Accepted Manuscript

Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition

Kai Xu, Zhenguo Yang, Peipei Kang, Qi Wang, Wenyin Liu

PII: S0010-4825(19)30110-6
DOI: https://doi.org/10.1016/j.compbiomed.2019.04.002
Reference: CBM 3259

To appear in: *Computers in Biology and Medicine*

Received Date: 7 December 2018
Revised Date: 1 April 2019
Accepted Date: 1 April 2019

Please cite this article as: K. Xu, Z. Yang, P. Kang, Q. Wang, W. Liu, Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition, *Computers in Biology and Medicine* (2019), doi: https://doi.org/10.1016/j.compbiomed.2019.04.002.

Document-Level Attention-Based BiLSTM-CRF Incorporating Disease Dictionary for Disease Named Entity Recognition

Kai Xu[a], Zhenguo Yang[a,b,*], Peipei Kang[a], Qi Wang[a], Wenyin Liu[a,*]

[a]*Department of Computer Science, Guangdong University of Technology, Guangzhou, China*
[b]*Department of Computer Science, City University of Hong Kong,Hong Kong,China*

Abstract

Background: Disease named entity recognition (NER) plays an important role in biomedical research. There are a significant number of challenging issues to be addressed; among these, the identification of rare diseases and complex disease names and the problem of tagging inconsistency (i.e., if an entity is tagged differently in a document) are attracting substantial research attention.

Methods: We propose a new neural network method named Dic-Att-BiLSTM-CRF (DABLC) for disease NER. DABLC applies an efficient exact string matching method to match disease entities with a disease dictionary; here, the dictionary is constructed based on the Disease Ontology. Furthermore, DABLC constructs a dictionary attention layer by incorporating a disease dictionary matching method and document-level attention mechanism. Finally, a bidirectional long short-term memory network and conditional random field (BiLSTM-CRF) with a dictionary attention layer is proposed to combine the disease dictionary to develop disease NER.

Results: Extensive experiments are conducted on two widely-used corpora: the NCBI disease corpus and the BioCreative V CDR corpus. We apply each test on 10 executions of each model, with a 95% confidence interval. DABLC achieves the highest F1 scores (NCBI: Precision = 0.883, Recall = 0.89, F1 = 0.886; BioCreative V CDR: Precision = 0.891, Recall = 0.875, F1 = 0.883), outperforming the state-of-the-art methods.

Conclusion: DABLC combines the advantages of both external dictionary resources and deep attention neural networks. This aids the identification of rare diseases and complex disease names; moreover, it reduces the impact of tagging inconsistency. Special disease NER and deep learning models addressing long sentences are noteworthy areas for future examination.

Keywords: Biomedical informatics, named entity recognition, string matching, machine learning, neural network.

*Corresponding authors

Email addresses: kaixu.gdut@foxmail.com (Kai Xu), zhengyang5-c@my.cityu.edu.hk (Zhenguo Yang), ppkanggdut@126.com (Peipei Kang), wangqi_6414@sina.com (Qi Wang), liuwy@gdut.edu.cn (Wenyin Liu)

1. Introduction

Named entity recognition (NER) is the task of identifying a specific mention, such as a geographical or individual's name. In the context of biomedical literature information analysis, NER is the fundamental pre-prepossessing step, e.g., identifying disease entities, drug entities, therapeutic entities, genes, proteins, etc. As an important class of medical named entities, disease NER is widely-used in medical research [1] in areas such as disease prevention, disease treatment, clinical diagnosis, disease causes, and relationship analysis [2]. The development of high-performance disease NER systems is highly significant for promoting medical research.

The complex composition of disease concepts hinders disease NER [2]; the challenging issues are mainly with respect to four aspects. First, certain disease names are likely to contain roots and affixes in Greek and Latin. Second, a large number of rare diseases are challenging to identify. Third, a significant number of disease names contain complex modifications such as human body parts, which increases the difficulty of NER. Fourth, a disease name frequently has multiple representations, i.e., the problem of tagging inconsistency. A large number of disease names in biomedical texts are recorded in abbreviated forms. Certain forms of the abbreviations are largely irregular, including a few author-defined abbreviations.

Early biomedical NER methods used rule-based dictionary matching and machine learning methods. Lin et al. [3] used a rule-based approach to identify biomedical entities including proteins and DNA. Jimeno et al. [4] used MetaMap [5] and a dictionary matching method to identify diseases. Lowe et al. [6] used dictionaries and grammar for the disease NER task. Conditional random fields (CRF) [7] were widely-used in sequence labelling tasks. Sun et al. [8] applied shallow syntactic features to a conditional random field (CRF) model in a biological NER task. Lee et al. [9] combined string matching and CRF methods for NER of Korean clinical texts. Leaman et al. [10] used CRF for biomedical NER tasks. Lee et al. [11] used two CRF models to identify disease named entities. The combination of multiple methods can combine the advantages of different methods to improve the overall performance. Campos et al. [12] used dictionary matching and machine learning and normalisation methods for a biomedical recognition task. Leaman et al. [2] used the MEDIC vocabulary [13] combined with a machine learning approach to identify diseases. Leaman et al. [14] combined a machine learning model and normalisation; the performance of the model is higher than that of DNorm. However, the aforementioned methods rely excessively on complex feature engineering, which is a skill-dependent task.

Recently, deep learning technologies substantially improve speech and visual object recognition, through multi-level data representation learning [15]. In the context of biomedical NER, such as for adverse drug reaction discovery [16], chemical NER [17], and disease NER [18], deep learning models have also been widely-used in textual data representation and in NER steps. In terms of the textual data representation step, word embedding models [19] are generally used as the first step of input in the NER task; this can effectively improve the performance [20]. Typically, the skip-gram word embedding method [19] can be adopted to obtain semantic information and contextual information from the original text. In terms of the NER step, the long short-term memory networks (LSTM) including bidirectional LSTM are widely-used; these are effective for capturing long-range related information. Furthermore, researchers generally adopt the

CRF to predict the sequence labels; this is known as BiLSTM-CRF. BiLSTM-CRF can effectively improve the performance with feature extraction and reduce the workload of feature selection. For example, Wei et al. [18] combined CRF and bidirectional recurrent neural networks to recognise named entities; they then fed the two results into a support vector machine classier. Gridach et al. [21] used a character level representation to identify biological entities by using BiLSTM-CRF.

Most methods [18] [21] [22] model the same named entities from the sentence-level perspective, i.e., different sentences of a document are considered as independent labelling tasks. However, the same named entities in a document generally represent an identical meaning, whereas the sentence-level NER methods are likely to tag the same named entities as different tags (tagging inconsistencies). Ratinov et al. [23] used rule-based post processing steps to enforce tagging consistency to improve tag consistency. Based on this work, Luo et al. [17] proposed a document-level attention model to solve the tagging inconsistency problem and achieved the highest performance in the chemical NER task. To our knowledge, the highest F1 scores for the current methods on NCBI disease corpus and BioCreative V CDR corpus disease NER are 0.862 [24] and 0.876 [25], respectively. A large possibility for further improvement in the performance still exists.

In this study, we aim to address the challenges of identifying rare and complex disease names in the context of disease NER. To achieve this, we propose a new Dic-Att-BiLSTM-CRF (DABLC) method that incorporates both disease dictionary matching and a document-level attention mechanism into BiLSTM-CRF for disease NER. The disease dictionary consists of both common and rare disease entities. Then, we adopt an efficient exact string matching method for word-level matching. The dictionary-based attention weight vectors are calculated by combining both the dictionary and attention by weights.

More specifically, we train the word embedding first; this is used as input to the BiLSTM. Second, we combine the dictionary matching score with the document-level attention score in a weighted manner. Third, the output score of the BiLSTM and the output weight value vector of the dictionary attention layer are combined to calculate the confidence scores for each word. Finally, we calculate the predicted score by summing the confidence scores and CRF transition scores at the document-level.

Therefore, our proposed DABLC can utilise the external disease dictionary resources effectively for entity matching; this aids the performance improvement. On the two widely-used disease datasets, i.e., NCBI disease corpus [1] and BioCreative V CDR corpus [26], the DABLC method obtains the highest F1 scores of 0.886 and 0.883, respectively; these are higher than those of the state-of-the-art methods including Dnorm[2], TaggerOne[14], and cTAKES[27], AuDis[11].

The main contributions of this study are summarised as follows:

- We propose to construct a dictionary of disease entities by utilizing the authoritative disease knowledge resources, which cover a large number of disease entities including rare and complex disease names.

- We design an effective dictionary matching method to utilise the results of the dictionary matching and the document-level attention mechanism; it assigns the disease entity with a degree of attention in a weighted manner, aiding the performance improvement.

- We propose the DABLC method incorporating both the dictionary matching technique and attention mechanism, and utilise the BiLSTM to capture the context and CRF to calculate the sequence tags simultaneously. The DABLC method achieves the highest performance on the two widely-used public datasets.

The rest of the paper is organised as follows: Section 2 describes the details of the proposed DABLC approach. Section 3 presents the experiments and analyses the experimental results. Section 4 summarises this paper.

2. Materials and method

In this section, we introduce the details of the DABLC, which consists of four modules (Figure 1): data pre-processing module, semantic word embedding training module, disease dictionary construction, and search algorithm module and Dic-Att-BiLSTM-CRF module combining disease dictionary and attention mechanism. First, we perform pre-processing on the input text, segmentation of words and sentences, and part-of-speech (POS) tagging. The tags of the part-of-speech tagging will be used by the CRF in the fourth module. Second, we use full-text biomedical resources, including PubMed and PMC OA, and adopt the skip-gram strategy to learn semantic word embeddings. Third, we merge five medical knowledge databases to construct a disease dictionary and adopt an efficient exact string matching method for dictionary matching. Finally, the predicted sequence tags are calculated by using BiLSTM in conjunction with the dictionary attention method and CRF.

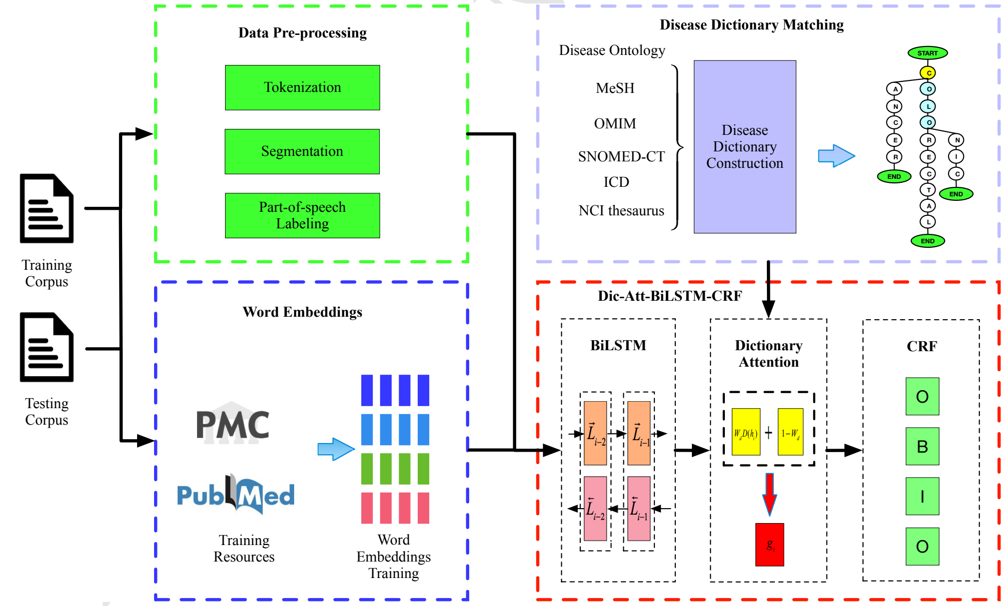

Figure 1: Architecture of Dic-Att-BiLSTM-CRF (DABLC) method. DABLC consists of four modules: data pre-processing module, disease dictionary matching module, word embeddings module, and Dic-Att-BiLSTM-CRF module.

2.1. Data pre-processing

We use two widely-used corpora, i.e., NCBI disease corpus and Bio Creative V CDR corpus, to evaluate the DABLC method. Both corpora have three datasets: training set, development set, and testing set. We first preprocess all the datasets, including sentence and article segmentation, tokenisation, and part-of-speech tagging. The datasets are segmented according to the tags of articles and sentences. For the tokenisation and POS tagging tasks, we used NLTK [28], a widely-used natural language toolkit.

2.2. Word embeddings

Hinton et al. [19] proposed a distributed representation method that used a vector to represent words in the semantic dimension. The distributed representation provides more accurate information for further machine learning models [29], achieving outstanding performance in a number of areas not limited to NLP. Furthermore, Mikolov et al. [19] proposed the skip-gram strategy to calculate the word embeddings from a large unlabelled corpus. We adopt the negative sampling (NEG) method to improve training speed and word embedding quality [30].

In the context of biomedical NER, word embedding methods are highly popular. For example, Pyysalo et al. [20] trained the word embedding model on the complete text of PubMed Central Open Access and the PubMed abstract datasets; moreover, the related resources had been released for biological NLP tasks. Habibi et al. [31] used 24 corpora for biomedical NER method evaluation, indicating that the method using word embedding is superior. Chiu et al. [32] used different corpora and word embedding training hyperparameters including sampling rate, learning rate, vector dimension, and context window size for evaluating the performance of word embedding. To obtain word embeddings more relevant to biomedicine, we have collected a total of 22,120,000 abstract records from the PubMed website and 672,000 full texts from PMC OA, based on our previous work [24]. Furthermore, we use all the collected data to train word embedding, which obtains a corresponding vector for each word; moreover, we apply the skip-gram model with a window size of five, and the dimension is set to 200.

2.3. Disease dictionary matching

2.3.1. Disease dictionary constructing

The available disease dictionary resources contain a large number of common and rare disease names; these can be effective in the context of disease NER. The construction of a disease dictionary can aid the attention model in capturing the disease named entities in the text more effectively; thereby, it can improve the performance of the DABLC disease NER method. To cover the maximum feasible number of disease named entities, it is necessary to construct the disease dictionary from a variety of resources. Disease Ontology (DO), provided by Lynn et al. [33], is an open source ontology that is associated with human disease. DO integrates a variety biomedical resources including Medical Subject Headings (MeSH) [34], Online Mendelian Inheritance in Man (OMIM) [35], SNOMED-CT [36], Classification of Diseases (ICD) [37], and NCI thesaurus [38] through extensive cross mapping.

We use the Jan 2019 version of DO [39] to construct the disease dictionary. There are three types of disease terms: disease name, exact synonym of the disease, and related synonym of the disease; e.g., the disease "Flinders Island spotted fever" (DOID 0050047) has an exact synonym 'Thai tick typhus' and a related synonym 'FISF'. We retrieve all the disease terms in the file of DO and obtain 11,142 disease terms, 17,167 exact synonyms, and 495 related synonyms. Furthermore, we remove the duplicate disease terms. Finally, we obtain 28,754 unique disease terms for the disease dictionary.

2.3.2. Matching method

In the context of biomedical NER, dictionary matching methods have been widely-used; moreover, named entities in different fields generally use the corresponding biomedical resources. Korkontzelos et al. [40] combined dictionary knowledge to capture common drug suffixes. Deléger et al. [41] identified concepts from the chemical and disease knowledge map database. Leaman et al. [2] used the MEDIC for disease NER based on pairwise learning to rank algorithm. Dogan et al. [42] mapped disease names to the corresponding concepts in MeSH and OMIM by using the synonym string similarity method. Demner-Fushman et al. [43] proposed MetaMap Lite; it focused on the speed of real-time processing and then mapped named entities to UMLS.

To implement an efficient dictionary matching method, we adopt the Trie dictionary data structure [44] to store the integrated medical resource dictionary; here, the matching method applies the Aho-Corasick algorithm [45]. The Trie tree, also called prefix tree, is a multi-fork tree structure that can efficiently store dictionaries for rapid searching. We use an associative array to store the characters of each word for implementing the Trie tree storage. Considering the query speed, we do not perform sub-strings matching. As shown in Figure 2, the root and leaf nodes of the Trie tree are the beginning and end, respectively, of the word boundary.

In particular, the Aho-Corasick algorithm is an extension of the Knuth—Morris–Pratt (KMP) string lookup algorithm [46] in the case of multi-pattern matching. The Aho-Corasick algorithm uses valid strings that have been partially matched previously; this prevents query traceback. Because the Aho-Corasick algorithm shifts the pattern strings to valid positions, it circumvents the need for the re-examination of previously matched characters. Unlike the Aho-Corasick algorithm, we match the complete word without matching the substring. Assume that the size of the query text is N and that the dictionary size is M. The time complexity of the entire word matching method is $O(N)$, which is low compared to that for substring matching ($O(MN)$); this improves the query efficiency.

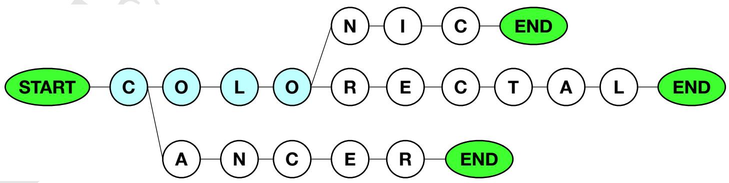

Figure 2: Example of data structure of a Trie tree. Three words are stored by a Trie tree: colonic, colorectal, and cancer. The green parts are the root node and leaf nodes of the Trie tree; these are the beginning and end, respectively, of the word boundary. The blue parts are identical characters among different words.

2.4. Dic-Att-BiLSTM-CRF method

2.4.1. BiLSTM-CRF

In the context of NER, the long short-term memory networks (LSTM) [47] have been widely-used to extract contextual text features. Lample et al. [48] proposed a model by combining bidirectional LSTM (BiLSTM) and conditional random fields (CRF) [7] for the NER task. BiLSTM was used to capture context information for words, and CRF was used to calculate optimal sequence combinations. The method was subsequently applied to disease NER [18] and biomedical NER [21] tasks. The following are details of BiLSTM and CRF.

BiLSTM can learn forward and backward information of input words, which aids in entity classification. More specifically, given a sentence X consisting of N words, which is represented as a set of vectors $(x_1, x_2, ..., x_n)$, a typical structure of an LSTM unit at each time t is calculated by the following formulas:

$$\begin{aligned} &i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\ &f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\ &c_t = (1 - i_t) \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ &o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\ &h_t = o_t \odot \tanh(c_t) \end{aligned} \tag{1}$$

where x_t, h_{t-1}, and c_{t-1} are inputs and h_t and c_t are outputs. i_t, o_t, and f_t are the gates of input, output, and forget, respectively. c_t is a memory cell vector. W_i, W_f, W_c, and W_o, with subscripts x, h, and c, are the weight matrices of the input word representation x_t, hidden state h_t, and memory cell c_t, respectively. b_i, b_f, b_c, and b_o denote the bias vectors. $\odot$ is the element-wise product. σ is the element-wise sigmoid function.

$\vec{y}_t$ and $\overleftarrow{y}_t$ are calculated by Equation (1); these represent the preceding and following words, respectively, of the word t. Furthermore, we concatenate $\vec{y}_t$ and $\overleftarrow{y}_t$ as L_t to represent the context information for the word t. L_t is calculated by Equation (2):

$$L_t = \tanh(W_{t_l}\vec{y}_t + W_{t_r}\overleftarrow{y}_t + b_t) \tag{2}$$

CRF [7] is a sequence tag model; it can compute the global optimal sequence. The tags sequences are related, e.g., an I-PER cannot appear after a B-LOC tag. For a sequence of predictions $(y_1, y_2, ..., y_n)$, we use the CRF layer to calculate the optimal sequence of tags. $T_{y_i,y_{i+1}}$ are the transition scores from the tag i to tag $i+1$. P_{i,y_i} is the output score matrix of the BiLSTM network. The score $s(X, y)$ is used to summarise the prediction score at the sentence-level. The sum of the scores from the transition scores and the BiLSTM network is as expressed in Equation (3):

$$s(X, y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=0}^{n} T_{y_i,y_{i+1}} \tag{3}$$

Then, we calculate the point estimate y^* of all the feasible outputs y such that the conditional log-likelihood probability softmax(s(X, y)) is maximised; this is illustrated in Equation (4):

$$y^* = \arg\max(\log(\text{softmax}(\text{s}(X, y)))) \tag{4}$$

2.4.2. Attention mechanism

Recently, the attention mechanism has been widely-used in the NLP field. Bahdanau et al. [49] and Luong et al. [50] used the attention mechanism in neural machine translation. Raffel et al. [51] demonstrated that the attention model could solve long-term memory problems of the long sequences. Yang et al. [52] used a hierarchical attention network for document classification tasks. The attention mechanism can effectively improve the performance of NER methods. There are two levels in terms of document processing, i.e., sentence-level and document-level. The sentence-level approaches primarily focus on sentence-level attention [53] [54] [55]; meanwhile, the document-level attention mechanism focuses on related entities in all the sentences of the documents, so that the label inconsistency problem can be handled. Luo et al. [17] and Xu et al. [56] used the document-level attention mechanism to address chemical and clinical NER tasks, respectively.

For a specified document $F = (h_1, ..., h_t, ..., h_N)$ consisting of N words, the document is represented by the words with their embeddings; this was introduced in Section 2.2. $h_i = (h_{i1}, ..., h_{in})$ and $h_j = (h_{j1}, ..., h_{jn})$ are n-dimensional vectors. We can calculate the distance between the current *i-th* word h_i and the *j-th* word h_j in the document by incorporating different distance measurements such as the Euclidean distance, Manhattan distance, and Cosine distance in the following equation:

$$dis(h_i, h_j) = \begin{cases} \sqrt{\sum\limits_{k=1}^{n} (h_{ik} - h_{jk})^2}, \\ \sum\limits_{k=1}^{n} |h_{ik} - h_{jk}|, \\ \dfrac{\sum\limits_{k=1}^{n} h_{ik} h_{jk}}{\sqrt{\sum\limits_{k=1}^{n} (h_{ik})^2}\sqrt{\sum\limits_{k=1}^{n} (h_{jk})^2}}. \end{cases} \tag{5}$$

Then, we adopt Equation (6) or Equation (7) to obtain the attention $score(h_i, h_j)$. For the Euclidean distance and Manhattan distance, we adopt Equation (6) to calculate the scores. $V_{\max}$ is the maximum score of the current word h_i compared with all the words in the entire document. Intuitively, a larger score indicates that the two vectors are more similar. For the Cosine distance, we adopt Equation (7) to calculate scores. The weight matrix W_A is the parameter to be learned in the training process.

$$score(h_i, h_j) = W_A \left(V_{\max} - dis(h_i, h_j)\right) \tag{6}$$

$$score(h_i, h_j) = W_A dis(h_i, h_j) \tag{7}$$

In the document-level attention layer, let A denote the weight vector and $A_{i,j}$ denote the similarity weight value between the current *i-th* word h_i and the *j-th* word h_j in the document. We can calculate $A_{i,j}$ by a softmax function in Equation (8):

$$A_{i,j} = \text{softmax}(score(h_i, h_j)) = \frac{\exp(score(h_i, h_j))}{\sum_k \exp(score(h_i, h_k))} \tag{8}$$

2.4.3. Dic-Att-BiLSTM-CRF approach

The attention model can address label inconsistency. However, the attention model described in Section 2.4.2 exhibits shortcomings in the identification of rare disease mentions that are absent in the training datasets; moreover, it is challenging to identify the disease mentions of long length. Therefore, we propose the Dic-Att-BiLSTM-CRF to incorporate both disease dictionary and Att-BiLSTM-CRF. Dic-Att-BiLSTM-CRF matches the disease entities in the document with the disease dictionary and assigns the matched entities with a certain weight; this weight is fused together with the attention weight. Dic-Att-BiLSTM-CRF can identify both rare disease mentions and disease mentions in long length.

More specifically, Dic-Att-BiLSTM-CRF consists of four layers (Figure 3). The first layer is the word embeddings layer. The word embeddings are trained through biomedical resources containing semantic information of words. The second layer is the BiLSTM layer, which is aimed at learning forward and backward information of input words. The third layer is the dictionary attention layer; it combines the disease dictionary matching method and attention mechanism. The fourth layer is the CRF layer; it is used to compute the global optimal sequence.

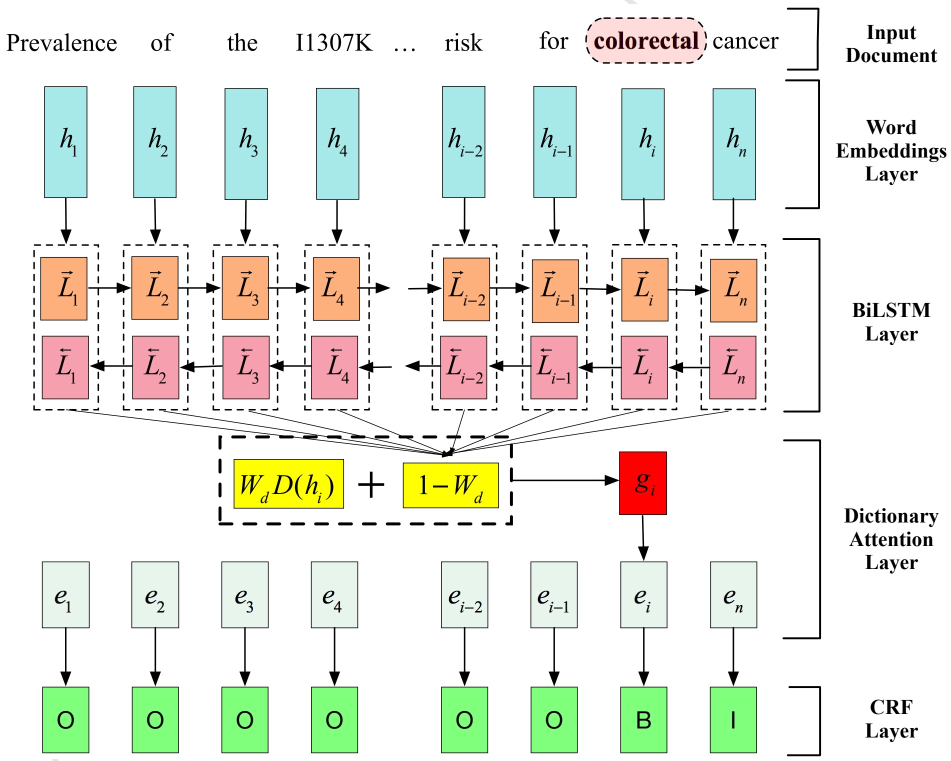

Figure 3: Detailed architecture of Dic-Att-BiLSTM-CRF (DABLC) method. DABLC consists of four layers: word embedding layer, BiLSTM layer (learning forward and backward information of the input words), dictionary attention layer (incorporating a disease dictionary matching and attention mechanism), and CRF layer.

For a document $F = (H_1, ..., H_t, ..., H_m)$ consisting of m sentences, each sentence $H = (h_1, ..., h_t, ..., h_n)$ consists of n words, and the number of words in the document is N. In terms of the dictionary attention layer, we introduce a dictionary-based attention weight vector A^D; it is different from the attention mechanism introduced in Section 2.4.2.

Firstly, $D(h_i)$ is a dictionary matching function as shown in Equation (9); its value is one if the word h_i is matched with the disease dictionary, whereas it is zero if they are unmatched.

$$D(h_i) = \begin{cases} 1, matched \\ 0, unmatched \end{cases} \tag{9}$$

Furthermore, we obtain the score $S(h_i, h_j)$ by calculating the weighted summation of the disease dictionary matching method and attention mechanism as shown in Equation (10); here, $score(h_i, h_j)$ is obtained by Equation (6) or Equation (7), W_d is the weight of dictionary matching function and $1 - W_d$ is the weight of the attention mechanism.

$$S(h_i, h_j) = W_d D(h_i) + (1 - W_d) score(h_i, h_j) \tag{10}$$

Let $A^D_{i,j}$ denote the similarity weight value vector of the dictionary-based attention method; it is calculated by the softmax function in Equation (11):

$$A^D_{i,j} = \frac{\exp(S(h_i, h_j))}{\sum_k \exp(S(h_i, h_k))} \tag{11}$$

Then, a dictionary-based global document vector g_i is calculated as the weighted average of $A^D_{i,j}$ and L_j (L_j is the BiLSTM output calculated by Equation (2)) as shown in Equation (12):

$$g_i = \sum_{j=1}^{N} A^D_{i,j} L_j \tag{12}$$

Subsequently, g_i and L_j are concatenated as a vector $[g_i; L_i]$, and a tanh function accepts the vector to produce the output z_i; here, W_g is a weight matrix learned in the training stage, as shown in Equation (13):

$$z_i = \tanh(W_g[g_i; L_i]) \tag{13}$$

Finally, a tanh function is used to calculate the confidence scores on the dictionary-based attention layer. The output score of the network is the probability score of each word, as shown in Equation (14); here, W_e is also a weight matrix learned during training:

$$e_i = \tanh(W_e z_i) \tag{14}$$

With respect to the final (CRF) layer, CRF is used to calculate the most effective tag path for preventing independent tagging, among all the feasible ones. For the disease NER task, we use the BIO method for each token labelling. Each token is tagged with one of the three labels, i.e., B, I or O; these indicate a token as being at the beginning, middle, or end, respectively, of a disease named entity. Let P denote the matrix of scores predicted by the Dic-Att-BiLSTM, and let the *i-th* column of P be a vector e_i, which is calculated by Equation (14). An entry $P_{i,j}$ is the score of the *j-th* tag of the *i-th* word in each sentence. Let T denote a tagging transition matrix, entry $T_{i,j}$ be the score of the transition from tag i to tag j and $T_{0,j}$ be an initial starting score of tag j. The transition matrix T is the parameter that the CRF model needs to learn in the training stage. Let $X = (h_1, ..., h_t, ..., h_n)$ denote a sentence, where n is the number of words in the sentence and m is the number of sentences in the document. Let $y = (y_1, ..., y_t, ..., y_n)$ denote the sequence of tag predictions of sentence X. Unlike Equation (3), we calculate the predicted score $s(X, y)$ in Equation (15) by summing the Dic-Att-BiLSTM scores and transition scores at the document-level. The probability of sequence y from all the feasible label sequences is calculated by Equation (4).

$$s(X,y) = \sum_{m} \left(\sum_{i=1}^{n} P_{i,y_i} + \sum_{i=0}^{n} T_{y_i,y_{i+1}} \right) \tag{15}$$

3. Results

3.1. Experimental datasets

To promote disease NER system research, American National Institutes of Health and BioCreAtIvE [57] released NCBI disease corpus [1] and BioCreative V CDR Challenge [26] [58] for disease NER research. We use these two widely-used disease corpora to evaluate the proposed DABLC method.

The NCBI disease corpus is large-scale and high-quality; it is based on the corpus released by Leaman et al. [59]. The NCBI disease corpus includes 14 experienced annotators for annotating. Each annotation in the corpus is completed by at least two annotators, ensuring the authority and accuracy of the annotations. The NCBI disease corpus contains 793 PubMed abstracts and 6,892 disease mentions, which are mapped to 2,136 unique disease mentions in the MEDIC. Table 1 presents the statistical information on the NCBI disease corpus.

The BioCreative V CDR corpus contains 1,500 PubMed articles with 12,850 disease mentions and 5,818 unique disease mentions, making it larger than the NCBI disease corpus. In the context of disease NER, we retain the disease entity labels and remove the chemical entity labels in the datasets. In terms of the disease named entity recognition task, four annotators with professional annotating backgrounds use MeSH [34] as a regulated vocabulary for annotation. Each article is annotated independently by two annotators, and the annotation result is finally determined by a high-level annotator. Table 2 presents the statistical information on the BioCreative V CDR corpus.

Table 1: Statistics of NCBI corpus.

Characteristics	Training	Development	Testing	Total
No. of PubMed article abstracts	593	100	100	793
No. of disease mentions	5145	787	960	6892
No. of unique disease mentions	1710	368	427	2136
Avg. sentences per abstract	10	10	10	10
Avg. words per sentence	20	22	22	21
Avg. words per abstract	217	226	232	225

Table 2: Statistics of BioCreative V CDR corpus.

Characteristics	Training	Development	Testing	Total
No. of articles	500	500	500	1500
No. of disease mentions	4182	4244	4424	12850
No. of unique disease mentions	1965	1865	1988	5818
Avg. sentences per article	11	11	11	11
Avg. words per sentence	26	26	25	26
Avg. words per article	305	309	306	307

3.2. Experimental settings

Similar to LeadMine [6] and RN+lm [25], we merge the development set and the training set. We apply each test over 10 executions of each model with a 95% confidence interval. The values in all the figures are the means of the experimental results. We randomly select 20% data from the training set to learn the optimal hyperparameter combination of the DABLC method; thereby, we obtain a list of optimised parameter values during the parameter tuning process (Table 3). We combine the word embedding training dataset with the PubMed abstract records and PMC OA full texts; the number of word embedding dimension is set as 200. We use a bidirectional LSTM network whose hidden-layer dimension is set as 100.

For the computational optimisation of neural network, we adopt the Adam algorithm [60]. Adam algorithm exhibits invariant gradient diagonal rescaling and a higher computational efficiency; this makes it more suitable for solving large-scale parameters. To obtain the highest F1 score performance, we conduct iterative training. If the number of iterations exceeds 20 and the F1 score does not increase further, it can be considered that the training is converged and need to be terminated. Once the highest F1 score on the development set is obtained, we can retain the parameters of the model to evaluate its performance on the test sets of the two corpora.

Table 3: Optimised parameter settings of the method.

Parameter	Setting	Description
Training Data	PubMed+PMC	Word embedding training dataset
Word_dim	200	Word embedding dimensions
LSTM_dim	100	LSTM hidden layer dimension
Bi-LSTM	TRUE	Using bidirectional LSTM
Learning method	Adam	Adam optimisation

We use three performance metrics for the evaluation: precision, recall, and F1 score. More specifically, precision is the proportion of retrieved disease instances that are relevant, recall is the proportion of relevant disease instances that are retrieved and F1 score is the harmonic mean of the precision and recall. For the evaluation, there are four likely conditions for a disease instance. True positive (TP) results imply the number of true disease instances classified as diseases. False positive (FP) results imply the number of non-disease instances classified as diseases. False negative (FN) results imply the number of true disease instances classified as non-diseases. True negative (TN) results imply the number of non-disease instances classified as non-diseases. Based on these conditions, precision, recall, and F1 score are defined in Equations (16) (17) (18), respectively:

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{18}$$

3.3. Evaluations of different alignment functions

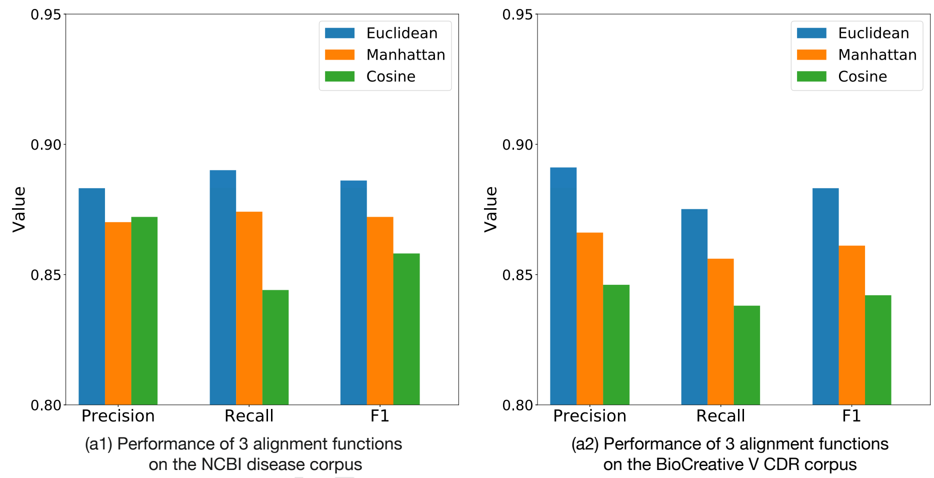

Figure 4: Performance of DABLC using different alignment functions on the NCBI disease corpus and BioCreative V CDR corpus.

In terms of distance measurement, we consider three popular ones as an example to evaluate the impact of using different measurements on the performance; these include the Euclidean distance, Manhattan distance, and Cosine distance (Equation (5)). We use the control variable method to test the performance of different alignment functions on the two corpora separately by fixing the optimal parameters. The word embedding vectors that contain biomedical semantics are inputted into the alignment function. The distance of different words is calculated by the alignment function. The performance of the alignment functions on the NCBI disease corpus and BioCreative V CDR Corpus are shown in Figure 4.

The figure shows that the highest F1 scores on the two datasets are obtained by the approach using the Euclidean distance; the scores are 0.886 and 0.883, respectively. The experimental results demonstrate that Euclidean distance is more effective than the other two distances for the considered tasks.

3.4. Evaluations of dictionary weights

Different dictionary weights are likely to influence the performance of the proposed DABLC. To determine the most effective dictionary weight, we conduct a series of experiments on different parameters. The control variable method is also used in this process; i.e., we change only the weight of the dictionary from low to high values at increments of 0.05 while keeping the other parameters fixed. We test the two corpora: NCBI disease corpus and BioCreative V CDR corpus. The performance while using different dictionary weights is shown in Figure 5.

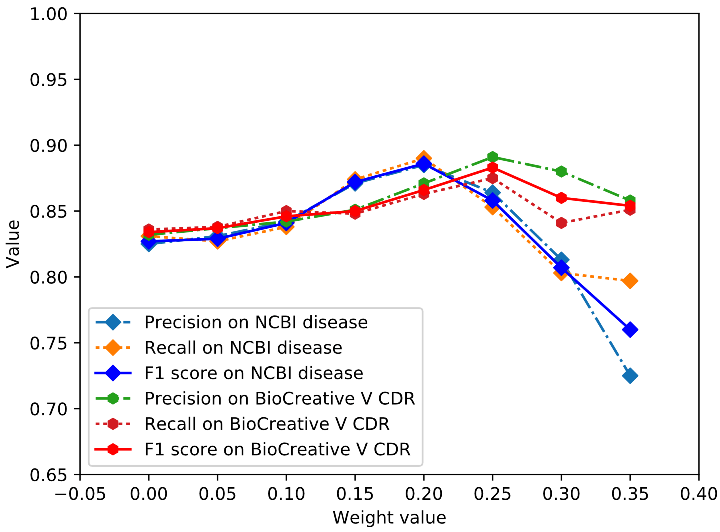

Figure 5: Performance of DABLC using different dictionary weights on NCBI disease corpus and BioCreative V CDR corpus.

The Figure 5 shows that the performance of DABLC first increases and then decrease with the increase in the dictionary weights, on both the corpora. DABLC achieves the highest performance by setting the dictionary weight as 0.2 (i.e., the attention weight as 0.8) on the NCBI disease corpus, which is 0.886 in terms of the F1 score. In terms of the BioCreative V CDR corpus, the highest performance of DABLC is 0.883 in terms of the F1 score, with the dictionary weight set as 0.25 (i.e., the attention weight as 0.75). A weight of zero denotes that DABLC uses only the attention weight, whereas a weight of one denotes that DABLC uses only the dictionary weight. The experimental results demonstrate that the use of the combinations of dictionary weight and attention weight achieves higher performances compared to cases using either one of them.

3.5. Comparisons of dictionary-based methods and non-dictionary-based methods

To verify the effectiveness of introducing dictionaries, we evaluate the performance achieved using dictionary-based and non-dictionary-based attention models. Unlike available disease identification methods, we integrate the external disease dictionary resources into the most update attention model to adapt it to disease NER; this aids the identification of rare and complex disease named entities. In particular, we adopt a document-level attention method, which can effectively alleviate the problem of tagging inconsistency [17] compared with the sentence-based attention methods. Next, we investigate the effectiveness of using sentence-level and document-level attention methods in combination with the dictionary.

For clarity, let ABLC (doc lev) and ABLC (sen lev) denote the proposed DABLC without dictionary, albeit using document-level and sentence-level attention, respectively. Let DABLC (doc lev) and DABLC (sen lev) denote the proposed DABLC with dictionary, albeit using document-level and sentence-level attention, respectively. The performance of the approaches on the two corpora is summarised in Table 4. The training datasets of the two corpora were used for training respectively. Each row of Table 4 corresponds to the same method for the two corpora. On the one hand, we observe that DABLC (doc lev) outperforms ABLC (doc lev) on both the corpora in terms of all the performance metrics. The document-level DABLC (doc lev) method achieves higher performance than the document-level ABLC (doc lev) method, with F1 scores increased by 0.046 and 0.045, respectively. The experimental results demonstrate that the dictionary-integrated DABLC method could identify more disease named entities; this indicates the effectiveness and necessity of introducing dictionary for the current tasks. On the other hand, we observe that DABLC (doc lev) outperforms DABLC (sen lev) on both the corpora in terms of the metrics; this is because the document-level attention method can solve the tagging inconsistency problem compared with sentence-level attention method. The experimental results demonstrate the significance of using the document-level attention method. Overall, DABLC (doc lev) achieves the highest performance on the two corpora; the F1 scores are 0.886 and 0.883, respectively.

Table 4: Performance of DABLC and ABLC methods with document-level and sentence-level attention. (mean ± 95% confidence interval)

	NCBI disease			BioCreative V CDR		
Methods	Precision	Recall	F1	Precision	Recall	F1
ABLC (sen lev)	0.804±0.002	0.801±0.001	0.803±0.001	0.811±0.001	0.807±0.002	0.809±0.001
DABLC (sen lev)	0.855±0.001	0.858±0.001	0.856±0.001	0.844±0.002	0.853±0.001	0.849±0.001
ABLC (doc lev)	0.842±0.002	0.839±0.001	0.84±0.001	0.845±0.001	0.832±0.001	0.838±0.001
DABLC (doc lev)	**0.883**±0.002	**0.89**±0.002	**0.886**±0.001	**0.891**±0.001	**0.875**±0.001	**0.883**±0.001

We provide an example to visualise the predicted results of DABLC and ABLC on the two corpora (Figure 6). The x-axis represents the predicted labels, and the y-axis represents the original input text. The grayscale blocks in the coordinate indicate the ground-truth label, and the darkness of a block indicates the predicted probability value on the corresponding ground-truth label for an input word. Intuitively, the darker the block is, the larger will be the predicted probability value. For illustrations, we select the sentence *'X-linked adrenoleukodystrophy ALD is an inherited disease characterised by progressive neurologic dysfunction'* as an example for the NCBI disease corpus. For the BioCreative V CDR corpus, we consider the sentence *'Protective effect of verapamil on gastric haemorrhagic ulcers in severe atherosclerotic rats'* as an example. The figure shows that the DABLC method is superior to the ABLC method for word label predictions with larger confidence.

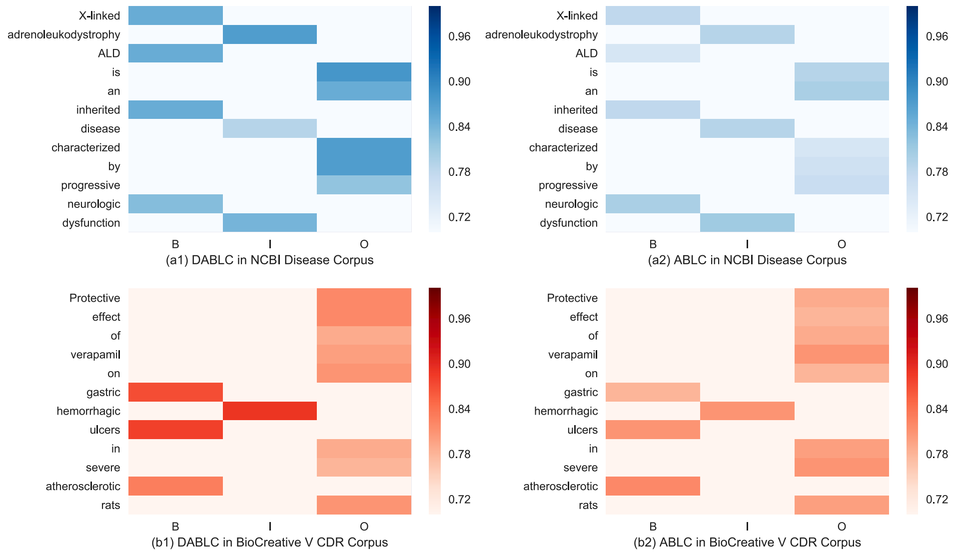

Figure 6: Performance of DABLC and ABLC on NCBI disease corpus and BioCreative V CDR corpus. The x-axis of each graph represents the predicted labels, and the y-axis represents the original text of the input. (a1) and (a2) show the performance of DABLC and ABLC, respectively, on the NCBI disease corpus. (b1) and (b2) show the performance of DABLC and ABLC on the BioCreative V CDR corpus, respectively. Each block shows the probability of annotating each input with its true label.

3.6. Performance comparison with other typical methods

To further demonstrate the effectiveness of our approach, we compare the performance of the DABLC method with nine state-of-the-art methods on the two corpora. More specifically, the dictionary look-up method [1] [58] used the SPECIALIST lexical tool in the MEDIC dictionary to identify the disease name. cTAKES [27] used UMLS to map the disease entities to the terms. Dnorm [2] was released by the National Center for Biotechnology Information (NIH) to calculate synonyms between concepts based on pairwise learning rankings (pLTR). C-BiLSTM-CRF [22] was based on character embedding to learn character-level word expression in combination with BiLSTM-CRF, to identify disease entities. TaggerOne [14] used a semi-Markov linear classifier while performing normalisation and NER during training and prediction. DNER [18] was based on CRF and Bi-RNN combined with a support vector machine classifier to identify disease named entities, and map normalisation based on dictionary matching of MeSH. LeadMine [6] used grammar rules and dictionary searching methods to correct spelling errors by linking entities to a specified grammar and dictionary. AuDis [11] used a robust CRF-based recognition model to identify important features associated with diseases and combined post-processing and dictionary lookup methods to improve the performance. RN+lm [25] proposed beam search and online structured learning to jointly perform disease named entity recognition and normalisation, permitting the use of non-local features to improve the performance of the method. SBLC [24] systematically combined word embedding, biLSTM and CRF for disease NER tasks, and it integrated Ab3P to identify disease abbreviations.

We further analyse the functional characteristics of the aforementioned methods in Table 5 using ‘Dictionary look-up’, ‘Disease name normalisation’, ‘Word embedding’, ‘Deep learning’, and ‘CRF’. ‘Y’ indicates that the method contained the corresponding character, whereas ‘N’ indicates that the method did not contain it. As illustrated in the table, most methods use the disease name normalisation method; moreover, more than half of them use CRF. DABLC, C-BiLSTM-CRF, and DNER also use deep learning techniques.

Table 5: Feature comparison of different baselines.

Methods	Dictionary look-up	Disease name normalisation	Word embedding	Deep learning	CRF
Dictionary look-up	Y	Y	N	N	N
cTAKES	Y	Y	N	N	Y
DNorm	Y	Y	N	N	N
C-BiLSTM-CRF	N	N	Y	Y	Y
TaggerOne	N	Y	N	N	N
DNER	N	Y	N	Y	Y
LeadMine	Y	Y	N	N	N
AuDis	Y	Y	N	N	Y
RN+lm	Y	Y	N	N	N
SBLC	N	N	Y	Y	Y
DABLC	Y	N	Y	Y	Y

The performance results of the baselines and our proposed DABLC are presented in Table 6. On both the corpora, the performance of the dictionary look-up method and cTAKES are relatively low; this reveals that it is not effective the adopt only of the strategy of dictionary look-up is in effective. Dnorm and TaggerOne, released by NIH, obtain higher performance; they benefit from the machine learning methods trained on the features of entities. C-BiLSTM-CRF and DNER use deep learning techniques, outperforming Dnorm and TaggerOne. Deep neural networks can learn more effective features of texts, compared with traditional machine learning techniques. LeadMine, AuDis and RN+lm use the complex normalisation method; moreover, they combine the complicated manual-setting grammar rules and exploit CRF and other machine learning technology to identify entities. The three methods acquire the advantages of multiple methods, achieving relatively high performance. However, they are customised and optimised on the particular training dataset; this is likely to be unsuitable for transferring them to other datasets. Our proposed DABLC method achieves the highest F1 scores (0.886 and 0.883) and relatively high performance on the two corpora. DABLC completely uses the external dictionary resources, the document-level attention method and the advantages of BiLSTM-CRF, to identify disease entities; this also circumvents manual feature engineering.

Table 6: Performance of our DABLC and baselines on the two corpora.

	NCBI disease			BioCreative V CDR		
Method	Precision	Recall	F1	Precision	Recall	F1
Dictionary look-up	0.213	0.718	0.316	0.427	0.675	0.523
cTAKES	0.476	0.541	0.506	0.513	0.552	0.532
DNorm	0.822	0.775	0.798	0.812	0.801	0.806
C-BiLSTM-CRF	0.848	0.761	0.802	—	—	—
TaggerOne	0.835	0.796	0.815	0.846	0.827	0.837
DNER	—	—	—	0.853	0.833	0.843
LeadMine	—	—	—	0.861	0.862	0.861
AuDis	—	—	—	**0.896**	0.835	0.865
RN+lm	**0.887**	0.773	0.826	**0.896**	0.857	0.876
SBLC	0.866	0.858	0.862	—	—	—
DABLC	0.883	**0.89**	**0.886**	0.891	**0.875**	**0.883**

3.7. Discussion

Compared to the baselines, the DABLC method identifies rare and longer complex diseases more effectively. Long disease names generally have over three words, e.g., the diseases 'glomerular basement membrane abnormalities' (PMID 9792860, D005921) and 'autosomal recessive Alport syndrome' (PMID 9792860, C536587). Rare and longer complex disease identification is a challenge in the context of disease NER. Our DABLC method introduces a disease dictionary matching method, which alleviates the problem posed by rare and complex disease NER.

First, we construct a disease dictionary that covers a large number of disease entities including rare and complex disease names. Secondly, we design an effective dictionary matching method to utilise the dictionary. Thirdly, we adopt the document-level attention mechanism to improve the performance of BiLSTM-CRF. Compared with the methods proposed by Luo et al. [17] and Xu et al. [56], we focus on the disease field; moreover, we elaborately integrate the dictionary matching technique into the attention model for recognising rare and complex disease names, thereby benefiting from the effective dictionary matching method and document-level attention mechanism. The revised DABLC method exhibits higher performance as demonstrated by the experiments on the NCBI and BioCreative V CDR datasets.

The proposed DABLC is trained on the NCBI and BioCreative V CDR disease corpora, benefiting from the dictionary matching algorithm; this enabled more accurate and comprehensive identification of disease name. The external dictionary contains rare and complex disease names, which aids disease NER. In particular, the integration of external dictionaries can provide a more accurate attention model. Experiments demonstrate that the dictionary matching method combined with the attention method improves the performance. We consider that a rich dictionary of diseases will play an important role in the tagging of complex and rare disease names.

Although our proposed DABLC achieves the highest performance for most of the cases in the experiments, a few shortcomings are likely to still exist. We examined the predicted results and observed that certain special disease names cannot be recognised correctly, e.g., disease names labelled in digits and letters, such as SCA2 (PMID 9506545, OMIM 183090), which are labelled as other entities rather than as disease named entities. This type of error may be prevented by adopting additional gene-based NER post-processing modules. Although the DABLC method can solve the label inconsistency problem, the position of the sixth occurrence of DM (PMID 9863607, D009223) is still not labelled successfully. This could be because of the high complexity of the context information at the location and because the LSTM model is vulnerable to the complex information. Designing a new LSTM network to address more complex long sentences is highly necessary.

4. Conclusion

Disease NER is an important and early step in the processing of biomedical information. In this work, we propose a novel dictionary-based and document-level attention mechanism with a deep neural network NER method, named as DABLC. The proposed DABLC tags the consistency of multiple instances in a document at the document level. DABLC combines an external disease dictionary that is constructed with five disease resources containing a rich collection of disease entities. We adopt the efficient exact string matching method for dictionary matching; this method can effectively and accurately match the disease names.

For evaluations, we compare DABLC with nine advanced methods on the widely-used NCBI disease corpus and BioCreative V CDR corpus. The DABLC method achieves 0.886 and 0.883 in terms of the F1 score on the two corpora, outperforming the baselines. The experimental results demonstrate that our document-level attention mechanism can solve the problem of label inconsistency and identify complex entities effectively.

Finally, we have discussed the DABLC method and a few feasible future works. Special disease NER and deep learning models addressing long sentences are noteworthy areas for future exploration.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61703109, No. 91748107, No. 500160134), the China Postdoctoral Science Foundation (No. 2018M643024) and the Guangdong Innovative Research Team Program (No. 2014ZT05G157).

References

[1] R. I. Dogan, R. Leaman, Z. Lu, NCBI disease corpus: a resource for disease name recognition and concept normalization., Journal of biomedical informatics 47 (2014) 1–10. doi:10.1016/j.jbi.2013.12.006.

[2] R. Leaman, R. Islamaj Doğan, Z. Lu, DNorm: disease name normalization with pairwise learning to rank, Bioinformatics 29 (22) (2013) 2909–2917. doi:10.1093/bioinformatics/btt474.
URL https://academic.oup.com/bioinformatics/article/29/22/2909/312804

[3] Y.-F. Lin, T.-H. Tsai, W.-C. Chou, K.-P. Wu, T.-Y. Sung, W.-L. Hsu, A Maximum Entropy Approach to Biomedical Named Entity Recognition, in: Proceedings of the 4th International Conference on Data Mining in Bioinformatics, BIOKDD'04, Springer-Verlag, Berlin, Heidelberg, 2004, pp. 56–61.
URL `http://dl.acm.org/citation.cfm?id=3000580.3000588`

[4] A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, D. Rebholz-Schuhmann, Assessment of disease named entity recognition on a corpus of annotated sentences, BMC Bioinformatics 9 (3) (2008) S3. `doi:10.1186/1471-2105-9-S3-S3`.
URL `https://doi.org/10.1186/1471-2105-9-S3-S3`

[5] A. R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program., in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17.
URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243666/`

[6] D. M. Lowe, R. A. Sayle, LeadMine: a grammar and dictionary driven approach to entity recognition., Journal of cheminformatics 7 (Suppl 1 Text mining for chemistry and the CHEMDNER track) (2015) S5. `doi:10.1186/1758-2946-7-S1-S5`.

[7] J. Lafferty, A. McCallum, F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, 2001, pp. 282–289.
URL `http://dl.acm.org/citation.cfm?id=645530.655813`

[8] C. Sun, Y. Guan, X. Wang, L. Lin, Rich features based Conditional Random Fields for biological named entities recognition, Computers in Biology and Medicine 37 (9) (2007) 1327–1333. `doi:10.1016/j.compbiomed.2006.12.002`.
URL `http://www.sciencedirect.com/science/article/pii/S0010482506002332`

[9] W. Lee, K. Kim, E. Y. Lee, J. Choi, Conditional random fields for clinical named entity recognition: A comparative study using Korean clinical texts, Computers in Biology and Medicine 101 (2018) 7–14. `doi:10.1016/j.compbiomed.2018.07.019`.
URL `http://www.sciencedirect.com/science/article/pii/S0010482518302105`

[10] R. Leaman, G. Gonzalez, BANNER: An executable survey of advances in biomedical named entity recognition, 2008.
URL `http://www.scopus.com/inward/citedby.url?scp=40549140499&partnerID=8YFLogxK`

[11] H.-C. Lee, Y.-Y. Hsu, H.-Y. Kao, AuDis: an automatic CRF-enhanced disease normalization in biomedical text, Database 2016. `doi:10.1093/database/baw091`.
URL `https://academic.oup.com/database/article/doi/10.1093/database/baw091/2630473`

[12] D. Campos, S. Matos, J. L. Oliveira, A modular framework for biomedical concept recognition, BMC Bioinformatics 14 (1) (2013) 281. `doi:10.1186/1471-2105-14-281`.
URL `https://doi.org/10.1186/1471-2105-14-281`

[13] A. P. Davis, T. C. Wiegers, M. C. Rosenstein, C. J. Mattingly, MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database, Database 2012. `doi:10.1093/database/bar065`.
URL `https://academic.oup.com/database/article/doi/10.1093/database/bar065/430135`

[14] R. Leaman, Z. Lu, TaggerOne: joint named entity recognition and normalization with semi-Markov Models., Bioinformatics 32 (18) (2016) 2839–2846. `doi:10.1093/bioinformatics/btw343`.

[15] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444. `doi:10.1038/nature14539`.
URL `https://www.nature.com/articles/nature14539`

[16] S. Santiso, A. Perez, A. Casillas, Exploring Joint AB-LSTM with embedded lemmas for Adverse Drug Reaction discovery, IEEE Journal of Biomedical and Health Informatics (2018) 1–1`doi:10.1109/JBHI.2018.2879744`.

[17] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, J. Wang, An attention-based BiLSTM-CRF approach to document-level chemical named entity, Bioinformatics (2017) btx761–btx761`doi:10.1093/bioinformatics/btx761`.
URL `http://dx.doi.org/10.1093/bioinformatics/btx761`

[18] Q. Wei, T. Chen, R. Xu, Y. He, L. Gui, Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks., Database : the journal of biological databases and curation 2016. `doi:10.1093/database/baw140`.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.
URL `http://dl.acm.org/citation.cfm?id=2999792.2999959`

[20] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, S. Ananiadou, Distributional Semantics Resources for Biomedical Text Processing, Tokyo, Japan, 2013, pp. 39–43.
URL `http://bio.nlplab.org/pdf/pyysalo13literature.pdf`

[21] M. Gridach, Character-Level Neural Network for Biomedical Named Entity Recognition, Journal of Biomedical Informatics.
URL `http://www.sciencedirect.com/science/article/pii/S1532046417300977`

[22] K. Xu, Z. Zhou, T. Hao, W. Liu, A bidirectional LSTM and conditional random fields approach to medical named entity recognition, in: Advances in Intelligent Systems and Computing, Vol. 639, Cairo, Egypt, 2017, pp. 355 – 365.
URL `http://dx.doi.org/10.1007/978-3-319-64861-3_33`

[23] L. Ratinov, D. Roth, Design Challenges and Misconceptions in Named Entity Recognition, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), Association for Computational Linguistics, Boulder, Colorado, 2009, pp. 147–155.
URL `http://www.aclweb.org/anthology/W09-1119`

[24] K. Xu, Z. Zhou, T. Gong, T. Hao, W. Liu, SBLC: a hybrid model for disease named entity recognition based on semantic bidirectional LSTMs and conditional random fields, BMC Medical Informatics and Decision Making 18 (5) (2018) 114. `doi:10.1186/s12911-018-0690-y`.
URL `https://doi.org/10.1186/s12911-018-0690-y`

[25] Y. Lou, Y. Zhang, T. Qian, F. Li, S. Xiong, D. Ji, A transition-based joint model for disease named entity recognition and normalization, Bioinformatics 33 (15) (2017) 2363–2371. `doi:10.1093/bioinformatics/btx172`.
URL `http://dx.doi.org/10.1093/bioinformatics/btx172`

[26] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, Z. Lu, BioCreative V CDR task corpus: a resource for chemical disease relation extraction., Database : the journal of biological databases and curation 2016. `doi:10.1093/database/baw068`.

[27] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, C. G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, Journal of the American Medical Informatics Association : JAMIA 17 (5) (2010) 507–513. `doi:10.1136/jamia.2009.001560`.
URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995668/`

[28] S. Bird, NLTK: the natural language toolkit, in: Proceedings of the COLING/ACL on Interactive presentation sessions, Association for Computational Linguistics, 2006, pp. 69–72.
URL `http://dl.acm.org/citation.cfm?id=1225421`

[29] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (8) (2013) 1798–1828.
URL `http://ieeexplore.ieee.org/abstract/document/6472238/`

[30] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.
URL `https://arxiv.org/abs/1301.3781`

[31] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, U. Leser, Deep learning with word embeddings improves biomedical named entity recognition, Bioinformatics 33 (14) (2017) i37–i48. `doi:10.1093/bioinformatics/btx228`.
URL `https://academic.oup.com/bioinformatics/article/33/14/i37/3953940`

[32] B. Chiu, G. Crichton, A. Korhonen, S. Pyysalo, How to Train good Word Embeddings for Biomedical NLP, in: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 166–174.
URL `http://anthology.aclweb.org/W16-2922`

[33] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, W. A. Kibbe, Disease Ontology: a backbone for disease semantic integration, Nucleic Acids Research 40 (D1) (2012) D940–D946. `doi:10.1093/nar/gkr972`.
URL `https://academic.oup.com/nar/article/40/D1/D940/2903651`

[34] C. E. Lipscomb, Medical Subject Headings (MeSH), Bulletin of the Medical Library Association 88 (3) (2000) 265–266.
URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC35238/`

[35] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic acids research 33 (suppl_1) (2005) D514–D517.
URL `https://academic.oup.com/nar/article-abstract/33/suppl_1/D514/2505259`

[36] P. L. Elkin, S. H. Brown, C. S. Husser, B. A. Bauer, D. Wahner-Roedler, S. T. Rosenbloom,

T. Speroff, Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists, Mayo Clinic Proceedings 81 (6) (2006) 741–748. doi:10.4065/81.6.741.
URL http://www.sciencedirect.com/science/article/pii/S002561961161728X

[37] S. Aymé, A. Rath, B. Bellet, WHO International Classification of Diseases (ICD) Revision Process: incorporating rare diseases into the classification scheme: state of art, Orphanet Journal of Rare Diseases 5 (1) (2010) P1. doi:10.1186/1750-1172-5-S1-P1.
URL https://doi.org/10.1186/1750-1172-5-S1-P1

[38] N. Sioutos, S. d. Coronado, M. W. Haber, F. W. Hartel, W.-L. Shaiu, L. W. Wright, NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information, Journal of Biomedical Informatics 40 (1) (2007) 30–43. doi:10.1016/j.jbi.2006.02.013.
URL http://www.sciencedirect.com/science/article/pii/S1532046406000311

[39] Repository for the Human Disease Ontology. (Jan. 2019).
URL https://github.com/DiseaseOntology/HumanDiseaseOntology

[40] I. Korkontzelos, D. Piliouras, A. W. Dowsey, S. Ananiadou, Boosting drug named entity recognition using an aggregate classifier, Artificial intelligence in medicine 65 (2) (2015) 145–153.
URL http://www.sciencedirect.com/science/article/pii/S0933365715000780

[41] E. Pons, B. F. H. Becker, S. A. Akhondi, Z. Afzal, E. M. van Mulligen, J. A. Kors, Extraction of chemical-induced diseases using prior knowledge and textual information., Database : the journal of biological databases and curation 2016. doi:10.1093/database/baw046.

[42] R. I. Dogan, Z. Lu, An inference method for disease name normalization, in: 2012 AAAI Fall Symposium Series, 2012.
URL http://www.aaai.org/ocs/index.php/FSS/FSS12/paper/viewPaper/5604

[43] D. Demner-Fushman, W. J. Rogers, A. R. Aronson, MetaMap Lite: an evaluation of a new Java implementation of MetaMap, Journal of the American Medical Informatics Association 24 (4) (2017) 841–844. doi:10.1093/jamia/ocw177.
URL https://academic.oup.com/jamia/article/24/4/841/2961848

[44] E. Fredkin, Trie Memory, Commun. ACM 3 (9) (1960) 490–499. doi:10.1145/367390.367400.
URL http://doi.acm.org/10.1145/367390.367400

[45] A. V. Aho, M. J. Corasick, Efficient String Matching: An Aid to Bibliographic Search, Commun. ACM 18 (6) (1975) 333–340. doi:10.1145/360825.360855.
URL http://doi.acm.org/10.1145/360825.360855

[46] D. Knuth, J. Morris, Jr., V. Pratt, Fast Pattern Matching in Strings, SIAM Journal on Computing 6 (2) (1977) 323–350. doi:10.1137/0206024.
URL https://doi.org/10.1137/0206024

[47] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.
URL http://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735

[48] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural Architectures for Named Entity Recognition, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270.
URL http://www.aclweb.org/anthology/N16-1030

[49] D. Bahdanau, K. Cho, Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, 2015, arXiv: 1409.0473.
URL http://arxiv.org/abs/1409.0473

[50] T. Luong, H. Pham, C. D. Manning, Effective Approaches to Attention-based Neural Machine Translation, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1412–1421.
URL http://aclweb.org/anthology/D15-1166

[51] C. Raffel, D. P. W. Ellis, Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems, in: ICLR 2016 workshop, 2016.
URL https://openreview.net/forum?id=81DD7ZNyxI6O2Pl0Ul5j

[52] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical Attention Networks for Document Classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1480–1489.
URL http://www.aclweb.org/anthology/N16-1174

[53] A. Bharadwaj, D. Mortensen, C. Dyer, J. Carbonell, Phonologically Aware Neural Model for Named

Entity Recognition in Low Resource Transfer Settings, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 1462–1472. doi:10.18653/v1/D16-1153.
URL http://www.aclweb.org/anthology/D16-1153

[54] M. Rei, G. Crichton, S. Pyysalo, Attending to Characters in Neural Sequence Labeling Models, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 309–318.
URL http://www.aclweb.org/anthology/C16-1030

[55] C. Pandey, Z. Ibrahim, H. Wu, E. Iqbal, R. Dobson, Improving RNN with Attention and Embedding for Adverse Drug Reactions, DH '17, ACM, New York, NY, USA, 2017, pp. 67–71. doi:10.1145/3079452.3079501.
URL http://doi.acm.org/10.1145/3079452.3079501

[56] G. Xu, C. Wang, X. He, Improving Clinical Named Entity Recognition with Global Neural Attention, in: Web and Big Data, Lecture Notes in Computer Science, Springer, Cham, 2018, pp. 264–279. doi:10.1007/978-3-319-96893-3_20.
URL https://link.springer.com/chapter/10.1007/978-3-319-96893-3_20

[57] L. Hirschman, A. Yeh, C. Blaschke, A. Valencia, Overview of BioCreAtIvE: critical assessment of information extraction for biology, BMC Bioinformatics 6 (1) (2005) S1. doi:10.1186/1471-2105-6-S1-S1.
URL https://doi.org/10.1186/1471-2105-6-S1-S1

[58] C.-H. Wei, Y. Peng, R. Leaman, A. P. Davis, C. J. Mattingly, J. Li, T. C. Wiegers, Z. Lu, Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task., Database : the journal of biological databases and curation 2016. doi:10.1093/database/baw032.

[59] R. Leaman, C. Miller, G. Gonzalez, Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark, in: Proceedings of the 2009 Symposium on Languages in Biology and Medicine, Vol. 82, 2009, p. 9.
URL https://books.google.com.sg/books/about/Enabling_Recognition_of_Disease_Mentions.html?id=Mf5AywAACAAJ&redir_esc=y

[60] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2014, arXiv: 1412.6980.
URL http://arxiv.org/abs/1412.6980

A dictionary of disease is constructed to identify rare and complex disease names

A document-level attention mechanism is used to solve tagging inconsistency

Our method performs better than other existing methods in F1 scores

ACCEPTED MANUSCRIPT