Heliyon 8 (2022) e11291

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Surgical procedure long terms recognition from Chinese literature incorporating structural feature

Nan Jiale, Dongping Gao *, Yuanyuan Sun, Xiaoying Li, Xifeng Shen, Meiting Li, Weining Zhang, Huiling Ren, Yi Qin

Institute of Medical Information, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, 100020, China

ARTICLE INFO

Keywords:
Natural language processing
Surgical procedure
Long term recognition
SoftLexicon

ABSTRACT

With rapid development of technologies in medical diagnosis and treatment, the novel and complicated concepts and usages of clinical terms especially of surgical procedures have become common in daily routine. Expected to be performed in an operating room and accompanied by an incision based on expert discretion, surgical procedures imply clinical understanding of diagnosis, examination, testing, equipment, drugs and symptoms, etc., but terms expressing surgical procedures are difficult to recognize since the terms are highly distinctive due to long morphological length and complex linguistics phenomena. To achieve higher recognition performance and overcome the challenge of the absence of natural delimiters in Chinese sentences, we propose a Named Entity Recognition (NER) model named Structural-SoftLexicon-Bi-LSTM-CRF (SSBC) empowered by pre-trained model BERT. In particular, we pre-trained a lexicon embedding over large-scale medical corpus to better leverage domain-specific structural knowledge. With input additionally augmented by BERT, rich multigranular information and structural term information is transferred from Structural-SoftLexicon to downstream model Bi-LSTM-CRF. Therefore, we could get a global optimal prediction of input sequence. We evaluate our model on a self-built corpus and results show that SSBC with pre-trained model outperforms other state-of-the-art benchmarks, surpassing at most 3.77% in F1 score. This study hopefully would benefit Diagnostic Related Groups (DRGs) and Diagnosis Intervention Package (DIP) grouping system, medical records statistics and analysis, Medicare payment system, etc.

1. Introduction

Surgical procedures contextualized in the medical field refer to a procedure that would be expected to be performed in an operating room and accompanied by an incision based on expert discretion. As a normative expression for clinical examination and treatment, surgical procedure terms, such as “ovary transplantation”, “pulmonary artery thrombolytic agent perfusion”, etc., become an important part of the standard medical terminology system, as well as standards with the smallest granularity and the highest necessity for medical and health information standard system. Those terms imply a clinical understanding of diagnosis, examination, testing, equipment, drugs, and symptoms, etc., and are a key building block for many intelligent medical systems.

Being a part of public standard medical terminology resources, surgical procedure terms have the problem of covering the real world, that is, the ability to describe real data [1]. With the booming development in the biomedical field, surgical procedures composing operative approaches, equipment, techniques, drugs, and such continuously change in practical use. Forms of usage like syntagmatic and cross use among these types caused complicated and diversified changes in terms' morphology, and the ever-increasing bytes of terms further exacerbates this problem. It grows difficult for static vocabularies and standards to discover variants and derivatives in reality thus fully covering surgical procedure terms promptly. How to address the problem, or more precisely, how to address the problem originated from the increased length becomes the heart of our concern. Previous studies provide excellent named entity recognition (NER) methods for detecting entities alike [2, 3, 4, 5].

Named entity recognition is the process of identifying targeted entities and their corresponding domain entity tags from unstructured unlabeled text. Suffice to say that improved NER model capability on surgical procedures directly links to higher identification performance of groupers of Diagnostic Related Groups (DRGs) and Diagnosis Intervention Package (DIP), and further facilitates Medicare payment system,

* Corresponding author.
E-mail address: gaodp_2022@126.com (D. Gao).

https://doi.org/10.1016/j.heliyon.2022.e11291
Received 24 July 2022; Received in revised form 4 October 2022; Accepted 24 October 2022

statistics, and analysis of Electronic Medical Records (EMRs) and Electronic Health Records (EHRs) in any potential use. Generally, rapid and accurate identification of such terms is of great significance to efficient hospital management, automatic update of terminology knowledge base, terminology standardization, and interoperability as well as the construction of a Chinese medical terminology system. However, an undeniable fact is that previous NER studies in the biomedical domain mainly concentrate on more common concepts [6, 7, 8, 9, 10, 11, 12], and bare efforts are put into this field or long entity recognition. Current NER models also find their incapability to process these long and complicated terms [13, 14]. Therefore, targeted studies are urgently needed for the automatic recognition of surgical procedure long terms.

This paper proposes a NER model named Structural-SoftLexicon-Bi-LSTM-CRF (SSBC) empowered by the pre-trained model BERT [15]. SSBC first maps characters in the input sequence into dense vectors, then the SoftLexicon feature is built and added to the character vector. The augmented character representations are later input into the sequence modeling layer and CRF (conditional random field) [16, 17] layer to generate final predictions. During the process, a pre-trained language model is applied to learn prior semantic knowledge. SSBC alleviates the long term recognition problem by using lexicon-character combination strategy merely on embedding layer without reconstructing the whole architecture and proves to be effective.

The contributions of this work are summarized as follows:

- We constructed a term corpus and combined term structural features with a deep learning method to overcome combination and cross ambiguity. Testing results outperformed baseline methods and reached a new high F1 score at 79.63%.
- We proposed Structural-SoftLexicon-Bi-LSTM-CRF (SSBC) model which integrates lexicon and character features and avoids complex sequence modeling. The overall architecture is simple, and the transferability is strong.
- We introduced a pre-trained domain lattice and pre-trained language model, which effectively tackle the problem of OOV (out-of-vocabulary) words. And the model performance can continually be improved by subsequent lexicon expansion.

2. Characteristics and difficulties of surgical procedures recognition

Considering the average length appeared in various surgical taxonomies and standards, we define surgical procedure long terms as terms with at least 5 Chinese characters. For entity recognition tasks, we summarize term features as follows:

(1) Highly territorial. Tiny replacement of certain characters may cause huge changes in the semantics of the words or even errors due to the high sensitivity. Thus, characters with rich information are demanded.
(2) More bytes. We define surgical procedure long terms as terms with at least 5 Chinese characters after considering the average length appeared in various surgical taxonomies and standards. For instance, in a normative document *Operations And Procedures Classification Code National Clinic Edition 3.0* [4], terms of at least 5 characters account for about 97% of all terms. The semantic dependency and complexity positively correlate to the length of terms.
(3) Highly structural. Surgical procedure terms treat leading terms as its core element and combine other kinds of components to form a term unit, seeing in Figure 1. In fact, almost all term elements follow national or industrial norms, making it possible to make full use of external lexicons to recognize terms. We defined term elements of 7 types: Ana (anatomy), App (operative approach), Dis (disease), Lea (leading term[1]), Equ (equipment), Drug (drug), and Sym (symptom) (see Figure 2).

Consequently, 3 challenges need overcome:

(1) For a term, App + Ana + Dis + Lea is the basic formula. However, syntagmatic and cross ambiguity in linguistics lead to word boundary changes, therefore generating cutting errors.
(2) Although terms in practical use are standardized, not all of them are. Due to the descriptive narrative and naming style, surgical procedure terms often have novel and diversified morphological patterns with weak name regularity and nomenclature, and are full of OOV words, which unfavorably embarrass recognition performance.
(3) A large number of symbols and multilingual words. For instance, English words, Arabic numerals, and hyphens are widely used in terms like “Bacon-Black术” (Bacon-Black), “AVAULTA全盆底重建术” (AVAULTA Total Pelvicfloor Reconstruction), and “重组人类活化C蛋白输注” (Recombinant Human Activated Protein C Perfusion).

3. Related works

NER is widely used in Chinese medical texts, involving electronic medical records [18, 19], traditional Chinese medicine texts [20, 21], clinical guidelines [22], disease subtypes [23], admission notes [24], PICO [25] disease and plant name [26], e. tc. Most approaches focus on feature extraction either by rule and dictionary-based means or by machine learning and deep learning means [27, 28]. detects entities from unstructured texts by using rules and dictionaries on the character-word level [29, 30]. utilize machine learning models extracting medical entities based on manually designed character-word level features like dictionaries, pos tagging, and n-grams. Rule-based approaches rely heavily on manual formulation of heuristic rules which requires thorough analysis of corpus and strong enumeration capability for possible rules. The dictionary-based approaches face pretty much the same problem in covering all related entities, which causes poor recognition performance [31]. Machine learning-based approach at earlier stage takes a naïve resolution by simply classifying each word independently [32]. The main problem with this resolution is it assumes that named entity labels are independent which is not the case. Later, the machine learning-based approach regards NER as a sequence labeling issue, and approaches, e.g., conditional random fields, Hidden Markov Models [33], support vector machine (SVM) [34], maximum entropy (ME) [35], are extensively used. Specifically, the machine learning-based approach builds dependencies in and out of input sequences based on labels and then is sent downstream to tackle labeling biases by a global optimization prediction layer.

Deep learning methods characterized by end-to-end learning, deep features extraction, and automatic modeling have comparatively surpassed outcomes maintained by traditional machine learning approaches [36]. Haixin Tan et al [37] propose an ensemble model composing Bi-LSTM, CRF, and transformer encoder [trans] for medical term recognition, incorporating Chinese character radical information and a pre-trained language model. Researchers gradually find that by optimizing the representation layer, models are easier to deploy and appear competitive results [MLNER and lexicon]. Articles [11, 38, 39] use encyclopedia, abstract, and patent corpus to input pre-trained word vectors to the model, and capture word features to form character-word hybrid embedding as model input; Yuan Li et al [40] apply static and dynamic attention mechanism to measure adjacent word weights around characters to obtain word importance information and concatenate into

[1] A surgical procedure term is usually ended with a leading term, that is, a word or a character of a clinical operative pattern. When coding operations or procedures of medical records in hospitals, deciding leading terms is the first step in the coding flow.

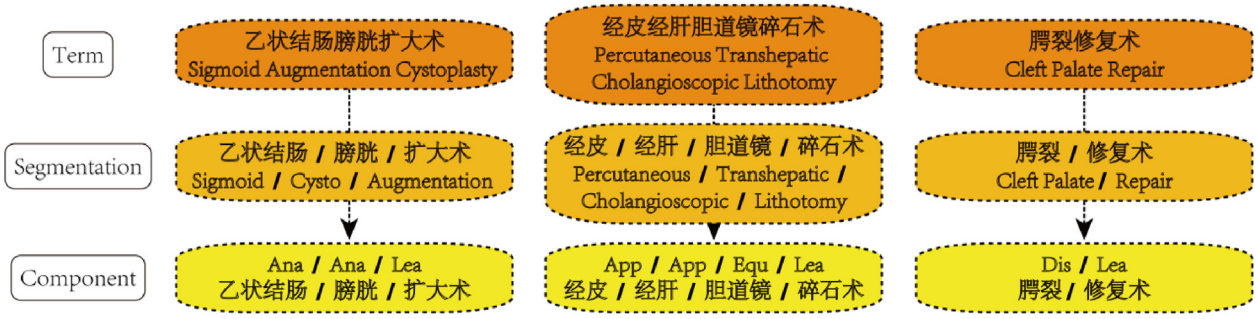

Figure 1. Examples of term segmentation by term type.

character vectors. To obtain clearer word boundaries, some use segmentation tagging [41] and tagging with different word boundaries generated by feature template segmentation [42] to assist in recognizing entities. Other scholars, based on domain feature, combine external thesaurus with character or word embeddings to facilitate entity recognition. With this thought, Heng-Yi Wu et al. [43] used Drugbank 4.0, Medical Subject Headings (MeSH) vocabulary, and medical literature to construct a dictionary of drug names and metabolism, meanwhile combined part-of-speech information to identify whether entities are related terms or substrings. LERNER I et al [44] use a terminology system based on Unified Medical Language System (UMLS) Metathesaurus and Systematized Nomenclature of Medicine (SNOMED) to obtain entity type information, and combine character and word embedding as model input to detect drug names. In addition to referring to external taxonomies, extracting boundary-specified subwords from known entities also becomes a feasible approach to enriching embeddings. For instance, C.Y.Hou [45] used the BPE iterative merging algorithm to generate subword units from known entities to form character-subword mixed embedding as the input of ID-CNNs-CRF model, which shortens the sequence length and facilitates the determination of boundaries of long entities, achieving superiority in long term recognition. Despite some unique techniques such as denoising algorithm [46] and adversarial training [47] being used, the main structures of these methods based on the embedding layer. Multi-level feature fusion, e.g., character, word, morphology, syntactic, pinyin, and radical features, is getting popular more than ever.

Although many deep neural network models have been applied to entity recognition, they all have different degrees of defects. First, methods that incorporate lexical information inevitably inherent word segmentation errors which can propagate further downstream. Considering the long span of surgical procedure terms, the word-based method will have a negative impact on recognition. Character-based method solves word segmentation problem, and empirically, its performance is better than word-based deep learning models [48, 49], but the former misses the word boundary information, and character sequence is often longer than word sequence, which increases difficulties. Naturally, more and more researchers consciously integrate word information into character-based NER methods to fully take advantage of both and avoid their shortcomings. This idea has two pathways: one is to dynamically improve sequence modeling layer [50, 51], and another is to improve embedding representation layer [52].

As a classical model for refactoring sequence modeling layers to utilize lexicon information, Lattice-LSTM [51] refreshes top scores for several tasks. However, Lattice-LSTM is slow in training speed, complex in architecture, and hard to combine with other deep neural networks. Ruotian Ma et al. [53] proposed SoftLexicon by improving Lattice-LSTM, which avoids complex sequence modeling and large-scale annotation corpus construction and can be combined with other neural network NER models by adjusting embedding representation layer. Therefore, we believe that it can transfer well to operation and procedure term recognition.

In this paper, we adopt a deep learning model incorporating character information, lexicon information, and term structural features. The embedding representation layer is fine-tuned to accept character embedding, BERT [54] embedding, and SoftLexicon embedding, and Bi-LSTM model is used for sequence modeling. In the decoding layer, we use CRF to solve the labeling bias problem in Bi-LSTM, thus obtaining global optimal tagging. Through experiments on the constructed corpus, our model reaches 79.63% on F1 value, outperformed by at most 3.77% than baseline models.

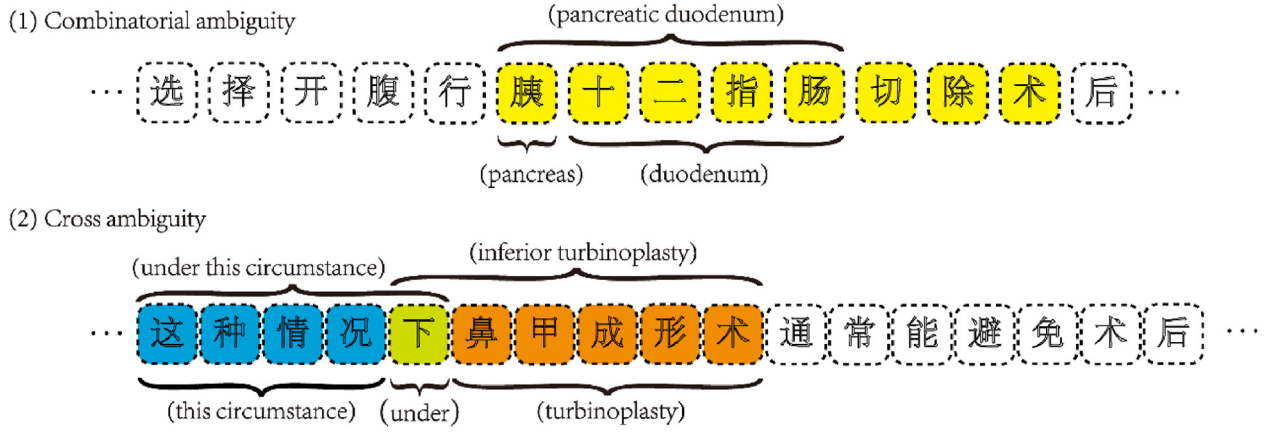

Figure 2. (1) Example indicates that the ground-truth entity (characters highlighted in yellow) may be mistakenly cut into an entity without pancreas (even if it is still legal at semantic level) owing a different semantic; (2) Example shows that a polysemous character "下" is legal in both word dependency relations but only one boundary cutting is the ground-truth entity (ENG: inferior turbinoplasty).

4. Model

We choose SoftLexicon to encode lexicon information in the character representations. As an encoding scheme, SoftLexicon gets all matched words rather than all matched tags of a certain character. By doing so, every character in the input sequence obtains all matched words when the character locates in the four positions tagged by 'BMES' (Begin, middle, end of a word, and single character forming a word) tagging scheme, subsequently, we introduced not only word boundaries but also word semantics.

The overall structure of this model is shown in Figure 3. First, the model extracts character $c_i \in s = (c_1, c_2, \ldots, c_n)$ from the input sequence containing surgical procedure mentions "Craniotomy Hematoma Clearance", as this case is "肿" (Swelling), then maps it into a dense vector and concatenates with character vector pre-trained by BERT. Meanwhile, sub-sequences where the character "肿" is located at, "血肿" (Hematoma), "颅血肿" (Intracranial Hematoma), etc., are matched with words in lexicon. Weighted representation is calculated according to the frequency of character "肿" in different positions in matched words. For example, "肿" is in the middle of the word "血肿清除术" (Hematoma Evacuation), then words containing this character are categorized into word set "M", where "M" originates from BMES tagging scheme. If there is no matching word starting with "肿", set "B" will be an empty set. To be adjusted to the medical field, we further expanded lexicon so as to obtain more domain knowledge. After the above steps, The SoftLexicon lexicon features are added to characters to form an augmented representation. We utilize the sequence modeling layer and finally have a CRF layer to predict results.

4.1. Character embedding

Each character in the Chinese input sequence is mapped into a dense vector, but due to the lack of large-scale annotated corpora, it is difficult to obtain rich information among characters. The presence of pre-trained language models like BERT is tested effectively on this issue since it learns rich prior semantic knowledge from large unlabeled unstructured text and improves model performance by passing it downstream, refreshing state-of-the-art results on a range of natural language processing tasks [55]. Let $s = (c_1, c_2, \ldots, c_n)$ denotes a sequence of tokens. After BERT embedding, each character c_i is denoted as $BERT(c_i)$ and is later concatenated with character embedding $e^c(c_i)$ to form an augmented embedding as shown in Eq. (1). Here, e^c stands for a lookup table for character embedding.

$$x_i^c = [e^c(c_i); BERT(c_i)] \tag{1}$$

4.2. SoftLexicon

SoftLexicon takes the following 3 steps to incorporate lexicon information.

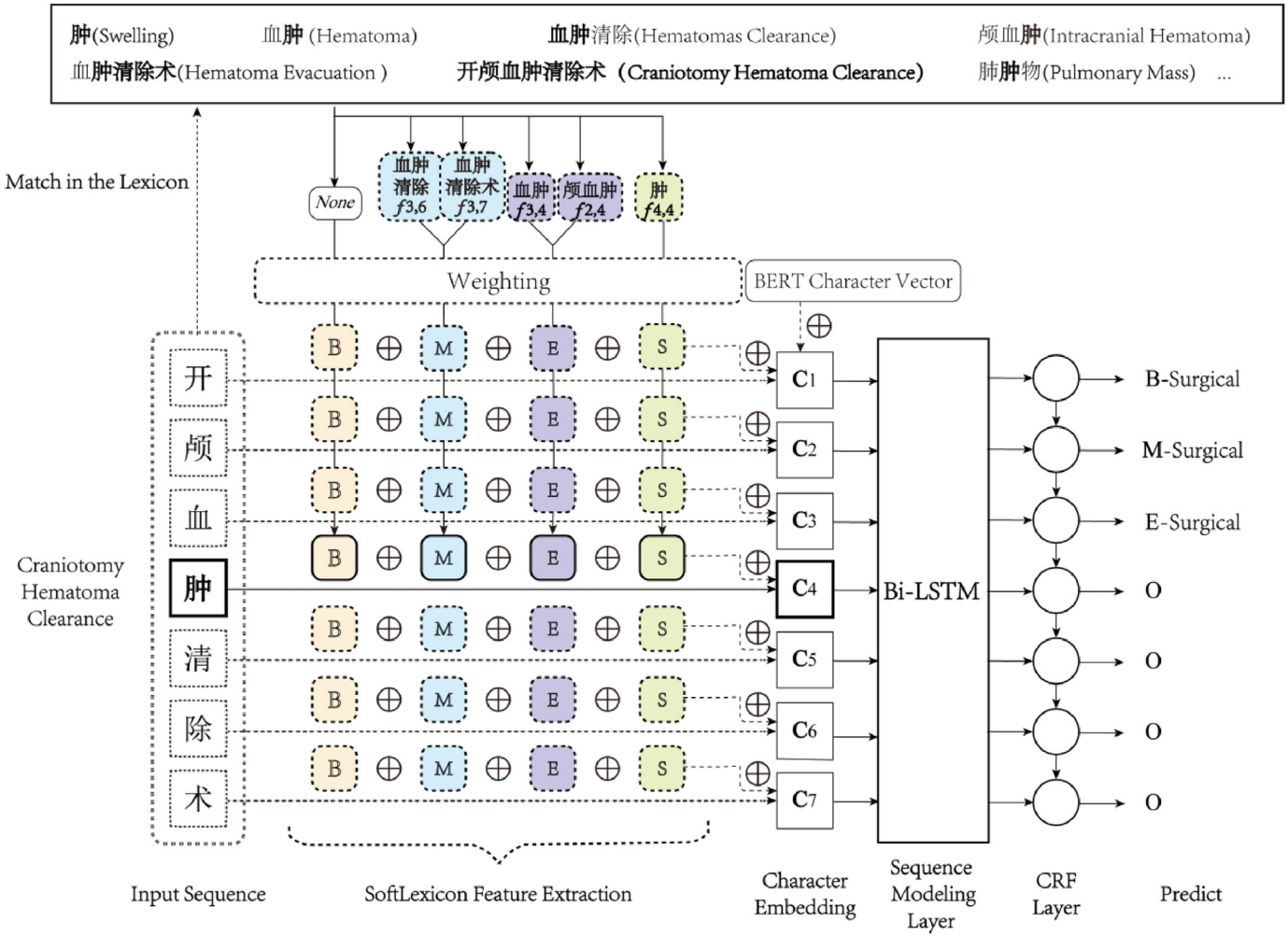

Figure 3. Architecture of SoftLexicon.

4.2.1. Categorizing the matched words

We denote a Lexicon as L. For a sequence of tokens $s = (c_1, c_2, \ldots, c_n)$, if the sub-sequence containing the character c_i matches words in L, it will be grouped into one of the four word sets tagged as 'B', 'M', 'E' and 'S' which is in accordance with "BMES" tagging scheme. The grouping process is in line with Eqs. (2), (3), (4), and (5). If a word set has no matched words inside L, then this set is filled with "*None*" representing an empty set.

$$\mathrm{B}(c_i) = \left\{\mathrm{w}_{i,k}, \forall \mathrm{w}_{i,k} \in \mathrm{L}, \mathrm{i} < \mathrm{k} \leq \mathrm{n}\right\} \tag{2}$$

$$\mathrm{M}(c_i) = \left\{\mathrm{w}_{j,k}, \forall \mathrm{w}_{j,k} \in \mathrm{L}, 1 \leq \mathrm{j} < \mathrm{i} < \mathrm{k} \leq \mathrm{n}\right\} \tag{3}$$

$$\mathrm{E}(c_i) = \left\{\mathrm{w}_{j,i}, \forall \mathrm{w}_{j,i} \in \mathrm{L}, 1 \leq \mathrm{j} < \mathrm{i}\right\} \tag{4}$$

$$\mathrm{S}(c_i) = \{c_i, \exists c_i \in \mathrm{L}\} \tag{5}$$

In Eq. (2), $\mathrm{w}_{i,k}$ represents a word (also a sub-sequence) spanning from a character indexed i to a character indexed k (the word span is surely less than or no more than n-length of input sequence). $\mathrm{w}_{i,k}$ shares the same starting index with the character c_i, which means c_i is the first component character in word $\mathrm{w}_{i,k}$. As Figure 3 shows that, given a certain character c_i from an input sequence, sub-sequences' first component characters start with c_i, i.e., sub-sequences labeled in red and blue in Figure 3, are recognized as $\mathrm{w}_{i,k}$. If these $\mathrm{w}_{i,k}$ exists in L, it will be grouped into word set B. Note that it's possible for a single c_i to produce several $\mathrm{w}_{i,k}$ and among which some are meaningless (sub-sequence labeled in black in Figure 4). Thus, any $\mathrm{w}_{i,k}$ qualifies Eq. (2) is considered a legal element in word set B.

Similarly, Eq. (3) describes that if character c_i is in the middle of all component characters of $\mathrm{w}_{j,k}$ and $\mathrm{w}_{j,k}$ appears in L, then $\mathrm{w}_{j,k}$ will be included in word set M. Eq. (4) describes that if there's a sub-sequence $\mathrm{w}_{j,i}$ ending with character c_i, and $\mathrm{w}_{j,i}$ appears in L, then $\mathrm{w}_{j,i}$ will be included into word set E. Eq. (5) describes that as long as the character c_i exists in L, then c_i is included in the word set S.

4.2.2. Condensing word sets

Words with different lengths will lead to inconsistent word set dimensions, so it is necessary to condense word sets to a fixed one. Considering the unbalanced importance of different words for entity recognition and speeding the training process, word frequency weighting is applied to obtain the weighted representation of word set S in Eq. (6): $v^s(S)$. $z(w)$ is the static word frequency of lexicon word w in statistics, e^w is the lookup table of word embedding, and Z in Eq. (7) is the sum of the frequency of all words in BMES sets.

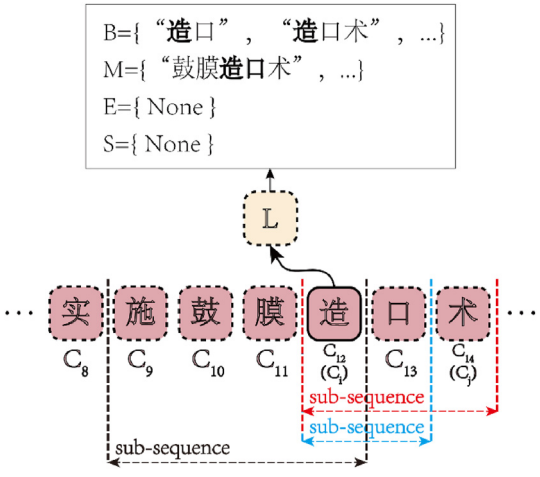

Figure 4. An example of grouping words into word set B.

$$v^s(S) = \frac{4}{Z}\sum_{w \in S} z(w) e^w(w) \tag{6}$$

where

$$Z = \sum_{w \in \mathrm{B} \cup \mathrm{M} \cup \mathrm{E} \cup \mathrm{S}} z(w) \tag{7}$$

4.2.3. Concatenating mutil-granularity features

After repeating (6) to obtain all four word set representations, we merge them into a fixed dimension feature $e^s(B, M, E, S)$ in Eq. (8) and concatenate with character vector to form a new embedding x^c in Eq. (9).

$$e^s(B, M, E, S) = [v^s(B); v^s(M); v^s(E); v^s(S)] \tag{8}$$

$$x^c = [x^c; e^s(B, M, E, S)] \tag{9}$$

4.3. Sequence modeling layer

Bi-directional Long Short-Term Memory (Bi-LSTM) [56] is the mainstream model used in sequence labeling tasks. Bi-LSTM synthesizes forward LSTM [57] and backward LSTM, uses two hidden layers to store information of input data from two directions respectively and concatenates information in step t as $h_t = [h_t^f; h_t^b]$. A LSTM hidden unit is composed of an input gate i_t, forget gate f_t, output gate o_t, and a memory cell c_t. We define $x = [x_1, x_2, \ldots, x_n]$ as input outputted by convolutional neural network (CNN) layer. In step t, the LSTM units can be expressed as:

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \tag{10}$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \tag{11}$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \tag{12}$$

$$\bar{c}_t = \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{13}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \bar{c}_t \tag{14}$$

$$h_t = o_t \odot \tanh(c_t) \tag{15}$$

where W_{xi}, W_{hi}, W_{xf}, W_{hf}, W_{xo} and W_{ho} in Eqs. (10), (11), and (12) represent the weight matrix of input gate i_t, forget gate f_t, and output gate o_t. Parameters b_i, b_f and b_o are bias vectors of the input gate, forget gate, and output gate. Parameters W_{xc}, W_{hc} and b_c in Eq. (13) are the weight matrix and bias vectors of new memory $\bar{c}_t$. $\bar{c}_t$ is the intermediate state of the memory cell waiting to be updated. h_{t-1} is the hidden state when step is $t-1$, σ in Eq. (14) is the sigmoid function, tanh in Eq. (15) is a hyperbolic tangent function [58], and the operator $\odot$ denotes the element-wise multiplication.

4.4. CRF layer

Conditional Random Field can make full use of context information of text labels, and model dependencies among sequence labels to obtain a global optimal prediction. Through tag transition probability, CRF adds sentence-level tag transition information on the basis of Bi-LSTM. For

Table 1. Corpus structure.

Article Title	Term	Context
复方曲肽注射液治疗颅脑术后脑水肿的临床效果及对不良反应、血清肿瘤坏死因子-α水平的影响	颅脑损伤开颅术	颅脑损伤开颅术是颅脑损伤科临床上常见的手术之一，是指通过机械设备来打开患者的颅骨后进行手术，从而达到清除颅内血肿，减轻颅内压力

Table 2. Detailed corpus statistics.

Type	Train	Dev	Test
Sentence	2.2k	0.2k	0.2k
Char	183.2k	22.9k	22.9k

input sequence $X = (x_1, x_2, \ldots, x_T,)$, the corresponding predicted tag sequence is $Y = (y_1, y_2, \ldots, y_T)$, and the corresponding tag sequence score $Score(X, I)$ is obtained by adding tag transition matrix A_{i_{t-1}, i_t} and tag probability matrix $F_{i_t, t}$ output by Bi-LSTM, as shown in Eq. (16). During training, the loss function is calculated using maximum likelihood estimation in Eq. (17), where U is the set of all possible tag sequences, and $I^{'}$ is a subset of U. Finally, the Viterbi algorithm in Eq. (18) is used to calculate the maximum probability sequence to obtain the optimal output I^{*}.

$$Score(X, I) = \sum_{t=1}^{T} \left(A_{i_{t-1}, i_t} + F_{i_t, t}\right) \tag{16}$$

$$Log(P(I|X)) = Score(X, I) - \log\left(\sum_{U} Score(X, I^{'})\right) \tag{17}$$

$$I^{*} = arg_U, \max(X, I^{'}) \tag{18}$$

5. Experiments

We evaluate the proposed model SSBC on a self-built corpus and compare its effectiveness with other state-of-the-art models. Corpus, parameter settings and results of our experiments are described below.

5.1. Corpus

Considering a large annotated corpus of this field is not available, we, therefore, built a corpus from scratch. By entering the search query "YE = 2018–2022 AND (TI = 手术 OR AB = 手术)"on CNKI (China National Knowledge Infrastructure) online journal database, literature numbered 1232 was finally included to meet the quantity requirement. Then we manually extracted surgical terms from the full text of literature under the guidance of clinical experts. We segmented the extracted corpus containing 2600 terms by BMES tagging scheme. Table 1 and Table 2 respectively portray corpus in rows and detailed statistics. In experiments, the corpus is randomly divided into train, dev, and test set in proportion to 8:1:1.

Selecting published journal articles other than others as our corpus source has 3 reasons: (1) peer-review and publication mechanism of journal articles enable surgical terms that appeared in literature to be more standardized and well-recognized; (2) closer to natural language; (3) more open to access and not restricted by medical safety and privacy rules.

5.2. Environment setting

We choose 2.10 GHz CPU, NVIDIA Tesla V100-16GB GPU, 64G RAM, Ubuntu 16.04 OS. In developing tool and compiling environment, we utilize Python 3.6 and PyCharm 2021.

Table 3. Hyper parameter setting.

Parameter	Value	Parameter	Value
char_emb_size	50	epoch	100
bichar emb size	50	dropout	0.5
lattice_emb_size	50	learning rate	0.005
batch_size	10	learing rate dacay	0.05
LSTM hidden_dim	200	LSTM_layer	1

Table 4. Encyclopedia entry words details.

Category	Size	Source
Leading term & Operative approach	372	*Surgical Procedure Classification Code National Clinical Edition3.0*
Anatomy (body)	1,084	*Chinese Medical Subject Headings (CMeSH)*
Disease	2,813	*Disease Classification and Codes (GB T14396-2016)*
Equipment	6,609	*Medical Device Classification Catalog 2017 Edition*
Comprehensive	18,749	*Tsinghua University Open Chinese Thesaurus (Medical)*

5.3. Hyper-parameters

Table 3 shows our hyper parameters setting during the training.

5.4. Lattice training

In addition to inheriting the general-domain lattice embedding used in paper [51], we construct a specialized lattice for surgical procedure terms using an encyclopedia and open access corpus. We choose terms in seven fields mentioned in section 2 as search queries to obtain encyclopedia content. Table 4 shows details about entry words. In addition, we utilize the Chinese Medical Conversational Question-Answer (CMCQA) dataset [59] as a supplementary training corpus, which is a huge conversational question answering dataset in Chinese medical field collecting dialogue materials from the medical dialogue Q&A website Chunyu covering 45 departments, 1.3 million complete dialogues and 19.83 million sentences with a total of 2.84GB. We fuse corpus from the two sources, then perform automatic word segmentation, and use Word2Vec [60] for 50-dimension word vector training. We compared model performance using different embeddings.

5.5. Evaluating matrix

The traditional practice to calculate the values of precision, recall, and $F1$-score is based on the classification results of the NER model. We follow this set of evaluation to reflect the performance of our model in classifying samples into the desired category. Precision (P) refers to the ratio of correct entities to predict entities. Recall (R) refers to the ratio of the entities in the test set which are correctly predicted. $F1$-score is the harmonic average value of P and R measuring overall performance, and is calculated by Eq. (19):

$$F1 = \frac{2{*}PR}{P + R} \times 100\% \tag{19}$$

In this work, we conduct several comprehensive experiments to verify the proposed module. The comparison models include (1) BERT-Bi-LSTM-CRF [61], a BERT model with a bidirectional long-short term memory network (LSTM) followed by a CRF layer that is the most popular architecture for NER task, but it still lacks the ability to leverage lexicon information; (2) BERT-FLAT-Lattice-Transformer [50], a BERT model with a flat-lattice Transformer incorporating lexicon information, but it is too complicated in structure and is hard to deploy just in

Table 5. The performances of different models in the test set of self-built corpus.

Model	%		F1
	P	R	
BERT-Bi-LSTM-CRF	64.71	**91.67**	75.86
BERT-FLAT-Lattice-Transformer	62.61	72.02	67.09
SSBC	70.52	85.91	77.46
Bichar-SSBC	71.47	86.58	78.30
BERT-SSBC	**73.24**	87.25	**79.63**

The bold font denotes the best result for each column

Table 6. Outcomes using different lattice embeddings.

Model	+General Lattice			+General & Medical Lattice			+Medical Lattice		
	P	R	F1	P	R	F1	P	R	F1
SoftLexicon (LSTM)	70.52	85.91	77.46	71.15	86.91	78.25	**75.45**	83.56	79.30
SoftLexicon (LSTM)+bichar	71.47	86.58	78.30	71.11	85.91	77.81	72.10	87.58	79.10
SoftLexicon (LSTM)+BERT	**73.24**	**87.25**	**79.63**	**71.43**	**87.25**	**78.55**	72.25	**88.26**	**79.46**

embedding layer; (4) SSBC, our proposed model simply fuse lexicon-character level feature without pre-trained language model; and (4) Bichar-SSBC, our proposed model with bichar feature embedding. The model comparison, in theory, can prove the importance of lexicon information and a pre-trained language model. The experiment results are shown in Table 5.

To see how domain knowledge to influence lexicon-based model, different combination of lattice embeddings is used, outcomes seeing in Table 6.

6. Discussion

First, our BERT-SSBC achieves the highest *F*1-score at 79.63% comparing the other four models, surpassing baseline models BERT-Bi-LSTM-CRF and BERT-FLAT-Lattice-Transformer by 3.77% and 12.54% respectively. The other 3 models based on SoftLexicon show competitiveness in evaluating indexes since their *F*1 score gap is no more than 2.5%, and the one with BERT achieves the highest *F*1 score of 79.46%, implicating the superiority of SoftLexicon structure and the effectiveness brought by rich prior semantic knowledge. SSBC also has the highest *P* score implying that the usage of lexicon greatly reduced segmentation errors.

Second, as proved in paper [51], adding word-level characteristics such as bigram is of great importance to represent character information. By utilizing pre-trained bigram embedding of Chinese Giga-Word, some entity boundary segmentation errors are reduced, and the model obtains lots of word information based on character. Compared with BERT-Bi-LSTM-CRF, bichar + SSBC behaves better in both *F*1 and *P* score.

Third, SoftLexicon behaves much better than FLAT-Lattice-Transformer even though they used the same lattice. This partly attributes to: (1) the unique structural feature specified in the surgical procedure field. (2) the design of the "Middle" group enables a character to exploit word information from the words that begin or end with it, but also the information of the words that contain the character.

Fourth, due to the characteristics of SoftLexicon, the use of different domain lexicons has an important impact on the performance of the model. Thus, we examine this effect by adding lexicons from different domains to the model. As shown in Table 6, models using medical lexicon tend to be better in most evaluating indexes. BERT-SSBC always achieves the best in every index even lexicon changes. Medical lexicon competes with the general lexicon fiercely implying that domain knowledge is as important as general domain knowledge. When considering that the size of the medical lexicon is only 1% big as the general lattice, we find high quality information representation can be obtained from specialized fields effectively.

Our model correctly predicts more than 70% of long terms especially with a small size of the pre-trained lexicon, showing its room for improvement by adding pre-trained corpus and its transferability to various domain-specific fields. But it is still not enough for application in daily routine processing since the precision rate is significantly lower than the recall rate, indicating its flaw in recognizing entities with longer lengths. Compare with small entities, long-term entities tend to be more unitary, and with more inter-entity nesting phenomena and blurred boundary truncation. Focusing on innovations beyond models and algorithms is hopefully a feasible way.

7. Conclusion and future work

Aiming at recognizing long terms in surgical procedure, this article applies a deep learning model that integrates field structural features, multi-granularities information, and pre-training language model BERT to alleviate the problem of challenging recognition of surgical terms in Chinese literature text. Through experiments on the self-built data set, our method proves to have an effective improvement in the performance of Chinese NER by reaching a new high F1 value of 79.63%. Moreover, we find the effect of domain-specific lexicon is no worse than that of general domain showing its room for continuous improvement and potential for transferability.

Our study has several limitations. This study builds data set by extracting related contexts from research articles which may cause a loss of information presented elsewhere in the full-text document. For instance, we found a prevailing description style of sentences where surgical procedure terms locate. This familiarity in contexts doesn't fit very well with scenarios in reality, therefore limiting the model's transferability in various domain-specific fields. For future work, it is important to consider extracting information from various full-text sources. Additionally, it is still not enough for application in daily routine processing since the precision rate is significantly lower than the recall rate, indicating its flaw in recognizing entities with longer lengths. For future work, to substantiate the usefulness of the system, it is important to test it on a larger lattice and more complex corpus. Finally, in-depth research of polysemy, abbreviations, relationships among entities, and the normalization of entities are the next tasks in our future work.

Declarations

Author contribution statement

Nan Jiale: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Gao Dongping: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Sun Yuanyuan: Contributed reagents, materials, analysis tools or data; Analyzed and interpreted the data.

Li Xiaoying: Analyzed and interpreted the data.

Shen Xifeng; Li Meiting; Zhang Weining; Ren Huiling; Qin Yi: Contributed reagents, materials, analysis tools or data.

Funding statement

Dongping Gao was supported by National Key Research and Development Program of China [2020AAA0104905].

Data availability statement

Data will be made available on request.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

This work was supported by the National Key Research and Development Program of China [grant ID 2020AAA0104905].

References

[1] Y. Cheng, T.L. Jiang, L.Z. Deng, L.M. Chen, K. Jiang, Research on the coverage of standard Chinese medical terminology to practical application, Chinese Journal of Health Informatics and Management 17 (5) (2020) 601–605+636.
[2] J. Yin, S. Luo, Z. Wu, L.M. Pan, Chinese named entity recognition with character-level BLSTM and soft attention model, J. Beijing Inst. Technol. 29 (1) (2020) 12.
[3] Z. Tang, B. Wan, L. Yang, Word-character graph convolution network for Chinese named entity recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 1520–1532.
[4] Q. Zhu, X. Li, A. Conesa, C. Pereira, GRAM-CNN: A deep learning approach with local context for named entity recognition in biomedical text, Bioinformatics 34 (9) (2018) 1547–1554.
[5] H.M. Yang, L. Li, R.D. Yang, Named entity recognition based on bidirectional long short-term memory combined with case report form, Chinese Journal of Tissue Engineering Research 22 (2018) 3237–3242.
[6] F.C. Zhang, Q.L. Qing, Y. Jiang, R.T. Zhuang, Research on Chinese EMR named entity recognition based on RoBERTa-WWM-BiLSTM-CRF, Data Analysis and Knowledge Discovery 6 (Z1) (2022) 251–262.
[7] S.Q. Jing, Y.L. Zhao, Recognizing clinical named entity from Chinese electronic medical record texts based on semi-supervised deep learning, J. Inf. Rec. Mater. 11 (6) (2021) 105–115.
[8] Q.Q. Qu, H.X. Kan, Named entity recognition of Chinese medical text based on Bert BiLSTM CRF, Electronic Design Engineering 29 (19) (2021) 5.
[9] H. Yu, C. Xu, Y.R. Liu, Y.W. Fu, D.P. Gao, BiLSTM and CRF-based extraction of therapeutic events from Chinese clinical guidelines, Chinese Journal of Medical Library and Information Science 29 (2) (2020) 6.
[10] L. Yang, X.S. Huang, J.Y. Wang, L.L. Ding, Z.X. Li, J. Li, Clinical trial disease subtype identification based on BERT-TextCNN, Data Analysis and Knowledge Discovery 6 (4) (2022) 69–81.
[11] H.Y. Wu, D. Lu, M. Hyder, S. Zhang, S.K. Quinney, DrugMetab: an integrated machine learning and lexicon mapping named entity recognition method for drug metabolite, CPT Pharmacometrics Syst. Pharmacol. 7 (2018).
[12] I. Lerner, N. Paris, X. Tannier, Terminologies Augmented Recurrent Neural Network Model for Clinical Named Entity Recognition, 2019.
[13] S.K. Wang, S.Z. Li, T.S. Chen, Recognition of Chinese medicine named entity based on condition random field, J. Xiamen Univ. 48 (3) (2009) 359–364.
[14] Y.H. Liang, W.J. Zhang, D.F. Zhou, A hybrid strategy for high precision long term extraction, J. Chin. Inf. Process. 23 (6) (2009) 26–30.
[15] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.
[16] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proc. 18th Int. Conf. Mach. Learn., Williamstown, MA, USA, 2001, pp. 282–289. Jun./Jul.
[17] National Health Commission of the People's Republic of China, Official Announcement for Hospital Administration, 2020. http://www.nhc.gov.cn/yzygj/s7659/202006/7912483be2784e2ca08a9ea4628369b8.shtml (accessed 8 December 2021).
[18] F.C. Zhang, Q.L. Qing, Y. Jiang, R.T. Zhuang, Research on Chinese EMR named entity recognition based on RoBERTa-WWM-BiLSTM-CRF, Data Analysis and Knowledge Discovery 6 (Z1) (2022) 251–262.
[19] S.Q. Jing, Y.L. Zhao, Recognizing clinical named entity from Chinese electronic medical record texts based on semi-supervised deep learning, J. Inf. Rec. Mater. 11 (6) (2021) 105–115.
[20] Q.Q. Qu, H.X. Kan, Named entity recognition of Chinese medical text based on Bert•BiLSTM•CRF, Electronic Design Engineering 29 (19) (2021) 5.
[21] Z. Liu, C. Luo, Z. Zheng, Y. Li, D. Fu, X. Yu, J. Zhao, TCMNER and PubMed: a novel Chinese character-level-based model and a dataset for TCM named entity recognition, J Healthc Eng (2021 Aug 7), 3544281.
[22] H. Yu, C. Xu, Y.R. Liu, Y.W. Fu, D.P. Gao, BiLSTM and CRF-based extraction of therapeutic events from Chinese clinical guidelines, Chinese Journal of Medical Library and Information Science 29 (2) (2020) 6.
[23] L. Yang, X.S. Huang, J.Y. Wang, L.L. Ding, Z.X. Li, J. Li, Clinical trial disease subtype identification based on BERT-TextCNN, Data Analysis and Knowledge Discovery 6 (4) (2022) 69–81.
[24] E. Shijia, Y. Xiang, M. Assoc Comp, Chinese Named Entity Recognition with Character-word Mixed Embedding. ACM Conference on Information and Knowledge Management (CIKM), Assoc Computing Machinery, Singapore, SINGAPORE, 2017, pp. 2055–2058.
[25] F.W. Mutinda, K. Liew, S. Yada, S. Wakamiya, E. Aramaki, Automatic data extraction to support meta-analysis statistical analysis: a case study on breast cancer, BMC Med. Inf. Decis. Making 22 (1) (2022 Jun) 158.
[26] H. Cho, W. Choi, H. Lee, A method for named entity normalization in biomedical articles: application to diseases and plantss, BMC Bioinf. 18 (1) (2017 Oct 13) 451.
[27] Yoshimasa Tsuruoka, Jun'ichi Tsujii, Improving the performance of dictionary based approaches in protein name recognition, J. Biomed. Inf. 37 (6) (2004) 461–470.
[28] R.T. Tsai, C.L. Sung, H.J. Dai, H.C. Hung, T.Y. Sung, W.L. Hsu, NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition, BMC Bioinf. 7 (Suppl 5) (2006 Dec 18) S11.
[29] Min Jiang, Yukun Chen, Mei Liu, S. Trent Rosenbloom, Subramani Mani, Joshua C. Denny, Hua Xu, A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries, J. Am. Med. Inf. Assoc. 18 (5) (2011) 601–606.
[30] Jianbo Lei, Buzhou Tang, Xueqin Lu, Kaihua Gao, Min Jiang, Hua Xu, A comprehensive study of named entity recognition in Chinese clinical text, J. Am. Med. Inf. Assoc. 21 (5) (2014) 808–814.
[31] P. Ma, B. Jiang, Z. Lu, N. Li, Z. Jiang, Cybersecurity named entity recognition using bidirectional long short-term memory with conditional random fields, Tsinghua Sci. Technol. 26 (3) (2021) 11–17.
[32] Q. Zhang, L. Hou, P. Lv, M. Zhang, H. Yang, Chinese medical entity recognition model based on character and word vector fusion, Sci. Program. 2021 (2021) 1–12, 5933652:1-5933652:12.
[33] G. Zhou, J. Su, Named entity recognition using an HMM-based chunk tagger, in: Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, Philadelphia, PA, USA, 2002, pp. 473–480.
[34] C. Cortes, V. Vapnik, Support vector network, Mach. Learn. 20 (1995) 273–297.
[35] J. Liang, X. Xian, X. He, et al., A novel approach towards medical entity recognition in Chinese clinical text, Journal of Healthcare Engineering (2017), 4898963. Article ID 4898963, 2017.
[36] Q. Wang, M. Iwaihara, Deep neural architectures for joint named entity recognition and disambiguation, in: IEEE International Conference on Big Data and Smart Computing, BigComp, 2019.
[37] H. Tan, Z. Yang, J. Ning, Z. Ding, Q. Liu, Chinese medical named entity recognition based on Chinese character radical features and pre-trained language models, in: 2021 International Conference on Asian Language Processing, IALP, 2021, pp. 121–124.
[38] M. Habibi, L. Weber, M. Neves, D.L. Wiegandt, U. Leser, Deep learning with word embeddings improves biomedical named entity recognition, Bioinformatics 33 (2017) i37–i48.
[39] M. Cho, J. Ha, C. Park, S. Park, Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition, J. Biomed. Inf. 103 (2020) 8.
[40] Y. Li, G. Du, Y. Xiang, S. Li, H. Chen, Towards Chinese clinical named entity recognition by dynamic embedding using domain-specific knowledge, J. Biomed. Inf. 106 (2020), 103435.
[41] Y. Jin, J. Xie, W. Guo, C. Luo, R. Wang, LSTM-CRF Neural Network with Gated Self Attention for Chinese NER, IEEE Access, 2019, p. 1.
[42] H. Zhao, C. Kit, Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition, in: Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, 2008.
[43] H.Y. Wu, D. Lu, M. Hyder, S. Zhang, S.K. Quinney, DrugMetab: an integrated machine learning and lexicon mapping named entity recognition method for drug metabolite, CPT Pharmacometrics Syst. Pharmacol. 7 (2018).
[44] I. Lerner, N. Paris, X. Tannier, Terminologies Augmented Recurrent Neural Network Model for Clinical Named Entity Recognition, 2019.
[45] C.Y. Hou, M.L. Wang, C.L. Li, Entity Subword Encoding for Chinese Long Entity Recognition. 4th China Conference on Knowledge Graph and Semantic Computing (CCKS), Springer-Verlag Singapore Pte Ltd, Hangzhou, PEOPLES R CHINA, 2019, pp. 123–135.
[46] H. Kim, J. Kang, How do your biomedical named entity recognition models generalize to novel entities? IEEE Access 10 (2022 Mar 8) 31513–31523.
[47] S. Zhao, Z. Cai, H. Chen, Y. Wang, A. Liu, Adversarial training based lattice lstm for Chinese clinical named entity recognition, J. Biomed. Inf. 99 (14) (2019), 103290.
[48] Y. Xu, Y.N. Wang, T.R. Liu, J.H. Liu, Y.B. Fan, Y. Qian, et al., Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries, J. Am. Med. Inf. Assoc. 21 (2014) E84–E92.
[49] K.X. Liu, Q.C. Hu, J.W. Liu, C.X. Xing, Ieee. Named Entity Recognition in Chinese Electronic Medical Records Based on CRF. 14th Web Information Systems and Applications Conference (WISA), IEEE, Liuzhou, PEOPLES R CHINA, 2017, pp. 105–110.
[50] Xiaonan Li, Hang Yan, Xipeng Qiu, Xuanjing Huang, FLAT: Chinese NER using flat-lattice transformer, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online. Association for Computational Linguistics, 2020, pp. 6836–6842.
[51] Y. Zhang, J. Yang, Chinese NER using lattice LSTM, in: 56th Annual Meeting of the Association-For-Computational-Linguistics (ACL), Assoc Computational Linguistics-Acl, Melbourne, Australia, 2018, pp. 1554–1564.
[52] J.R. Cheng, J.X. Liu, X.B. Xu, D.W. Xia, L. Liu, V.S. Sheng, A review of Chinese named entity recognition, KSII Trans Internet Inf Syst 15 (2021) 2012–2030.
[53] Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, Xuanjing Huang, Simplify the usage of lexicon in Chinese NER, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online. Association for Computational Linguistics, 2020, pp. 5951–5960.
[54] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.
[55] 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, 2016.
[56] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Network. 18 (5–6) (2005) 602–610.
[57] Sepp Hochreiter, Jürgen Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[58] M. Rhif, A. B. Abbes, B. Martinez, and I. R. Farah, A deep learning approach for forecasting non-stationary big remote sensing time series, Arabian J. Geosci., vol. 13.
[59] Yixuan Weng, CMCQA, 2022. https://github.com/WENGSYX/CMCQA (accessed 8 December 2021).
[60] Mikolov I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed Representations of Words and Phrases and Their Compositionality, in: Proc. Adv. Neural Inf. Process. Syst., Lake Tahoe, NV, USA, Dec. 2013, pp. 3111–3119.
[61] Hui Li, Yu Lin, Jie Zhang, Ming Lyu, Fusion deep learning and machine learning for heterogeneous military entity recognition, Wireless Commun. Mobile Comput. 2022 (2022) 1–11. Article ID 1103022, 11 pages, 2022.