A BERT-Span model for Chinese named entity recognition in rehabilitation medicine

Jinhong Zhong*, Zhanxiang Xuan*, Kang Wang and Zhou Cheng

School of Management, Hefei University of Technology, Hefei, Anhui, China
* These authors contributed equally to this work.

ABSTRACT

Background: Due to various factors such as the increasing aging of the population and the upgrading of people's health consumption needs, the demand group for rehabilitation medical care is expanding. Currently, China's rehabilitation medical care encounters several challenges, such as inadequate awareness and a scarcity of skilled professionals. Enhancing public awareness about rehabilitation and improving the quality of rehabilitation services are particularly crucial. Named entity recognition is an essential first step in information processing as it enables the automated extraction of rehabilitation medical entities. These entities play a crucial role in subsequent tasks, including information decision systems and the construction of medical knowledge graphs.
Methods: In order to accomplish this objective, we construct the BERT-Span model to complete the Chinese rehabilitation medicine named entity recognition task. First, we collect rehabilitation information from multiple sources to build a *corpus* in the field of rehabilitation medicine, and fine-tune Bidirectional Encoder Representation from Transformers (BERT) with the rehabilitation medicine *corpus*. For the rehabilitation medicine *corpus*, we use BERT to extract the feature vectors of rehabilitation medicine entities in the text, and use the span model to complete the annotation of rehabilitation medicine entities.
Result: Compared to existing baseline models, our model achieved the highest F1 value for the named entity recognition task in the rehabilitation medicine *corpus*. The experimental results demonstrate that our method outperforms in recognizing both long medical entities and nested medical entities in rehabilitation medical texts.
Conclusion: The BERT-Span model can effectively identify and extract entity knowledge in the field of rehabilitation medicine in China, which supports the construction of the knowledge graph of rehabilitation medicine and the development of the decision-making system of rehabilitation medicine.

Subjects Artificial Intelligence, Graphics, Text Mining, Neural Networks
Keywords Rehabilitation medicine, Named entity recognition, Text annotation, BERT-Span

Submitted 17 May 2023
Accepted 21 July 2023
Published 21 August 2023

Corresponding author
Zhanxiang Xuan,
xuanzhanxiang@163.com

Academic editor
Shibiao Wan

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj-cs.1535**

INTRODUCTION

With the steady growth of the economy and the continuous improvement of living standards, there is an escalating demand to enhance the quality of life for individuals. In addition to disease prevention, clinical treatment, and physical health care, rehabilitation medicine is receiving growing attention as a crucial approach to assist patients with

How to cite this article Zhong J, Xuan Z, Wang K, Cheng Z. 2023. A BERT-Span model for Chinese named entity recognition in rehabilitation medicine. **PeerJ Comput. Sci. 9:e1535 DOI 10.7717/peerj-cs.1535**

disabilities in regaining their daily living abilities and enhancing their overall quality of life. According to statistics from a study by The Lancet (*Feng et al., 2020*), China has the highest rehabilitation needs globally, totaling 460 million people. According to the China Health and Wellness Statistical Yearbook from 2012 to 2020 (*Committee NHaW, 2012–2020*), the number of outpatient visits to rehabilitation hospitals in China surged from more than six million to more than 11 million. The increasing demand for rehabilitation medical care in China is evident. However, China's rehabilitation medical facilities suffer from inadequate resources, with many hospitals lacking dedicated rehabilitation medicine departments. Additionally, there is a shortage of professional rehabilitation physicians, with an average of less than six rehabilitation physicians and fewer than ten rehabilitation therapists per 100,000 population in China (*Committee NHaW, 2021*). Consequently, a substantial imbalance exists between the supply and demand for rehabilitation services in the country.

Knowledge graph, as a state-of-the-art technology, plays a pivotal role in transforming complex unstructured data into a structured "entity-relationship-entity" format (*Ji et al., 2021*). This transformation facilitates downstream tasks such as question and answer systems, decision systems, and information management systems. Leveraging the knowledge graph to organize rehabilitation data enables us to accelerate the advancement of the rehabilitation healthcare system, enhance the training of rehabilitation doctors, and enhance national awareness of rehabilitation, ultimately alleviating the shortage of rehabilitation resources in the country. As a crucial component in knowledge graph construction, named entity recognition (NER) plays a vital role in automatically extracting predefined entities from extensive and intricate texts, thereby facilitating the structuring of information (*Li et al., 2020a*). Earlier NER is primarily performed based on rules and dictionaries (*Fukuda et al., 1998*; *Keretna, Lim & Creighton, 2014*). While these methods demonstrate high accuracy, their heavy reliance on expert-defined rules hinders their extensibility. Later, the technology of machine learning break through and scholars start to apply statistical methods to NER, such as hidden Markov models, support vector machines (SVM), conditional random fields (CRF) (*Keretna et al., 2014*; *Lei et al., 2014*), *etc*. These methods demonstrate excellent performance with the support of rich corpora. Nevertheless, constructing such corpora requires substantial manual effort, and the *corpus*-specific features pose challenges in terms of portability and generalizability. With the enhancement of computer computing power and the surge of massive data, deep learning with powerful generalization ability has gradually become a research point in the field of named entity identification.

At present, apart from some voluntary forms of rehabilitation communities on the Internet such as, Hong Kong Society for Rehabilitation (https://www.rehabsociety.org.hk/zh-hans/) and Medical Pulse (https://www.medlive.cn/), no unified research in the field of rehabilitation medicine has been found for the time being. Scattered information about rehabilitation on web pages lacks credibility, so we chose to extract reliable rehabilitation information from thesis literature and books. We first search for CNKI literature (https://www.cnki.net) and authoritative guide books to build a *corpus* in the field of rehabilitation medicine in a comprehensive and multifaceted way, and use the annotation consistency principle to verify the accuracy of the *corpus*. In terms of model selection, the BERT demonstrates robust linguistic representation and feature extraction capabilities, and Span can handle long medical entities and nested entities very well. Hence we combine the features of BERT and span to construct the BERT-Span model for the named entity recognition task in the rehabilitation domain. The experimental results indicate that the model achieves a higher F1 score in named entity identification studies in rehabilitation medicine compared to several baseline models.

The main contributions of this article are as follows:

(1) In view of the complexity and diversity of rehabilitation medical texts and the remoteness of vocabulary, we have formulated corresponding annotation specifications, adopted BIO annotation method, and have formed a certain scale of *corpus* in rehabilitation medicine.

(2) Combined with the results of annotation and the knowledge in the field of rehabilitation medicine, we propose the BERT-Span model that is more suitable for the task of Chinese named entity recognition in rehabilitation medicine, which fills the gap in NER in the field of rehabilitation medicine to a certain extent.

RELATED WORK

Basic neural network

Initially, convolutional neural networks (CNN), recurrent neural networks (RNN), Bi-directional long short-term memory (BiLSTM) and other basic neural networks can achieve good recognition results in NER tasks (*Li et al., 2021*; *Schuster & Paliwal, 1997*; *Zhang et al., 2015*). *Dong et al. (2016)* propose a multi-classification method for Chinese electronic medical records based on CNN, and achieve better results than traditional machine learning methods such as SVM and CRF in the self-constructed *corpus*. Later, *Dong et al. (2017)* propose to extract entities from Chinese electronic medical records using a multi-task bidirectional RNN model. It can further improve the performance of NER by determining the parameter weight through multi-layer bidirectional RNN. *Xu et al. (2018b)* train the BiLSTM-CRF model in the supervised *corpus* for medical named entity recognition. *Xu et al. (2018a)* also propose a similar BiLSTM-CRF model to specifically identify disease entities, which reaches 86.20% of F1 in NCBI disease data set. *Qin & Zeng (2018)*, on the basis of BiLSTM-CRF, use character based word embedding and continuous Bag-of-Words embedding (CBOW) as the input of the model to capture more representations. *Florez et al. (2018)* also improve the input of the BiLSTM-CRF model. They connect character level representation and part of speech features to form a comprehensive word representation, and then input the BiLSTM network to identify medical entities from clinical records.

Neural network integrating attention mechanism

With the development of the research, the simple neural network model has exposed some problems. For example, global features can not be obtained, and information between long sentences is seriously missing. The concept of weight distribution in attention mechanism can effectively solve such problems. Some scholars commence adding attention mechanism to BiLSTM-CRF for named entity recognition. *Ji et al. (2019)* add the attention mechanism to the BiLSTM-CRF model to conduct medical NER research on Chinese electronic medical records, and obtain 90.82% of F1 scores in the open dataset. *Wu et al. (2020)* use attention mechanism to complete NER and intention analysis tasks in medical problems based on the BiLSTM-CRF model. *Li et al. (2019a)* combine attention mechanism with neural network (BiLSTM ATT-CRF), and introduce medical dictionaries and part of speech features, which confirmed the effectiveness of attention mechanism in obtaining long sentence information and text features. *Wu et al. (2019)* propose the Att-BiLSTM-CRF model to capture long-range dependencies using a self-attentive mechanism that combines a new character-level representation and POS annotation information to capture the semantic information of the input sentences, taking into account the own characteristics of Chinese. In order to better explain the meaning of Chinese characters, *Yin et al. (2019)* use CNN to extract radical-level features on the basis of the BiLSTM-CRF model, aiming to capture the internal relevance of characters, and use the self-attention mechanism to capture the dependency between characters. *An et al. (2022)* add the multi-head self-attention mechanism on the basis of BiLSTM-CRF and proposed the MUSA BiLSTM-CRF model, which further improve the NER in terms of global information and multi-level semantic information acquisition. In addition, some scholars try to improve the BiLSTM layer of traditional models to improve the model recognition effect. *Zhao et al. (2019)* propose a convolutional neural network (AT lattice LSTM-CRF) based on attention mechanism. Lattice-LSTM integrates character level information and word-level information, making up for the defect of single information. *Li et al. (2020b)* put forward the ELMo lattice LSTM-CRF model, which uses lattice LSTM to obtain word and character information. The variant ELMo model uses Chinese characters as input to learn context information in specific fields. Some scholars use Transformer to code instead of BiLSTM. *Wan et al. (2020)* propose a method based on ELMo-ET-CRF model. Compared with ELMo-LSTM-CRF model, ELMo-ET-CRF has better performance.

Transfer learning

However, numerous tasks in NER require a large amount of annotated data, and NER in some new fields do not meet these conditions. Transfer learning focuses on storing existing problem solving models and using them on other different but related problems (*Pan & Yang, 2010*). Therefore, some scholars put forward the idea of transfer learning to complete this task. *Bhatia, Arumae & Busra Celikkaya (2020)* learn the appropriate parameter sharing scheme between the source dataset and the target dataset through the dynamic transmission network (DTN), so as to achieve the best NER result on a smaller target dataset. *Dong et al. (2019)* propose to train the shallow bidirectional RNN network on the general domain data set first, then migrate the weights of the trained parameters to the deep bidirectional RNN network, and train again on the Chinese electronic medical record. *Gong, Zhang & Chen (2020)* use a *corpus* of the same domain to pre-train word embedding and subsequently fine-tune the pre-training model of the target *corpus* using the model parameters obtain from the training to overcome the challenge of a small annotated *corpus* of Chinese clinical entities. Later, with its strong language representation ability and feature extraction ability, BERT make outstanding achievements in the field of named entity recognition (*Devlin et al., 2018*). *Xue et al. (2019)* use BERT to extract the word vector of the text as the input of the subsequent model. *Zhang et al. (2019)* propose the BERT-BiLSTM-CRF pre-training language model. The experimental results show that compare with the previous model, the BiLSTM-CRF model combines with BERT has the most outstanding effect in Chinese electronic medical record entity recognition. *Li, Zhang & Zhou (2020)* use unlabeled Chinese clinical text to pre-training BERT model to learn domain specific knowledge, and use BiLSTM and CRF to extract text features and decode predictive tags respectively. *Liu et al. (2022)* propose the Med BERT-Span FLAT model, in which Med BERT optimizes the representation of long medical entities, and Span-FLAT replaces the traditional CRF and has a better effect on nested entity annotation.

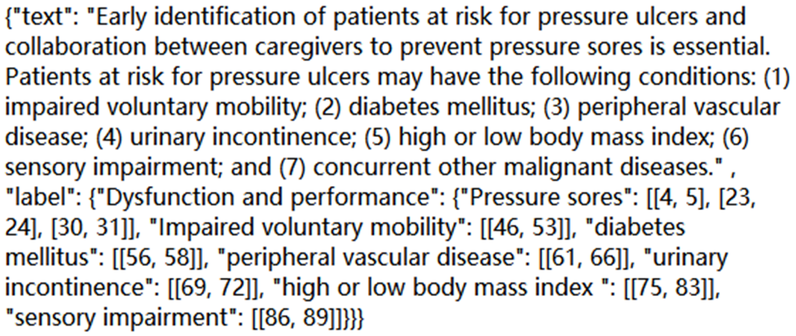
In summary, the algorithmic technology employed for NER tasks is steadily advancing, and its application in the medical field has yielded high accuracy results. However, in certain specialized fields, the presence of intricate and less familiar vocabulary, coupled with limited available corpora, poses challenges to named entity recognition. Currently, no relevant NER studies have been found in the field of rehabilitation medicine. With the maturity of migration learning, the recognition of named entities in these new fields can no longer require a large amount of annotated *corpus* to be prepared in advance to have a good recognition effect. Therefore, in this article, we focus on the task of named entity recognition in the field of Chinese rehabilitation medicine. In view of the fact that there is no *corpus* in this field at present, we collect rehabilitation data from CNKI literature database, rehabilitation books *etc*. to build a small-scale rehabilitation *corpus*, and verify the reliability of the *corpus* by the annotation consistency principle. Aiming at the characteristics of entities in the field of rehabilitation medicine, we construct the BERT-Span model to solve the recognition problem of long and nested entities, and verify the validity of the model by comparing with the classical baseline model.

METHODS

Data preparation

Portions of this text were previously published as part of a preprint (*Zhong et al., 2023*). Stroke is one of the most important public health issues of global concern and is also a hot topic in the field of rehabilitation medicine (*Zhou et al., 2019*). Our research will also focus on stroke in rehabilitation. Experimental data for this article are mainly from stroke rehabilitation guidelines and medical literature. Considering the richness and authority of the data, we select CNKI (https://www.cnki.net/) as our retrieval database. Set keywords "stroke", "rehabilitation" and corresponding synonyms in CNKI advanced search, and select "CSSCI", "core journals" and "CSCD" literature from 2011–2020. After manual screening, some irrelevant and repetitive literatures were removed, and 160 literatures were derived as partial data sources. Finally, a total of 10,000 corpora are collected from rehabilitation guides and CNKI literature, totaling 612,807 words.

Since there is no unified standard for naming entity identification in the field of rehabilitation medicine, on the basis of our previous experience in entity labeling in the field of medicine and the characteristics of texts in the field of rehabilitation, we divide rehabilitation entities into: "dysfunction and performance", "rehabilitation assessment",



```
{"text": "Early identification of patients at risk for pressure ulcers and
collaboration between caregivers to prevent pressure sores is essential.
Patients at risk for pressure ulcers may have the following conditions: (1)
impaired voluntary mobility; (2) diabetes mellitus; (3) peripheral vascular
disease; (4) urinary incontinence; (5) high or low body mass index; (6)
sensory impairment; and (7) concurrent other malignant diseases." ,
"label": {"Dysfunction and performance": {"Pressure sores": [[4, 5], [23,
24], [30, 31]], "Impaired voluntary mobility": [[46, 53]], "diabetes
mellitus": [[56, 58]], "peripheral vascular disease": [[61, 66]], "urinary
incontinence": [[69, 72]], "high or low body mass index ": [[75, 83]],
"sensory impairment": [[86, 89]]}}}
```

Figure 1 **A *corpus* diagram in a JSON file.** Full-size DOI: 10.7717/peerj-cs.1535/fig-1

"rehabilitation methods", "rehabilitation equipment", "body" and "drugs" six categories. Among them, the first four categories are summarized from a large number of rehabilitation literature, and the last two categories are obtained by combining the common points of rehabilitation medicine and clinical medicine.

Corpus construction

Referring to the existing medical *corpus*, we have built a *corpus* in the field of rehabilitation medicine. In the context of *corpus* annotation, the article follows the following specifications for the annotation of four types of entities in rehabilitation medicine: non-overlapping annotation, full foreign language vocabulary not marked, and special symbols and conjunctions can be annotated.

In order to ensure the accuracy of annotation, we primarily relies on manual annotation and verification when creating the rehabilitation *corpus*. We conduct three rounds of manual annotation and verification, and extract 500, 1,250, and 1,250 corpora respectively, until there are no entities that are difficult to regulate boundaries or ambiguous categories in the annotation process. For the 2,500 corpora in the last two rounds, we randomly select 500 corpora and submit them to the second person to label independently. The results of the two labels on the 500 corpora are compared and evaluated according to the Inter annotator Agreement (IAA) principle (*Roberts et al., 2009*). After calculation, the F1 value of these 500 rehabilitation corpora reaches 85%, so it is believed that the rehabilitation *corpus* independently constructed in this article is reliable.

Based on the above standard principles and workflow of annotation, we complete annotation work for 2,500 corpora. Find out the entities that appear in each sentence, and mark the category and location of each entity; Correct the wrong annotation form and overlapping annotation, and realize the error detection and omission of manually annotated entities; The manual annotation results are then evaluated. The format of a certain *corpus* in the final JSON file is shown in Fig. 1:

To sum up, we collects 10,000 corpora in the field of rehabilitation medicine, and marks 2,500 of them manually. The number of entities in six categories in the 2,500 corpora is shown in Table 1:

Table 1 **Number of entities in *corpus*.**

Entity type	Dysfunction and performance	Rehabilitation assessment	Rehabilitation methods	Rehabilitation equipment	Body	Drugs
Number of entity	3,881	2,751	3,069	445	2,312	130

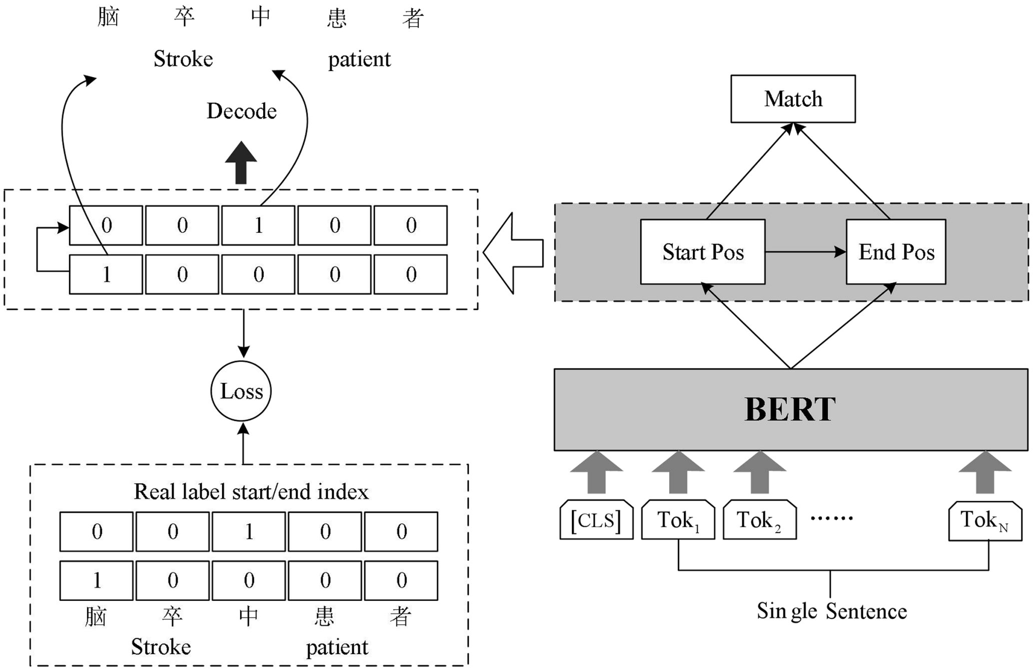


Figure 2 **Schematic diagram of BERT-Span model.** Full-size DOI: 10.7717/peerj-cs.1535/fig-2

BERT-Span model

Chinese rehabilitation medicine field vocabulary is rare, word structure is complex, long noun phrases and nested noun phrases are often found in the context. In order to identify these entity phrases more accurately and improve the accuracy of named entity recognition, we build BERT-Span model for named entity recognition task in rehabilitation medicine. The model structure is shown in Fig. 2.

The sentences in the text are vectorized and fed into BERT, which leverages the masked language model (MLM) pre-training technique. By employing multiple deep bi-directional Transformers, BERT generates deep bi-directional language representation vectors that capture contextual information. These vectors serve as text features and enable BERT to extract meaningful representations from the input sentences.

There are a greater number of long medical entities and nested medical entities in rehabilitation medicine texts. Sequence annotation can only be assigned to the BIO tag corresponding to each token, and becomes insufficient when appearing nested entities. Some researchers propose an idea referring to machine reading comprehension (MRC) model framework technology (*Li et al., 2019b*), specifically dealing with nested entity recognition. We follow this inspiration and use Span model to implement the annotation task.

The Span model uses double sequence tagging in the decoding process. One sequence is used to mark the start position of all entities, and the other sequence is used to mark the end position of all entities. Two linear layers are used for decoding operations. Decoding operation using two linear layers to extract the head and tail part of the entity.

Since there may be multiple entities belonging to the same category in the input text, the model will get several start indexes and end indexes simultaneously during decoding. Hence the model needs to match the decoded start index and end index appropriately. We complete the matching of the start and end index of the entities using the binary classifier. The classifier inputs the possible start and end index of the entity in the sentence and outputs 0 or 1 to indicate whether it matches or not. The formula in Span is as follows:

$$
\begin{aligned}
&P_{start} = soft\max eachrow(E \cdot T_{start}) \in R^{n\times 2} \\
&P_{end} = soft\max eachrow(E + one - hot(P_{start}) \cdot T_{end}) \in R^{n\times 2} \\
&\hat{\mathrm{I}}_{start} = \left\{ i \middle| \arg\max(P_{start}^{(i)}) = 1, i = 1, 2 \cdots, n \right\} \\
&\hat{\mathrm{I}}_{end} = \left\{ j \middle| \arg\max(P_{end}^{(j)}) = 1, j = 1, 2 \cdots, n \right\} \\
&P_{i_{start}, j_{end}} = sigmoid\left(m \cdot concat\left(E_{i_{start}}, E_{j_{end}} \right) \right) \qquad (1)
\end{aligned}
$$

Obtain the representation matrix E from the output of BERT, the model first predicts the probability of each token as the starting index $\mathrm{p}_{\mathrm{start}}$, where $\mathrm{T}_{\mathrm{start}}$ is the weight of learning. Each row of $\mathrm{p}_{\mathrm{start}}$ represents the probability distribution of each index as the starting position of the entity for a given query. The $\mathrm{p}_{\mathrm{start}}$ unique hot code introduced in the $\mathrm{P}_{\mathrm{end}}$ calculation process adds the starting position information to the ending sequence. Apply argmax function to each row of and to get the prediction index that may be the start or end position, $\hat{\mathrm{I}}_{\mathrm{start}}$ and $\hat{\mathrm{I}}_{\mathrm{end}}$, where the $\mathrm{P}^{(\mathrm{i})}$ superscript i represents the ith row of the matrix. Given random starting index $\mathrm{i}_{\mathrm{start}} \in \hat{\mathrm{I}}_{\mathrm{start}}$ and ending index $\mathrm{j}_{\mathrm{end}} \in \hat{\mathrm{I}}_{\mathrm{end}}$, the matching probability of the two indexes is predicted by sigmoid function binary classification model, where m is the weight obtained from learning.

The loss function L consists of three parts: the entity start position loss function $\mathrm{L}_{\mathrm{start}}$, the entity end position loss function $\mathrm{L}_{\mathrm{end}}$, and the entity interval loss function $\mathrm{L}_{\mathrm{span}}$. $\mathrm{L}_{\mathrm{start}}$ denotes the deviation of the predicted entity start position from the true entity start position. $\mathrm{L}_{\mathrm{end}}$ denotes the deviation of the predicted entity end position from the true entity start position. $\mathrm{L}_{\mathrm{span}}$ denotes the deviation of the predicted entity location interval from the true entity location interval. The loss function for each part uses a cross-entropy loss function, where P denotes the model prediction probability and Y is the sample label. The weighted sum of these three loss components is the final overall loss value, where α, β, and γ are the model hyperparameters. The specific formula is as follows:

$$
\begin{aligned}
&L_{start} = CE(P_{start}, Y_{start}) \\
&L_{end} = CE(P_{end}, Y_{end}) \\
&L_{span} = CE\left(P_{start,\, end}, Y_{start,\, end}\right) \\
&L = \alpha L_{start} + \beta L_{end} + \gamma L_{span} \qquad (2)
\end{aligned}
$$

RESULTS AND DISCUSSION

In this section, we compare our propose model with several baseline models and analyze the superiority of our approach from different perspectives. For our model, we try to explore the properties of entities in rehabilitation medicine texts from the perspective to further explore the performance of the model for text processing in the Chinese rehabilitation medicine domain. BERT-softmax and BERT-BiLSTM-CRF are the most widely used baseline models in the field of named entity recognition at the present time, and are also the comparison models chosen for this experiment. During the experiment, we use Adam optimizer with initial learning rate of 2×10^{-5} and warmup_proportion of 0.15. Train batch size is 8 and the train epoch is 20, and the ratio of training set and test set is set to 9:1. The rest of the hyperparameters are default values. Our code and data are released at https://github.com/jamesXuan/BERT-Span.

We use the most commonly used evaluation metrics in the field of named entity recognition: precision rate *P*, recall rate R and F1 value, where, precision rate *P* refers to the probability of the actual correct samples among all the samples predicted to be correct; recall rate R refers to the probability of the predicted samples being correct among the actual correct samples; F1 value is the summed average of recall rate and precision rate, and the calculation formula is given in:

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$
$$F1 = \frac{2PR}{P + R} \tag{3}$$

A comparative experiment of different models in the same *corpus*

Figure 3 illustrates the performance of the BERT-Span compared to the BERT-softmax, BERT-BiLSTM-CRF models on a self-built rehabilitation *corpus* for the rehabilitation named entity recognition task. To mitigate the potential impact of a small dataset leading to significant experimental errors, a 5-fold cross-validation is employed.

It can be seen that with the auxiliary of the pre-trained language model BERT, these models generally have better results for the recognition of named entities in rehabilitation medicine. On this basis, BiLSTM-CRF is advantageous for entity recognition compared to softmax, where the former focuses more on the global information of the sequence. BERT-BiLSTM-CRF outperforms BERT-softmax in terms of recall (R) and F1-score metrics. In addition, there are long medical entities and nested medical entities in the field of rehabilitation medicine, and CRF is difficult to distinguish them because of the limitation of its own structure and cannot play a good effect, while Span can solve this problem well, so the BERT-Span model achieves the optimal effect for the recognition of named entities in the field of rehabilitation medicine.

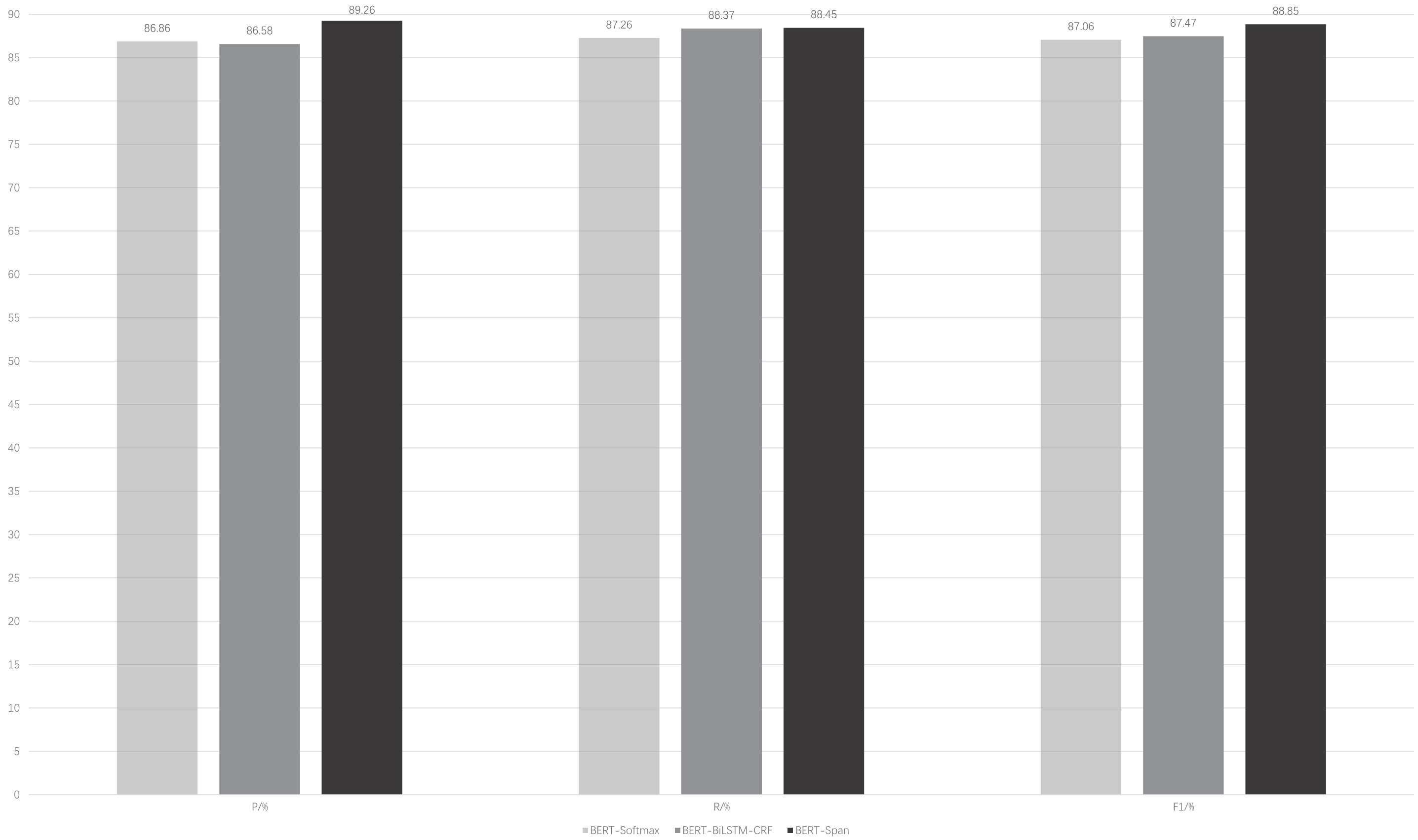


Figure 3 Comparison of model performance for the same data set. BERT-softmax, bidirectional encoder representation from transformers-softmax; BERT-BiLSTM-CRF, bidirectional encoder representation from transformers-bi-directional long short-term memory-conditional random fields; BERT-Span, bidirectional encoder representation from transformers-span. Full-size DOI: 10.7717/peerj-cs.1535/fig-3

A comparative experiment on different corpora

Table 2 illustrates the effect of different *corpus* sizes on model performance. The BERT-Span model was first used to predict the annotation of 7,500 unannotated data, and a total of 2,500 manually annotated data and 7,500 model annotated data were obtained. Among the 2,500 manually labeled data, 1,250 data were randomly selected as a common test set for two sets of experiments. The first set of experiments uses the remaining 1,250 manually labeled data as the training set; the training set of the second set of experiments consists of the first training set and 7,500 model labeled data.

From the table, it is clear that the performance of all three models is significantly improved after increasing the data size of the training set. In terms of F1 metrics, BERT-softmax increases by 1.7%, BERT-BiLSTM-CRF increases by 1.37%, and BERT-Span has the highest increase of 1.8%. The amount of data is still an important factor limiting the performance of the model. Notably, the BERT-Span model consistently achieves optimal performance in the named entity identification task within rehabilitation medicine, both before and after the dataset expansion.

Table 2 Effect of training data size on model performance.

Groups	Model	*P*	R	F1
1	BERT-Softmax	82.40	85.12	83.74
	BERT-BiLSTM-CRF	82.68	85.92	84.27
	BERT-Span	85.58	83.92	84.75
2	BERT-Softmax	84.16	86.77	85.44
	BERT-BiLSTM-CRF	84.42	86.89	85.64
	BERT-Span	86.01	87.11	86.55

Table 3 Entity recognition performance statistics.

Entity	*P*	R	F1
Dysfunction and performance	93.19	92.95	93.07
Rehabilitation assessment	88.80	85.38	87.06
Rehabilitation methods	87.98	87.98	87.98
Rehabilitation equipment	85.29	85.29	85.29
Body	86.78	87.87	87.32
Drugs	100	83.33	90.91

Exploration of the recognition accuracy of different entities

Table 3 presents a comprehensive analysis of the BERT-Span model's performance in identifying the six categories of entities within rehabilitation medicine.

As can be seen from the table, there are considerable differences in the recognition effects of different entities. The identification accuracy of rehabilitation assessment, rehabilitation methods, rehabilitation equipment, and body was low, with F1 values below 90%, where the identification effect of rehabilitation equipment was the worst, with an F1 value of 85.29%. The recognition accuracy of dysfunction and performance, and drug entities was higher, with F1 values over 90%, and "dysfunction and performance" had the best recognition effect, with an F1 value of 93.07%.

The size of the dataset and the distribution of the labels have an impact on the final prediction results of the model. Therefore, we try to explain the variability of identification results between different classes of entities in terms of these two aspects.

We collated the number of each entity in the dataset, the number of low-frequency entities (the number of entity occurrences is 1), and the percentage of low-frequency entities in the total number of entities, as shown in Table 4.

We can see that "dysfunction and performance" has the lowest percentage of low frequencies. The model can extract the feature information of these entities more accurately and give correct prediction results. The highest percentage of low frequencies for "rehabilitation equipment" indicates that most of the data of "rehabilitation equipment" entities only appear once in the data set, and the model cannot obtain their feature information well enough to give high prediction accuracy. Meanwhile, the BERT

Table 4 **Entity number and frequency statistics.**

Entity type	No. of low-frequency entities	No. of entities	Proportion
Dysfunction and performance	666	3,881	17%
Rehabilitation assessment	551	2,751	20%
Rehabilitation methods	866	3,069	28%
Rehabilitation equipment	144	445	32%
Body	472	2,312	20%
Drugs	42	130	32%

pre-training model we use is based on the general clinical medicine *corpus*, while the two types of entities "drug" and "body" exist in clinical medicine, and they can obtain some feature information in advance. The four entities of "dysfunction and performance", "rehabilitation assessment", "rehabilitation methods", and "rehabilitation equipment" exist only in the field of rehabilitation medicine and can only learn feature information during model training. Therefore, although the low frequency of "drug" is as high as 32% as the low frequency of "rehabilitation equipment", the drug has learned some of its characteristic information in advance, so it still has a high recognition effect. It can be expected that dealing with the low-frequency word problem will have further improvement on the performance of the model.

CONCLUSIONS

Named entity recognition in rehabilitation medicine is an important step in many rehabilitation text information extraction tasks. In order to advance the research of named entity recognition in rehabilitation field, we collect data from various ways, discuss the scheme with rehabilitation physicians, and build a rehabilitation *corpus*, which solves the problem that there is no ready-made *corpus* available in the field of rehabilitation medicine. To avoid the influence of long medical entities and nested entities on Chinese named entity recognition in rehabilitation medicine, we proposed the Bert-Span model, using span tokens instead of traditional CRF sequence tokens. The experimental results demonstrate that the BERT-Span model is more suitable for named entity recognition studies in rehabilitation medicine compared with the BERT-Softmax and BERT-BiLSTM-CRF baseline models.

However, there are some limitations of our study. On the one hand, we have only validated the performance of the Chinese rehabilitation medicine named entity recognition task on a self-built *corpus*, and its effectiveness on Chinese NERs from other rehabilitation datasets needs further validation. On the other hand, the performance of the model on named entity recognition tasks in other language rehabilitation domains is unclear.

In the future, we would apply our model to the actual rehabilitation medicine text annotating tasks, so as to help build the knowledge graph of rehabilitation medicine. Moreover, we plan to extend our approach to other relevant rehabilitation medicine text datasets to further validate and improve the model's scalability and generalization capabilities.

ACKNOWLEDGEMENTS

The authors would like to thank all members of the research team for their technical support during the research activities.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Jinhong Zhong conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Zhanxiang Xuan conceived and designed the experiments, performed the experiments, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Kang Wang analyzed the data, prepared figures and/or tables, and approved the final draft.
- Zhou Cheng conceived and designed the experiments, prepared figures and/or tables, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Code and raw data are available at GitHub and Zenodo:

https://github.com/jamesXuan/BERT-Span.

Zhanxiang Xuan. (2023). BERT-Span dataset [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8088554.

The experimental data in this article are from the rehabilitation guide of stroke and medical literature of CNKI platform (https://www.cnki.net/).

REFERENCES

An Y, Xia X, Chen X, Wu FX, Wang J. 2022. Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF. *Artificial Intelligence in Medicine* **127(2)**:102282 DOI 10.1016/j.artmed.2022.102282.

Bhatia P, Arumae K, Busra Celikkaya E. 2020. Dynamic transfer learning for named entity recognition. *Precision Health and Medicine: A Digital Revolution in Healthcare* **843**:69–81 DOI 10.1007/978-3-030-24409-5_7.

Committee NHaW. 2012–2020. *China health and wellness statistical yearbook*. Beijing: Peking Union Medical University Press.

Committee NHaW. 2021. Opinions on accelerating the development of rehabilitation medical work. *Available at http://www.nhc.gov.cn/yzygj/s7653pd/202106/a8aa207082ce4efd998929f3ad6170fe.shtml*.

Devlin J, Chang M-W, Lee K, Kristina T. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *ArXiv preprint* DOI 10.48550/arXiv.1810.04805.

Dong X, Chowdhury S, Qian L, Guan Y, Yang J, Yu Q. 2017. Transfer bi-directional LSTM RNN for named entity recognition in Chinese electronic medical records. In: *2017 IEEE 19Th International Conference on E-Health Networking, Applications and Services (Healthcom)*. Piscataway: IEEE, 1–4.

Dong X, Chowdhury S, Qian L, Li X, Guan Y, Yang J, Yu Q. 2019. Deep learning for named entity recognition on Chinese electronic medical records: combining deep transfer learning with multitask bi-directional LSTM RNN. *PLOS ONE* **14(5)**:e0216046 DOI 10.1371/journal.pone.0216046.

Dong X, Qian L, Guan Y, Huang L, Yu Q, Yang J. 2016. A multiclass classification method based on deep learning for named entity recognition in electronic medical records. In: *2016 New York Scientific Data Summit (NYSDS)*. Piscataway: IEEE, 1–10.

Feng Z, Glinskaya E, Chen H, Gong S, Yip WJTL. 2020. Long-term care system for older adults in China: policy landscape, challenges, and future prospects. *The Lancet* **396**:1362–1372 DOI 10.1016/S0140-6736(20)32136-X.

Florez E, Precioso F, Riveill M, Pighetti R. 2018. Named entity recognition using neural networks for clinical notes. In: *International Workshop on Medication and Adverse Drug Event Detection: PMLR*. 7–15.

Fukuda K-I, Tsunoda T, Tamura A, Takagi T. 1998. Toward information extraction: identifying protein names from biological papers. *Pacific Symposium on Biocomputing* **707(18)**:707–718.

Gong L, Zhang Z, Chen S. 2020. Clinical named entity recognition from Chinese electronic medical records based on deep learning pretraining. *Journal of Healthcare Engineering* **2020(2)**:8829219 DOI 10.1155/2020/8829219.

Ji B, Liu R, Li S, Yu J, Wu Q, Tan Y, Wu J. 2019. A hybrid approach for named entity recognition in Chinese electronic medical record. *BMC Medical Informatics and Decision Making* **19(S2)**:64 DOI 10.1186/s12911-019-0767-2.

Ji S, Pan S, Cambria E, Marttinen P, Yu PS. 2021. A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems* **33**:494–514 DOI 10.1109/TNNLS.2021.3070843.

Keretna S, Lim CP, Creighton D. 2014. A hybrid model for named entity recognition using unstructured medical text. In: *2014 9th International Conference on System of Systems Engineering (SOSE)*. Piscataway: IEEE, 85–90.

Keretna S, Lim CP, Creighton D, Shaban KB. 2014. Classification ensemble to improve medical named entity recognition. In: *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Piscataway: IEEE, 2630–2636.

Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. 2014. A comprehensive study of named entity recognition in Chinese clinical text. *Journal of the American Medical Informatics Association* **21**:808–814 DOI 10.1136/amiajnl-2013-002381.

Li Z, Liu F, Yang W, Peng S, Zhou J. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems* **33**:6999–7019 DOI 10.1109/TNNLS.2021.3084827.

Li J, Sun A, Han J, Li C. 2020a. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data* **34**:50–70 DOI 10.1109/TKDE.2020.2981314.

Li Y, Wang X, Hui L, Zou L, Li H, Xu L, Liu W. 2020b. Chinese clinical named entity recognition in electronic medical records: development of a lattice long short-term memory model with contextualized character representations. *JMIR Medical Informatics* **8(9)**:e19848 DOI 10.2196/19848.

Li X, Zhang H, Zhou X-H. 2020. Chinese clinical named entity recognition with variant neural structures based on BERT methods. *Journal of Biomedical Informatics* **107(5)**:103422 DOI 10.1016/j.jbi.2020.103422.

Li L, Zhao J, Hou L, Zhai Y, Shi J, Cui F. 2019a. An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records. *BMC Medical Informatics and Decision Making* **19(S5)**:395 DOI 10.1186/s12911-019-0933-6.

Li X, Feng J, Meng Y, Han Q, Wu F, Li J. 2019b. A unified MRC framework for named entity recognition. *ArXiv preprint* DOI 10.48550/arXiv.1910.11476.

Liu N, Hu Q, Xu H, Xu X, Chen M. 2022. Med-BERT: a pretraining framework for medical records named entity recognition. *IEEE Transactions on Industrial Informatics* **18(8)**:5600–5608 DOI 10.1109/TII.2021.3131180.

Pan SJ, Yang Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* **22**:1345–1359 DOI 10.1109/TKDE.2009.191.

Qin Y, Zeng Y. 2018. Research of clinical named entity recognition based on Bi-LSTM-CRF. *Journal of Shanghai Jiaotong University (Science)* **23(3)**:392–397 DOI 10.1007/s12204-018-1954-5.

Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, Setzer A. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics* **42(5)**:950–966 DOI 10.1016/j.jbi.2008.12.013.

Schuster M, Paliwal KK. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* **45**:2673–2681 DOI 10.1109/78.650093.

Wan Q, Liu J, Wei L, Ji B. 2020. A self-attention based neural architecture for Chinese medical named entity recognition. *Mathematical Biosciences and Engineering* **17(4)**:3498–3511 DOI 10.3934/mbe.2020197.

Wu C, Luo G, Guo C, Ren Y, Zheng A, Yang C. 2020. An attention-based multi-task model for named entity recognition and intent analysis of Chinese online medical questions. *Journal of Biomedical Informatics* **108(17)**:103511 DOI 10.1016/j.jbi.2020.103511.

Wu G, Tang G, Wang Z, Zhang Z, Wang Z. 2019. An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition. *IEEE Access* **7**:113942–113949 DOI 10.1109/ACCESS.2019.2935223.

Xu K, Zhou Z, Gong T, Hao T, Liu W. 2018a. SBLC: a hybrid model for disease named entity recognition based on semantic bidirectional LSTMs and conditional random fields. *BMC Medical Informatics and Decision Making* **18(S5)**:114 DOI 10.1186/s12911-018-0690-y.

Xu K, Zhou Z, Hao T, Liu W. 2018b. A bidirectional LSTM and conditional random fields approach to medical named entity recognition. In: *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics*. vol. 2017, 355–365.

Xue K, Zhou Y, Ma Z, Ruan T, Zhang H, He P. 2019. Fine-tuning BERT for joint entity and relation extraction in Chinese medical text. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Piscataway: IEEE, 892–897.

Yin M, Mou C, Xiong K, Ren J. 2019. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism. *Journal of Biomedical Informatics* **98(1)**:103289 DOI 10.1016/j.jbi.2019.103289.

Zhang W, Jiang S, Zhao S, Hou K, Liu Y, Zhang L. 2019. A BERT-BiLSTM-CRF model for Chinese electronic medical records named entity recognition. In: *2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA)*. 166–169.

Zhang S, Zheng D, Hu X, Yang M. 2015. Bidirectional long short-term memory networks for relation classification. In: *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*. 73–78.

Zhao S, Cai Z, Chen H, Wang Y, Liu F, Liu A. 2019. Adversarial training based lattice LSTM for Chinese clinical named entity recognition. *Journal of Biomedical Informatics* **99(14)**:103290 DOI 10.1016/j.jbi.2019.103290.

Zhong J, Xuan Z, Wang K, Cheng Z. 2023. A BERT-span model for named entity recognition in rehabilitation medicine. *Research Square*. DOI 10.21203/rs.3.rs-2588490/v1.

Zhou M, Wang H, Zeng X, Yin P, Zhu J, Chen W, Li X, Wang L, Wang L, Liu YJTL. 2019. Mortality, morbidity, and risk factors in China and its provinces, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. **394**:1145–1158 DOI 10.1016/S0140-6736(19)30427-1.