applied sciences

MDPI

Article

Fusion of SoftLexicon and RoBERTa for Purpose-Driven Electronic Medical Record Named Entity Recognition

Xiaohui Cui [1,2,†], Yu Yang [1,2,†], Dongmei Li [1,2,*], Xiaolong Qu [1,2], Lei Yao [3], Sisi Luo [1,2] and Chao Song [1,2]

[1] School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China; cuixiaohui@bjfu.edu.cn (X.C.); yangyu22@bjfu.edu.cn (Y.Y.); quxiaolong@bjfu.edu.cn (X.Q.); 17812069970@163.com (S.L.); songchao1025@bjfu.edu.cn (C.S.)

[2] Engineering Research Center for Forestry-Oriented Intelligent Information Processing of National Forestry and Grassland Administration, Beijing 100083, China

[3] College of Engineering and Applied Science, University of Wisconsin-Milwaukee, Milwaukee, WI 53202, USA; leiyaouwm@gmail.com

* Correspondence: lidongmei@bjfu.edu.cn

† These authors contributed equally to this work.

Abstract: Recently, researchers have extensively explored various methods for electronic medical record named entity recognition, including character-based, word-based, and hybrid methods. Nonetheless, these methods frequently disregard the semantic context of entities within electronic medical records, leading to the creation of subpar-quality clinical knowledge bases and obstructing the discovery of clinical knowledge. In response to these challenges, we propose a novel purpose-driven SoftLexicon-RoBERTa-BiLSTM-CRF (SLRBC) model for electronic medical records named entity recognition. SLRBC leverages the fusion of SoftLexicon and RoBERTa to incorporate the word lexicon information from electronic medical records into the character representations, enhancing the model's semantic embedding representations. This purpose-driven approach helps achieve a more comprehensive representation and avoid common segmentation errors, consequently boosting the accuracy of entity recognition. Furthermore, we employ the classical BiLSTM-CRF framework to capture contextual information of entities more effectively. In order to assess the performance of SLRBC, a series of experiments on the public datasets of CCKS2018 and CCKS2019 were conducted. The experimental results demonstrate that SLRBC can efficiently extract entities from Chinese electronic medical records. The model attains F1 scores of 94.97% and 85.40% on CCKS2018 and CCKS2019, respectively, exhibiting outstanding performance in the extraction and utilization efficiency of clinical information.

Keywords: electronic medical record; named entity recognition; purpose-driven; SoftLexicon; RoBERTa

check for updates

Citation: Cui, X.; Yang, Y.; Li, D.; Qu, X.; Yao, L.; Luo, S.; Song, C. Fusion of SoftLexicon and RoBERTa for Purpose-Driven Electronic Medical Record Named Entity Recognition. *Appl. Sci.* **2023**, *13*, 13296. https://doi.org/10.3390/app132413296

Academic Editor: Luis Javier Garcia Villalba

Received: 10 November 2023
Revised: 11 December 2023
Accepted: 12 December 2023
Published: 15 December 2023

1. Introduction

Due to the rapid proliferation of electronic medical records and the varied data formats they encompass, individuals are encountering growing difficulties in the pursuit of clinical knowledge. Typically, electronic medical records encompass a wide range of data types, including hospitalization records, medical procedure records, and more, collectively chronicling a patient's entire treatment history. It holds a crucial position in medical decision-making and research. Against this background, creating an intelligent named entity recognition (NER) system to identify medical entities with crucial information provides a solid foundation for constructing electronic medical record clinical knowledge bases and promoting knowledge discovery [1].

NER constitutes the initial critical stage in the realm of natural language processing for information extraction [2]. Compared to general-domain texts, electronic medical record texts contain more domain-specific terminologies. For the electronic medical record NER task, the entity categories to be identified include anatomical location, disease diagnosis, symptom description, etc. Our purpose is to recognize these types of entities from clinical sentences. The formalization process of NER can be defined as follows: given an annotated sequence $S = < w_1, w_2, \ldots, w_n >$, we can obtain a list of triples after the recognition, where each triple contains the information of an entity [3]. For example, in the triple $< I_s, I_e, t >$, $I_s \in [1, n]$ and $I_e \in [1, n]$ respectively refer to the start index and the end index of an entity, and t is one of the predefined entity types. Figure 1 shows an example of the NER process. The input is a Chinese electronic medical record, “患者一月前出现腹部阵发性疼痛不适 The patient experienced paroxysmal abdominal pain and discomfort 1 month ago”, and two triples have been obtained after the NER system. The resulting triples indicate that “腹部 abdomen” is the entity of anatomical location, and “疼痛不适 pain and discomfort” is the entity of symptom description. In the NER system, people frequently overlook the fact that every component of the model is purposefully driven to generate the desired results. The automatic identification of these entities in electronic medical records plays a pivotal role in the development of medical informatics.

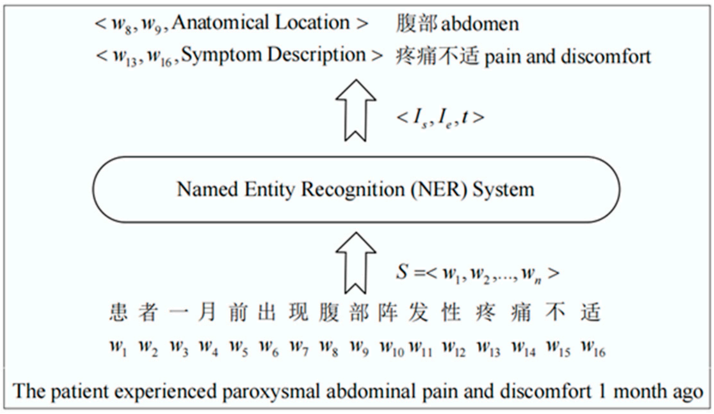

$< w_8, w_9,$ Anatomical Location $>$ 腹部 abdomen
$< w_{13}, w_{16},$ Symptom Description $>$ 疼痛不适 pain and discomfort
$< I_s, I_e, t >$
Named Entity Recognition (NER) System
$S = < w_1, w_2, \ldots, w_n >$
患 者 一 月 前 出 现 腹 部 阵 发 性 疼 痛 不 适
w_1 w_2 w_3 w_4 w_5 w_6 w_7 w_8 w_9 w_{10} w_{11} w_{12} w_{13} w_{14} w_{15} w_{16}
The patient experienced paroxysmal abdominal pain and discomfort 1 month ago

Figure 1. An example of NER process.

Most of the previous NER systems have adhered to the traditional Data–Information–Knowledge–Wisdom (DIKW) architecture [4,5]. Over the past few years, deep learning-based techniques for electronic medical record NER have gained immense popularity as one of the leading approaches. Researchers have endeavored to integrate the DIKW with a knowledge graph [6–9]. Meanwhile, thanks to the abundance of English-language corpora and the relatively straightforward processing, research in English NER has seen significant advancements. In the case of Chinese NER, a straightforward method involves initially conducting word segmentation and subsequently applying word-level sequence labeling models to the segmented sentences [10]. However, the unavoidable occurrence of incorrect word segmentation can result in NER errors, which can then lead to the propagation of these errors. Numerous studies also proposed that character-based Chinese NER techniques tend to outperform their word-based counterparts [11–13]. Nevertheless, character-based methods have their limitations as well, as they do not leverage the semantic information of words. To address this issue, Zhang et al. [14] first proposed the Lattice-LSTM model for mixed characters and lexicon words. In contrast to conventional character-based and word-based models, this approach can attain superior performance by harnessing explicit word information instead of relying on character sequence tags. Ma et al. [15] improved the Lattice-LSTM model and proposed a SoftLexicon method by considering more lexical information, achieving the best performance so far. Nonetheless, these methods may not prioritize the significance of the intended purpose and might not make the most of entities within Chinese electronic medical records, potentially leading to the exclusion of crucial lexical information. To address these gaps, this paper adopts the novel purpose-driven DIKW [16], which connects the diverse models of DIKW through purpose and unifies them as a whole.

Specifically, this paper proposes a novel purpose-driven SoftLexicon-RoBERTa-BiLSTM-CRF (SLRBC) NER model for electronic medical records. SLRBC initially acquires a word-level representation with the fusion of SoftLexicon and RoBERTa, subsequently improving the recognition of long-distance entities using Bidirectional Long Short-Term Memory (BiLSTM). Finally, a Conditional Random Field (CRF) is employed for decoding. The main contributions of this paper can be summarized as follows:

- We designed an SLRBC NER model for Chinese electronic medical records, fusing SoftLexicon and RoBERTa at the representation layer and adopting the classical BiLSTM-CRF framework to improve the model's performance;
- SLRBC employs a novel purpose-driven DIKW architecture. Within the SoftLexcion representation, we established four sets for each character to incorporate word lexicon information from electronic medical records into character representations and assigned weights to achieve a more comprehensive purpose representation;
- We conducted extensive experiments on the CCKS2018 and CCKS2019 public datasets to verify the effectiveness of SLRBC, and the results demonstrate its superiority in Chinese electronic medical record NER. The code can be accessed at https://github.com/QuXiaolong0812/SLRBC (accessed on 8 December 2023).

The rest of this paper is organized as follows. In Section 2, we present an overview of the literature covering character-based, word-based, and hybrid methods. Section 3 details our proposed approach, encompassing the representation layer, the encoding layer, and the label decoding layer. Section 4 includes our experimental procedures and a discussion of the results. Section 5 discusses strengths and weaknesses of the proposed approach. Finally, in Section 6, we summarize our findings and provide insights into potential avenues for future research.

2. Related Work

Early electronic medical record NER tasks adopted a combination of rule-based and dictionary-based methods, such as the MedEx system [17]. Yang et al. [18] discussed the linguistic and structural features of Chinese medical records and analyzed both rule-based and machine-learning approaches. With the development of deep learning, various deep neural networks such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Transformer have been applied to NER tasks. These deep-learning techniques can be categorized into character-based, word-based, and hybrid methods according to the various levels of granularity in the extraction of the representation layer.

2.1. Character-Based Methods

Character sequence labeling has always been the primary method for NER. In the early stages, Collobert et al. [19] proposed the CNN-CRF model and obtained competitive performance for various best NER models. Huang et al. [20] built the BiLSTM-CRF model, which achieved state-of-the-art results in NER tasks at that time. With the development of pre-trained language models (PLM), such as BERT [21], many researchers used PLMs to obtain character-level representations. Wang et al. [22] used BERT to address the missing contextual information problem and added an extra mechanism to capture relations between words. Wu et al. [23] proposed a model based on RoBERTa and character root features. They utilized RoBERTa to acquire medical features and applied BiLSTM to extract partial radical features. With the emergence of a lighter PLM, ALBERT, Yao et al. [24] proposed a model based on ALBERT-AttBiLSTM-CRF and transfer learning for fine-grained entity recognition of manufactured text. Aiming at the problem of word ambiguity in Chinese clinical NER tasks, Li et al. [25] proposed an ALBERT-based method and introduced a multi-head attention mechanism to capture inter-character dependencies. Due to their simplicity and effectiveness, character-based methods serve as the foundation for most existing NER approaches.

2.2. Word-Based Methods

In the context of Chinese NER, the effective integration of word lexicon information has consistently been a focal point of research. A conventional approach involves initial word segmentation followed by the application of a word-level sequence labeling model [26]. He et al. [27] proposed a unified NER model for Chinese social media and explored the advantages and disadvantages of three methods: character embedding; word embedding; and character position embedding. In addition, Rei [28] used word-level language modeling to enhance NER training and performed multi-task learning on large raw texts. Since boundary detection and type prediction for the NER task can cooperate with each other, Li et al. [29] proposed a modularized interaction network model, which can utilize both segment-level and word-level dependency information. With the application of the prompt learning paradigm in natural language processing, He et al. [30] proposed a prompt-based word-level information injection BERT to integrate prompt-guided lexicon information into a PLM. However, due to the inherent ambiguity of the Chinese language and the less well-defined granularity of Chinese words compared to languages like English, the majority of current Chinese NER methods typically do not separately address word-level features [13].

2.3. Hybrid Methods

Character-based approaches may not harness word information to its fullest extent, whereas word-based methods are vulnerable to errors stemming from segmentation problems. In order to overcome the drawbacks of both approaches, Dong et al. [31] employed the skip-gram model and bi-directional LSTM RNN model to extract word embeddings and character embeddings, respectively. These embeddings were subsequently merged to create the ultimate representations. Nonetheless, employing such a basic concatenation method can result in an excessive amount of redundant information. To address this issue, Zhang et al. [14] proposed the Lattice-LSTM model based on mixed characters and lexicon words for the first time. In contrast to character-based models, this model effectively exploits the semantic connections among neighboring characters within words. Furthermore, when compared to word-based models, it mitigates the adverse consequences stemming from word segmentation errors. Adversarial training avoids model overfitting by adding noise, which has been combined with the Lattice-LSTM model by successive research [32,33]. Ma et al. [15] also extended work on the Lattice-LSTM model and proposed a SoftLexicon method by integrating more lexical information without modifying the internal structure of LSTM. This approach incorporates all the words corresponding to each character within the representation layer and can be employed in different sequence labeling frameworks. In the field of agriculture, to alleviate the problems of agricultural text professionalism and uneven distribution of entity types, Zhang et al. [34] combined SoftLexicon with an attention mechanism and proposed the AttSoftlexicon to help the model effectively utilize lexical information. Building upon prior research, we integrate the SoftLexicon approach with RoBERTa and utilize the traditional BiLSTM-CRF architecture to achieve the intelligent extraction of entities from Chinese electronic medical records.

3. Proposed Method

The framework of the proposed SLRBC is shown in Figure 2, which consists of three main modules: (1) The representation layer incorporates the basic character representation, Softexicon representation, and RoBERTa representation to obtain a more comprehensive representation; (2) The encoding layer adopts the classical BiLSTM to capture contextual information and long-distance dependencies within sequences; (3) The label decoding layer uses the CRF mechanism to recognize entities. Taking the input sentence “腹部疼痛不适 abdominal pain and discomfort” as an example, character “腹 location” will be assigned the label “B-T1”, indicating that it is the beginning position (B) of the anatomical location (T1) entity. At the same time, the character “部” will be assigned the label “I-T1”, indicating that it is the inside position (I) of the anatomical location (T1) entity. As a result, the whole entity “腹部 abdomen” can be recognized. The details of each module are described below.

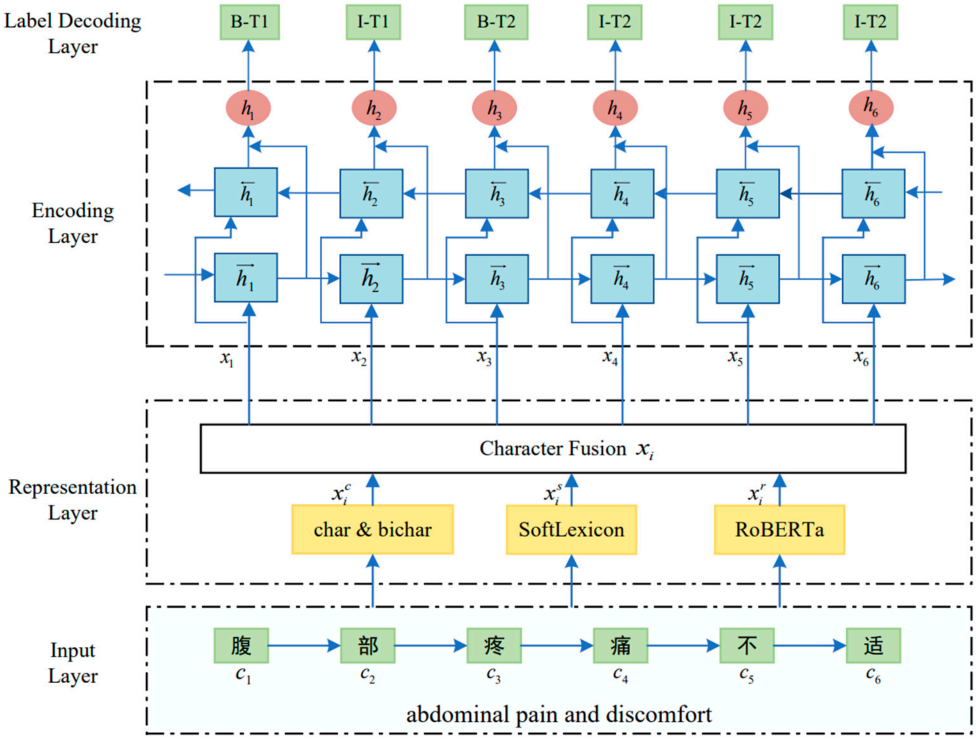


Figure 2. The framework of SLRBC.

3.1. Representation Layer

In addition to the basic character representation, both SoftLexicon and RoBERTa to obtain a more comprehensive representation are introduced in the representation layer. SoftLexicon effectively leverages a built word frequency dictionary to integrate word lexicon information into character representations, while RoBERTa excels at capturing more comprehensive semantic information.

In this paper, each sentence is treated as $s = \{c_1, c_2, \ldots, c_n\} \in V_c$, where V_c represents the set of all characters, and n represents the length of the sentence.

3.1.1. Basic Character Representation

The Lattice-LSTM model proposed by Zhang et al. [14] has proved the effectiveness of character embedding using both character embedding (char) and double character embedding (bichar). Therefore, we adopt both char and bichar in this step, which can be calculated as Equation (1):

$$x_i^c = [e^c(c_i); e^b(c_i, c_{i+1})] \tag{1}$$

where e^c denotes the character embedding lookup table, and e^b denotes the double-character embedding lookup table.

3.1.2. SoftLexicon Representation

Relying solely on the basic character representation described earlier is inadequate for fully integrating entity information into the model. Because traditional character-based methods face challenges in integrating medical word lexicon information into the model without the aid of word segmentation. On the other hand, word-based models heavily rely on the precision of word segmentation outcomes and have encountered difficulties in achieving desirable performance levels. Inspired by Ma et al. [15], we adopt the SoftLexicon approach to embed known medical entities into the model.

Individuals may perceive the same thing differently, and varying purposes can yield distinct outcomes. Similarly, as for the character c_i, we use the word frequency dictionary constructed in the data pre-processing stage to obtain the relevant words, and thus, construct four sets, $B(c_i)$, $M(c_i)$, $E(c_i)$, and $S(c_i)$, where $B(c_i)$ is the set including all words starting with c_i; $M(c_i)$ is the set including all words with c_i as the middle part; $E(c_i)$ is the set including all words ending with c_i, and $S(c_i)$ is the set of a single character c_i. These four sets signify four distinct purposes and encompass semantic details pertaining to various facets of entities. As shown in Figure 3, in the sequence of "患者诊断为直肠癌病 The patient was diagnosed with rectal cancer disease", these four sets corresponding to the character "肠 intestine" are as follows: B(肠intestine) = {"肠癌 intestinal cancer", "肠癌病 intestinal cancer disease"}; M(肠intestine) = {"直肠癌 rectal cancer", "直肠癌病 rectal cancer disease"}; E(肠intestine) = {"直肠 rectum"}; and S(肠intestine) = {"肠 intestine"}. It is obvious that the same lexical set may contain more than one word, such as B(肠intestine), which indicates that there are two words beginning with "肠 intestine". To achieve a more harmonious balance among these four purposes and facilitate their fusion, it is essential to assign weights to them based on word frequency.

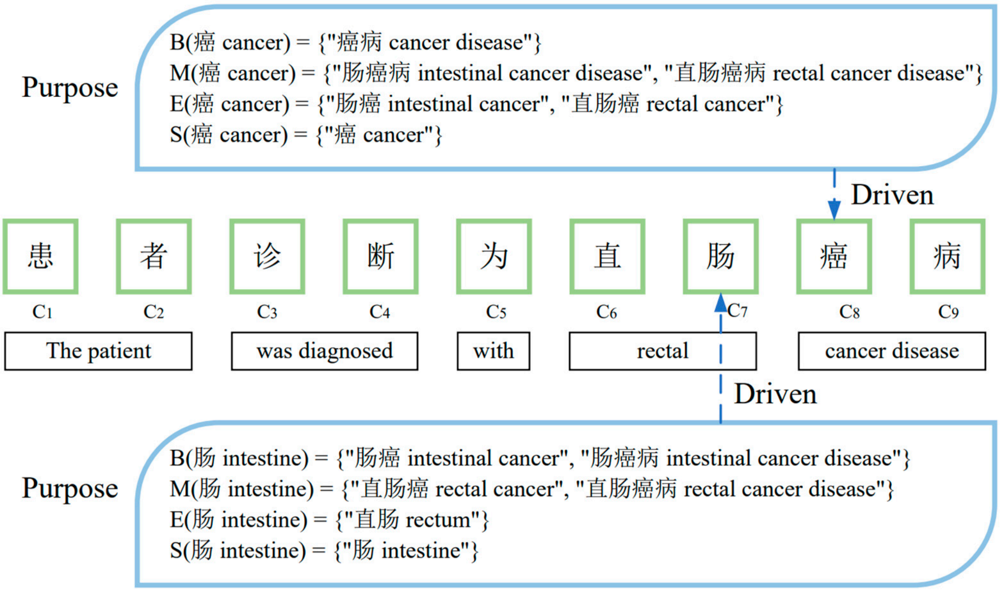

Purpose
B(癌 cancer) = {"癌病 cancer disease"}
M(癌 cancer) = {"肠癌病 intestinal cancer disease", "直肠癌病 rectal cancer disease"}
E(癌 cancer) = {"肠癌 intestinal cancer", "直肠癌 rectal cancer"}
S(癌 cancer) = {"癌 cancer"}
Driven
患
者
诊
断
为
直
肠
癌
病
C1
C2
C3
C4
C5
C6
C7
C8
C9
The patient
was diagnosed
with
rectal
cancer disease
Driven
Purpose
B(肠 intestine) = {"肠癌 intestinal cancer", "肠癌病 intestinal cancer disease"}
M(肠 intestine) = {"直肠癌 rectal cancer", "直肠癌病 rectal cancer disease"}
E(肠 intestine) = {"直肠 rectum"}
S(肠 intestine) = {"肠 intestine"}

Figure 3. An example of SoftLexicon method.

As for the input sequence $s = \{c_1, c_2, \ldots, c_n\}$, assuming that there is a vocabulary set $L \in \{B, M, E, S\}$, the weighted average is expressed as $v^s(L)$, which can be calculated as Equations (2) and (3):

$$v^s(L) = \frac{4}{Z} \sum_{w \in L} Z(w) e^w(w) \tag{2}$$

$$Z = \sum_{w \in B \cup M \cup E \cup S} Z(w) \tag{3}$$

where w is the word to be embedded; e^w is the word embedding lookup table; $Z(w)$ is the frequency of w counted in the word frequency dictionary.

To retain a greater amount of information, the representations of these four lexical sets are combined through concatenation. Finally, the embedding representation of c_i transformed by the SoftLexicon can be calculated as Equation (4):

$$x_i^s = [v^s(B(c_i)); v^s(M(c_i)); v^s(E(c_i)); v^s(S(c_i))] \tag{4}$$

3.1.3. RoBERTa Representation

In order to obtain better word-level semantic representation, we also introduce RoBERTa-wwm [35], a robust Chinese PLM. RoBERTa-wwm thoroughly addresses the necessity of word segmentation in the Chinese language and covers not only individual characters but entire words when applying masking. Table 1 shows an example of the whole word masking mechanism of RoBERTa-wwm. In contrast to BERT, three characters, “直肠癌 rectal cancer”, in RoBERTa-wwm, are considered as a single unit, and these characters are masked together. Following the pre-training phase, the acquired semantic representation operates at the lexical level, significantly enhancing the overall representational capacity of the model beyond the character level.

Table 1. An Example of whole word masking mechanism of RoBERTa-wwm.

Masking Strategy	Result
Original Text	患者诊断为直肠癌病 The patient was diagnosed with rectal cancer disease
BERT	患者诊断为直肠[MASK]病 The patient was diagnosed with rectal [MASK] disease
RoBERTa-wwm	患者诊断为[MASK] [MASK] [MASK]病 The patient was diagnosed with [MASK] [MASK] disease

With the text sequence $s = \{c_1, c_2, \ldots, c_n\}$ as input, the embedding representation with RoBERTa-wwm can be calculated as Equation (5):

$$x_i^r = e^r(c_i) \tag{5}$$

where e^r is the vector lookup table of RoBERTa-wwm.

3.1.4. Representation Fusion

There are two prevalent methods for fusing the aforementioned embedding vectors. One approach involves performing a weighted summation operation on the vectors, while the other entails concatenating these vectors. To simplify the computation, we opt for the latter. This approach offers a degree of flexibility since it does not need to be concerned about the dimensions of each embedding vector in the representation layer. The final embedding representation of c_i can be calculated as Equation (6):

$$x_i = [x_i^c; x_i^s; x_i^r] \tag{6}$$

3.2. Encoding Layer

In the encoding layer, we utilize the conventional BiLSTM to understand the fused representation obtained from the representation layer. BiLSTM stands out as an effective variant of RNN, capable of retaining information from both preceding and subsequent neural nodes, allowing for it to capture contextual information and long-distance dependencies within sequences [20].

The internal structure of the LSTM cell is shown in Figure 4. Each neural node of LSTM consists of an input gate, a forgetting gate, and an output gate. Utilizing these gate control units makes it feasible to determine whether to preserve or discard node-related information at each stage, thereby enabling the detection of long-distance dependencies. Specifically, the input gate is responsible for determining whether to retain the current node’s input information, while the forgetting gate determines whether to retain information from the previous neural node’s hidden layer. Lastly, the output gate decides whether to pass on the current node’s output to the subsequent node.

Assuming that the output of the hidden layer of the last node is h_{t-1}, and the input of the current node is x_t, the forgetting gate f_t, the input gate i_t, and the output gate o_t can be calculated as Equations (7)–(9):

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{7}$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{8}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{9}$$

where σ is the sigmoid activation function; W_{xf}, W_{hf}, W_{xi}, W_{hi}, W_{xo}, and W_{ho} are trainable weights, and b_f, b_i, and b_o are biases.

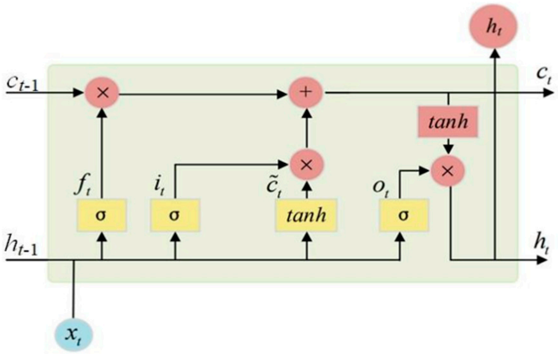


Figure 4. LSTM cell structure.

With these calculated gates, the memory cell corresponding to the current node can be calculated as Equation (10):

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \tag{10}$$

where $tanh$ is the activation function; W_{xc} and W_{hc} are trainable weights, and b_c is a bias.

Then, we can obtain the output of the hidden layer of the current node with the output gate o_t, which can be calculated as Equation (11):

$$h_t = o_t tanh(c_t) \tag{11}$$

The described process outlines the computation of LSTM. BiLSTM comprises both a forward LSTM and a backward LSTM, allowing for the simultaneous capture of information from both the preceding and following moments. It can be calculated as Equations (12)–(14):

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(x_t) \tag{12}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(x_t) \tag{13}$$

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \tag{14}$$

where $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ denote the outputs in both directions.

3.3. Label-Decoding Layer

The NER task can be viewed as a label prediction task, and these predicted labels have associations with neighboring labels. For instance, when the current predicted label is "I-Type1", the preceding label can only be "I-Type1" or "B-Type1". Within the encoding layer, we solely focus on the contextual information present in the electronic medical record text and do not account for label dependencies. As a result, we introduce a CRF layer after the neural network layer to determine the globally optimal sequence of labels, thus identifying possible entities.

Assuming that there is an input sequence $x = \{x_1, x_2, \ldots, x_n\}$ and the corresponding predicted label sequence is $y = \{y_1, y_2, \ldots, y_n\}$, the score of the predicted label sequence y corresponding to the input sequence x can be calculated by Equation (15):

$$S(x, y) = \sum_{i=1}^{n} A_{y_{i-1}, y_i} + \sum_{i=1}^{n} P_{i, y_i} \tag{15}$$

where A is the transfer score matrix; A_{y_{i-1},y_i} is the score of the label y_{i-1} transfer to the label y_i; P is the character label score matrix obtained from the output of the BiLSTM layer, and P_{i,y_i} is the score of the i-th character predicted as y_i by the BiLSTM layer.

Then, the probability distribution of the label sequence y is obtained by normalizing $S(x,y)$ with the softmax function, which can be calculated as Equation (16):

$$p(y|x) = \frac{e^{s(x,y)}}{\sum_{\widetilde{y} \in Y_x} e^{s(x,\widetilde{y})}} \tag{16}$$

where y denotes the true label sequence, and Y_x denotes all the predicted label sequences.

To improve loss calculation, this paper opts for maximum likelihood estimation, aiming to maximize the probability of the actual label sequences, which can be calculated as Equation (17):

$$log(p(y|x)) = S(x,y) - log(\sum_{\widetilde{y} \in Y_x} e^{S(x,\widetilde{y})}) \tag{17}$$

Ultimately, we employ the Viterbi algorithm to identify the label sequence with the highest score, which can be calculated as Equation (18):

$$y' = \underset{\widetilde{y} \in Y_x}{argmax}(S\left(x,\widetilde{y}\right)) \tag{18}$$

4. Experiments

4.1. Datasets

We assess the performance of SLRBC using the publicly available electronic medical record datasets from CCKS2018 (https://www.sigkg.cn/ccks2018/?page_id=16 (accessed on 10 November 2023)) and CCKS2019 (https://www.sigkg.cn/ccks2019/?page_id=62 (accessed on 10 November 2023)). The entity distribution statistics of the datasets are shown in Tables 2 and 3. CCKS2018 consists of 5 entity types and comprises 3251 sentences for training, 358 for validating, and 432 for testing. CCKS2019, on the other hand, defines 6 entity types and includes 5708 sentences for training, 755 for validation, and 743 for testing.

Table 2. CCKS2018 dataset entity distribution.

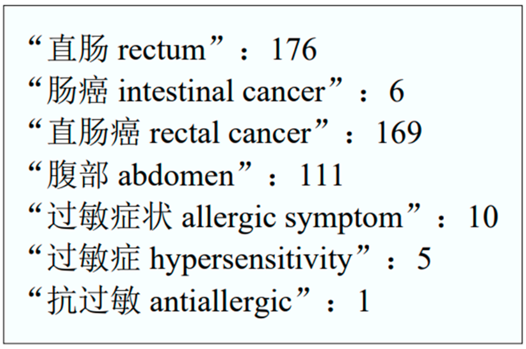
CCKS2018	Anatomical Location	Symptom Description	Independent Symptom	Medicine	Surgery
Number	7838	2066	3055	1005	1125
Proportion	51.95%	13.69%	20.25%	6.66%	7.46%

Table 3. CCKS2019 dataset entity distribution.

CCKS2019	Disease Diagnosis	Image Inspection	Laboratory Test	Surgery	Medicine	Anatomical Location
Number	4212	969	1195	1029	1768	8426
Proportion	23.93%	5.51%	6.79%	5.85%	10.05%	47.88%

The datasets are labeled with the BIO annotation method, in which the first character constituting the entity is labeled as "B-Type"; the rest characters constituting the entity are labeled as "I-Type", and the other non-entity characters are labeled as "O". As a result, there are 11 different labels in CCKS2018 and 13 in CCKS2019.

Since the word frequency dictionary needs to be constructed in the representation layer of SLRBC, we extract words from the datasets and count the frequency of each word. In this regard, the CCKS2018 word frequency dictionary encompasses 7619 words, while the CCKS2019 word frequency dictionary comprises 10,566 words. Exemplary words and their respective frequencies from the word frequency dictionary are depicted in Figure 5.



“直肠 rectum”： 176
“肠癌 intestinal cancer”： 6
“直肠癌 rectal cancer”： 169
“腹部 abdomen”： 111
“过敏症状 allergic symptom”： 10
“过敏症 hypersensitivity”： 5
“抗过敏 antiallergic”： 1

Figure 5. Examples from word frequency dictionary.

4.2. Experimental Setup

SLRBC is implemented on a single RTX 2080 Ti GPU with PyTorch version 1.10.0. We carry out an extensive array of experiments and record the parameter values at which the model achieves its optimal performance. The network weights are optimized by the Adam algorithm, and other details of the parameters are specified in Table 4.

Table 4. Specific setting of experimental parameters.

Parameter	Value
character vector dimension	50
word vector dimension	50
RoBERTa vector dimension	1024
LSTM hidden layer dimension	300
learning rate	0.015
batch size	6
epoch	30
dropout	0.5

In our experiments, we use standard Precision (P), Recall (R), and F1-score (F1) to evaluate the model performance:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{19}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{20}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \times 100\% \tag{21}$$

where TP represents the number of entities correctly identified by the model; FP represents the number of entities identified by the model as unrelated, and FN represents the number of correct entities not identified by the model.

4.3. Baselines

To verify the effectiveness of SLRBC, we conduct comparison experiments on the electronic medical record datasets of CCKS2018 and CCKS2019 with the following models: the first three are character-based methods, while the remaining two are hybrid methods. Note that there are few existing Chinese word-based methods, and they do not perform as well as the other two types of methods, so we do not consider this type of method.

- **CNN-CRF** [19] uses a CNN-based character embedding layer and employs CRF to decode;

- **BiLSTM-CRF** [20] uses a BiLSTM-based character embedding layer and employs CRF to decode;
- **BERT-BiLSTM-CRF** [22] is an improvement in the BiLSTM-CRF model, which introduces BERT on the top of the encoder;
- **SoftLexicon-BiLSTM-CRF** [15] combines character-level and word-level representations and adopts the classical BiLSTM-CRF framework;
- **SoftLexicon-BERT** [15] combines SoftLexicon and BERT to achieve better representations.

4.4. Main Results

In this section, we compare and analyze the performance of SLRBC with other baseline models. The experimental results are shown in Table 5. It is noticeable that when compared to these baseline models, SLRBC has enhanced the F1 score on the CCKS2018 dataset by 7.81%, 9.71%, 0.8%, 1.62%, and 0.78%, respectively, and on the CCKS2019 dataset by 16.93%, 2%, 1.43%, 2.36%, and 1.32%, respectively.

Table 5. Experimental results of different models on CCKS2018 and CCKS2019.

Model	CCKS2018			CCKS2019		
	P	**R**	**F1**	**P**	**R**	**F1**
CNN-CRF	83.72%	90.88%	87.16%	61.97%	76.49%	68.47%
BiLSTM-CRF	86.39%	84.15%	85.26%	82.92%	83.88%	83.40%
BERT-BiLSTM-CRF	94.37%	93.87%	94.17%	83.81%	84.14%	83.97%
SoftLexicon-BiLSTM-CRF	93.22%	93.48%	93.35%	81.98%	84.14%	83.04%
SoftLexicon-BERT	94.06%	94.32%	94.19%	83.21%	84.97%	84.08%
SLRBC	**94.70%**	**95.23%**	**94.97%**	**84.62%**	**86.19%**	**85.40%**

The winner is in bold.

Through horizontal comparison, we observe that the performance of all models is superior on CCKS2018 compared to CCKS2019. This discrepancy is evidently due to variations in the granularity of entity annotation between the two datasets. The clearer the annotation of training data, the better the performance of the model. Among these models, the CNN-CRF model displays the most prominent performance variation between the two datasets. Its performance on CCKS2019 notably lags behind that of the other models. This discrepancy arises from the fact that the named entities in CCKS2019 are longer, and the CNN model struggles to capture long-distance dependencies effectively. It is evident that our SLRBC can effectively address this issue.

Through vertical comparison, we observe a significant improvement in model performance after incorporating SoftLexicon compared to the traditional CNN and BiLSTM model structures. This result further validates the positive impact of the purpose-driven approach on enhancing NER accuracy. Simultaneously, it is evident that the F1 scores for BERT-BiLSTM-CRF and SoftLexicon-BERT with PLM, as well as our SLRBC, are notably superior to those without PLM. It proves the effectiveness of introducing large PLMs. The remarkable performance of SLRBC underscores the effectiveness of the model structure we proposed, which integrates SoftLexicon and RoBERTa in the realm of Chinese electronic medical record NER.

In addition, we conduct a comparison of the training speed and testing speed of each model, as shown in Table 6. It can be found that the model without BERT or RoBERTa, including CNN-CRF, BiLSTM-CRF, and SoftLexicon-BiLSTM-CRF, has faster processing speed, primarily due to the simpler model structures and fewer parameters. However, SLRBC operates at a slower pace, indicating that RoBERTa, with its increased parameter count, contributes to the slower processing speed compared to BERT. Meanwhile, the output vector dimension of BERT is 768, while the output vector dimension of RoBERTa is 1024, which further increases the overall computational load of the model.

Table 6. Comparison of training speed and testing speed of each model.

Model	CCKS2018		CCKS2019	
	Training Speed	**Testing Speed**	**Training Speed**	**Testing Speed**
CNN-CRF	18.39 st/s	46.04 st/s	21.97 st/s	54.29 st/s
BiLSTM-CRF	18.80 st/s	48.21 st/s	21.65 st/s	54.31 st/s
BERT-BiLSTM-CRF	15.53 st/s	20.86 st/s	17.73 st/s	22.22 st/s
SoftLexicon-BiLSTM-CRF	18.56 st/s	46.45 st/s	21.51 st/s	54.36 st/s
SoftLexicon-BERT	15.21 st/s	20.37 st/s	17.14 st/s	21.96 st/s
SLRBC	11.85 st/s	13.56 st/s	13.57 st/s	14.13 st/s

"st/s" represents the number of sentences processed by the model per second.

4.5. Ablation Study

4.5.1. Analysis of Different Representation Layers

To further investigate the impact of four modules in representation layers on model performance, we conduct ablation experiments. Acknowledging character embedding (char) as the most basic representation, we make the deliberate decision to retain it consistently in each experiment. Subsequently, we progressively eliminate either one or both bichars, SoftLexicon, and RoBERTa. Table 7 shows the specific experimental results.

Table 7. Effects of different modules in representation layer on model performance.

Four Modules in Representation Layer				F1 on Two Datasets	
Char	**Bichar**	**SoftLexicon**	**RoBERTa**	**CCKS2018**	**CCKS2019**
√	×	√	√	94.35%	84.12%
√	√	×	√	93.88%	84.75%
√	√	√	×	93.35%	83.04%
√	×	×	√	93.73%	84.70%
√	×	√	×	93.12%	83.38%
√	√	×	×	92.38%	82.83%
√	√	√	√	94.97%	85.40%

"√" means to add the module, and "×" means to remove the module.

In the representation layer of our SLRBC, char, bichar, SoftLexicon, and RoBERT are used simultaneously to achieve the best results on both datasets, once again verifying the validity of SLRBC. The removal of SoftLexicon or RoBERTa results in a significant decrease in F1 compared to SLRBC. F1 reaches its lowest point when both are removed simultaneously, providing conclusive evidence that these two modules play a crucial role in the representation layer of SLRBC. However, when bichar is removed, F1 does not decrease significantly compared with SLRBC. This is due to the fact that bichar, serving as an adjunct to char, makes a relatively smaller contribution to the representation. Experiments conducted on the two datasets reveal that the F1 for CCKS2018 exhibited less fluctuation compared to that of CCKS2019. This result indicates that when the entity annotation of training data is clearer, the change in the representation layer has less influence on the final recognition performance.

4.5.2. Analysis of Different Neural Networks

In the SLRBC proposed in this paper, BiLSTM is selected as the encoding layer. In order to explore the influence of different neural networks on the overall model and subsequently optimize the neural network layer, we replace the BiLSTM with CNN and Transformer in ablation experiments. The experimental results are shown in Table 8, which indicates that the model performs best when using BiLSTM in the encoding layer. When employing CNN and Transformer, there is a discernible disparity in F1 on both datasets compared to BiLSTM. It can be inferred that the constrained receptive field of CNN hinders its ability to capture long-distance dependencies, and the ability of Transformer to capture such dependencies is also marginally inferior to that of BiLSTM. Therefore, we choose to use BiLSTM in the encoding layer, resulting in the best overall model performance.

Table 8. Experimental results of different encoding layer models on CCKS2018 and CCKS2019.

Encoding Layer Model	**CCKS2018**			**CCKS2019**		
	P	**R**	**F1**	**P**	**R**	**F1**
CNN	91.25%	94.32%	92.76%	76.23%	84.46%	80.13%
Transformer	91.83%	92.22%	92.02%	79.57%	80.54%	80.05%
BiLSTM	94.70%	95.23%	94.97%	84.62%	86.19%	85.40%

4.5.3. Analysis of Different Hidden Layer Dimensions

We set the hidden layer dimensions as 100, 200, 300, 400, and 500 to investigate its impact on model performance. At the same time, we compare the total loss and F1 of the model under different hidden layer dimensions on CCKS2018 and CCKS2019, which can be seen in Figures 6 and 7. With a training epoch count below 10, the total loss experiences the most rapid decline, and F1 shows the most noticeable increase. Beyond 20 training epochs, both the total loss and F1 begin to gradually stabilize. Consequently, we opt to set the epoch to 30. In general, the dimension of the hidden layer exhibits no discernible impact on F1. Setting the hidden layer dimension to 100 results in the fastest convergence, and the speed diminishes with an increase in the hidden layer dimension. When the hidden layer dimension is set to 300, SLRBC demonstrates relatively excellent overall performance, reaching the peak of F1. Therefore, we set the dimension of the hidden layer to 300 in this paper.

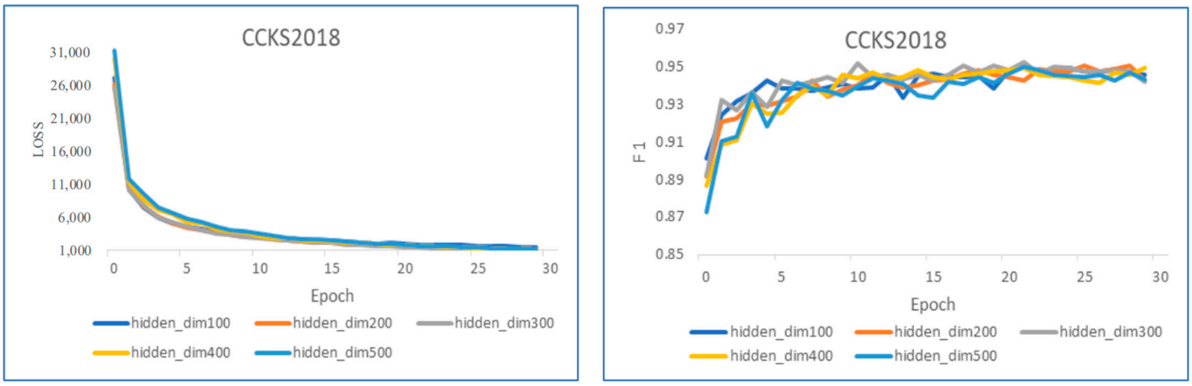


Figure 6. LOSS and F1 under different hidden layer dimensions on CCKS2018.

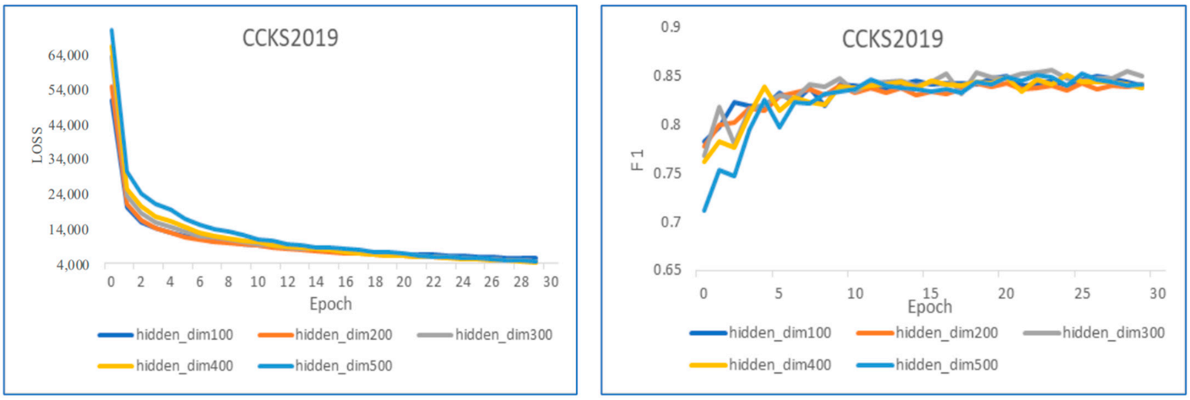


Figure 7. LOSS and F1 under different hidden layer dimensions on CCKS2019.

5. Discussion

Our experimental results demonstrate that the proposed SLRBC can provide more accurate recognition performance by using a purpose-driven approach with the fusion of SoftLexicon and RoBERTa. We compare SLRBC with five baseline models: CNN-CRF [19]; BiLSTM-CRF [20]; BERT-BiLSTM-CRF [22]; SoftLexicon-BiLSTM-CRF [15]; and SoftLexicon-BERT [15]. By comparing CNN-CRF, BiLSTM-CRF, and BERT-BiLSTM-CRF, we prove the importance of lexical information in Chinese electronic medical record NER. Then, we compare SLRBC with SoftLexicon-BiLSTM-CRF and SoftLexicon-BERT, which belong to hybrid methods. The latter two only consider one or both representations, while our SLRBC considers basic character representation, SoftLexicon representation, and RoBERTa representation at the same time, thus achieving a more comprehensive representation. Moreover, the classical BiLSTM-CRF framework is adopted in our SLRBC, which helps to capture contextual information of entities more effectively.

However, we also notice that SLRBC operates at a slower pace compared with other baseline models. This may be attributed to the fusion method adopted by the representation layer. In SLRBC, we use the simple concatenate method to fuse different representations. This approach preserves as much of the original information as possible but increases the dimensions of representation features. Although it is helpful to improve the performance of entity recognition, it also significantly increases the time and space cost. Therefore, different fusion methods can be explored to further improve SLRBC.

6. Conclusions

In this paper, we propose a novel purpose-driven model called SLRBC for the Chinese electronic medical record NER model, namely, SLRBC, which uses the fusion of SoftLexicon and RoBERTa within the representation layer. Compared to existing character-based, word-based, and hybrid methods, SLRBC fuses SoftLexicon and RoBERTa, constructing four sets with different purposes to achieve a more comprehensive purpose representation. Such a purpose-driven approach helps to incorporate the word lexicon information from electronic medical records into the character representations, thereby endowing the model with more extensive semantic embedding representations. Simultaneously, SLRBC introduces the classical BiLSTM-CRF framework to enhance the model's ability in medical entity recognition. We conducted several comparative experiments on two public NER datasets. The experimental results show that each module plays an indispensable role in our proposed SLRBC. Specifically, the F1 of SLRBC increases by 0.78% to 9.71% on CCKS2018 and 1.32% to 16.93% on CCKS2019, which validates the effectiveness of the model and lays the foundation for clinical knowledge discovery in electronic medical records.

Although SLRBC has achieved advanced performance in experiments, there are still challenges and space for improvement. In the future, faced with increasingly complex medical knowledge, we will delve deeper into integrating more domain knowledge and merging the latest prompt learning paradigm to further elevate the model's performance and its practical value in clinical applications.

Author Contributions: Conceptualization, X.C. and Y.Y.; data curation, Y.Y.; formal analysis, L.Y.; funding acquisition, X.C.; investigation, Y.Y. and C.S.; methodology, Y.Y. and S.L.; project administration, Y.Y.; resources, S.L.; software, Y.Y. and X.Q.; supervision, D.L.; validation, X.C., Y.Y. and D.L.; visualization, S.L.; writing—original draft, Y.Y.; writing—review and editing, D.L., X.Q. and L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Key R&D Program of China (2022YFF1302700), the Outstanding Youth Team Project of Central Universities (QNTD202308) and the Ant Group through the CCF-Ant Research Fund (CCF-AFSG 769498 RF20220214).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. Fries, J.A.; Steinberg, E.; Khattar, S.; Fleming, S.L.; Posada, J.; Callahan, A.; Shah, N.H. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nat. Commun.* **2021**, *12*, 2017. [CrossRef]
2. Li, D.M.; Luo, S.S.; Zhang, X.P.; Xu, F. Review on named entity recognition. *J. Front. Comput. Sci. Tech.* **2022**, *16*, 1954–1968.
3. Li, J.; Sun, A.X.; Han, J.L.; Li, C.L. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 50–70. [CrossRef]
4. Rowley, J. The wisdom hierarchy: Representations of the DIKW hierarchy. *J. Inf. Sci.* **2007**, *33*, 163–180. [CrossRef]
5. Li, Y.B.; Li, Z.; Duan, Y.C.; Spulber, A.B. Physical artificial intelligence (PAI): The next-generation artificial intelligence. *Front. Inf. Technol. Electron. Eng.* **2023**, *24*, 1231–1238. [CrossRef]
6. Song, Z.Y.; Duan, Y.C.; Wan, S.X.; Sun, X.B.; Zou, Q.; Gao, H.; Zhu, D. Processing optimization of typed resources with synchronized storage and computation adaptation in fog computing. *Wirel. Commun. Mob. Comput.* **2018**, *2018*, 3794175. [CrossRef]
7. Duan, Y.C.; Sun, X.B.; Che, H.Y.; Cao, C.J.; Li, Z.; Yang, X. Modeling data, information and knowledge for security protection of hybrid IoT and edge resources. *IEEE Access* **2019**, *7*, 99161–99176. [CrossRef]
8. Lei, Y.; Duan, Y.C. Trusted service provider discovery based on data, information, knowledge, and wisdom. *Int. J. Softw. Eng. Knowl. Eng.* **2021**, *31*, 3–19. [CrossRef]
9. Gao, H.H.; Duan, Y.C.; Shao, L.X.; Sun, X.B. Transformation-based processing of typed resources for multimedia sources in the IoT environment. *Wirel. Netw.* **2021**, *27*, 3377–3393. [CrossRef]
10. Wu, F.Z.; Liu, J.X.; Wu, C.H.; Huang, Y.F.; Xie, X. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 3342–3348.
11. He, J.Z.; Wang, H.F. Chinese named entity recognition and word segmentation based on character. In Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing, Hyderabad, India, 11–12 January 2008; pp. 128–132.
12. Liu, W.; Xu, T.G.; Xu, Q.H.; Song, J.Y.; Zu, Y.R. An encoding strategy based word-character LSTM for Chinese NER. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 2379–2389.
13. Ding, R.X.; Xie, P.J.; Zhang, X.Y.; Lu, W.; Li, L.L.; Si, L. A neural multi-digraph model for Chinese NER with gazetteers. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1462–1467.
14. Zhang, Y.; Yang, J. Chinese NER using lattice LSTM. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1554–1564.
15. Ma, R.T.; Peng, M.L.; Zhang, Q.; Wei, Z.Y.; Huang, X.J. Simplify the usage of lexicon in Chinese NER. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5951–5960.
16. Li, Y.B.; Duan, Y.C.; Maamar, Z.; Che, H.Y.; Spulber, A.B.; Fuentes, S. Swarm differential privacy for purpose-driven Data-Information-Knowledge-Wisdom architecture. *Mob. Inf. Syst.* **2021**, *2021*, 6671628. [CrossRef]
17. Xu, H.; Stenner, S.P.; Doan, S.; Johnson, K.B.; Waitman, L.R.; Denny, J.C. MedEx: A medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 19–24. [CrossRef] [PubMed]
18. Yang, J.F.; Yu, Q.B.; Guan, Y.; Jiang, Z.P. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. *Acta Autom. Sin.* **2014**, *40*, 1537–1562.
19. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
20. Huang, Z.H.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
21. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
22. Wang, Q.C.; E, H. A BERT-based named entity recognition in Chinese electronic medical record. In Proceedings of the 2020 9th International Conference on Computing and Pattern Recognition, Xiamen, China, 30 October–1 November 2020; pp. 13–17.
23. Wu, Y.; Huang, J.; Xu, C.; Zheng, H.L.; Zhang, L.; Wan, J. Research on named entity recognition of electronic medical records based on RoBERTa and radical-level feature. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 2489754. [CrossRef]
24. Yao, L.G.; Huang, H.S.; Wang, K.W.; Chen, S.H.; Xiong, Q.Q. Fine-grained mechanical Chinese named entity recognition based on ALBERT-AttBiLSTM-CRF and transfer learning. *Symmetry* **2020**, *12*, 1986. [CrossRef]
25. Li, D.M.; Long, J.; Qi, J.T.; Zhang, X.P. Chinese clinical named entity Recognition with ALBERT and MHA mechanism. *Evid-Based Complement. Altern. Med.* **2022**, *2022*, 2056039. [CrossRef] [PubMed]
26. Yang, J.; Teng, Z.Y.; Shang, M.S.; Zhang, Y. Combining discrete and neural features for sequence labeling. In Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics, Konya, Turkey, 3–9 April 2016; pp. 140–154.

27. He, H.F.; Sun, X. A unified model for cross-domain and semi-supervised named entity recognition in Chinese social media. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 3216–3222.
28. Rei, M. Semi-supervised multitask learning for sequence labeling. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 2121–2130.
29. Li, F.; Wang, Z.; Hui, S.C.; Liao, L.J.; Song, D.D.; Xu, J.; He, G.; Jia, M. Modularized interaction network for named entity recognition. 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; pp. 200–209.
30. He, Q.; Chen, G.W.; Song, W.C.; Zhang, P.Z. Prompt-based word-level information injection BERT for Chinese named entity recognition. *Appl. Sci.* **2023**, *13*, 3331. [CrossRef]
31. Dong, X.S.; Chowdhury, S.; Qian, L.J.; Guan, Y.; Yang, J.F.; Yu, Q. Transfer bi-directional LSTM RNN for named entity recognition in Chinese electronic medical records. In Proceedings of the 19th International Conference on e-Health Networking, Applications and Services, Dalian, China, 12–15 October 2017; pp. 1–4.
32. Zhao, S.; Cai, Z.P.; Chen, H.W.; Wang, Y.; Liu, F.; Liu, A. Adversarial training based lattice LSTM for Chinese clinical named entity recognition. *J. Biomed. Inform.* **2019**, *99*, 103290. [CrossRef]
33. Su, S.; Qu, J.; Cao, Y.; Li, R.Q.; Wang, G. Adversarial training lattice LSTM for named entity recognition of rail fault texts. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 21201–21215. [CrossRef]
34. Zhang, L.L.; Nie, X.L.; Zhang, M.M.; Gu, M.Y.; Geissen, V.; Ritsema, C.J.; Niu, D.; Zhang, H. Lexicon and attention-based named entity recognition for kiwifruit diseases and pests: A Deep learning approach. *Front. Plant Sci.* **2022**, *13*, 1053449. [CrossRef] [PubMed]
35. Cui, Y.M.; Che, W.X.; Liu, T.; Qin, B.; Yang, Z.Q. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3504–3514. [CrossRef]

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