.nist.gov/publications/2017/additional.papers/TAC2017.ADR_overview.proceedings.pdf (accessed on 7 March 2020).
27. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*; The Mit Press: Cambridge, MA, USA, 2013; pp. 3111–3119.
28. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Qatar, Doha, 25–29 October 2014; pp. 1532–1543.
29. Wright, D. NormCo: Deep Disease Normalization for Biomedical Knowledge Base Construction. Ph.D. Thesis, University of California, San Diego, CA, USA, 2019.
30. Phan, M.C.; Sun, A.; Tay, Y. Robust representation learning of biomedical names. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3275–3285.
31. Mondal, I.; Purkayastha, S.; Sarkar, S.; Goyal, P.; Pillai, J.; Bhattacharyya, A.; Gattu, M. Medical entity linking using triplet network. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019; pp. 95–100.
32. Yan, C.; Zhang, Y.; Liu, K.; Zhao, J.; Shi, Y.; Liu, S. Biomedical Concept Normalization by Leveraging Hypernyms. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3512–3517.
33. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv181004805.
34. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
35. Sohn, S.; Comeau, D.C.; Kim, W.; Wilbur, W.J. Abbreviation definition identification based on automatic precision estimates. *BMC Bioinform.* **2008**, *9*, 402. [CrossRef] [PubMed]
36. Yadav, V.; Bethard, S.; Surdeanu, M. Alignment over Heterogeneous Embeddings for Question Answering. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019.
37. Zhang, Y.; Chen, Q.; Yang, Z.; Lin, H.; Lu, Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci. Data* **2019**, *6*, 1–9. [CrossRef] [PubMed]
38. Zhang, X.; Zhao, J.; Lecun, Y. *Character-Level Convolutional Networks for Text Classification*; MIT Press: Cambridge, MA, USA, 2015.
39. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
40. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
41. Vaswani, A.; Shazeer, N.; Parmar, N. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *2*, 5998–6008.
42. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.