PeerJ Computer Science

A named entity recognition model embedding rehabilitation medicine Chinese character knowledge graph

Jinhong Zhong[1], Zhou Cheng[1], Shengping Liu[2], Mengting Sun[1], Zhanxiang Xuan[1] and Wenxing Lu[1]

[1] Hefei University of Technology, Hefei, China
[2] Unisound AI Co., Ltd., Beijing, China

ABSTRACT

Background: Named entity recognition (NER) is pivotal for medical information extraction and clinical decision support. However, most studies on Chinese medicine NER account for single-feature and multi-feature embeddings of Chinese characters, neglecting the multi-feature correlations inherent in Chinese characters. This limitation is exacerbated in rehabilitation medicine due to sparse annotated data, complex terminology, and significant character-level polysemy, which traditional models struggle to address effectively.

Methods: To bridge this gap, this article constructs a novel NER framework integrating a rehabilitation medicine chinese character knowledge graph (RMCCKG). The RMCCKG not only encompasses Chinese character features, including character, radical, pinyin, stroke, structure, part of speech, and morphology, but also embraces interrelationships between these features. Expanding on this foundation, we propose the RMCCKG+BERT-BiLSTM-CRF NER model. We employ RMCCKG embeddings and Bidirectional Encoder Representations from Transformers (BERT) embeddings as joint input representations, where the RMCCKG embeddings serve as prior knowledge to assist the model in better understanding the text content. This model enables character-level semantic enhancement This model enables character-level semantic enhancement through a hybrid architecture combining bidirectional long short-term memory (BiLSTM) and conditional random field (CRF) layers.

Results: Experiments on our self-developed Rehab dataset and the public CMeEE dataset demonstrate that our model outperforms baseline methods, achieving a 3.96% F1-score improvement. Further, our studies further reveal that low-dimensional fusion of RMCCKG embeddings yields optimal performance, with significant gains in low-frequency entity recognition.

Subjects Data Mining and Machine Learning, Text Mining
Keywords Named entity recognition, Rehabilitation medicine, Chinese character multi-feature, Knowledge graph embedding

Submitted 8 August 2024
Accepted 7 August 2025
Published 23 October 2025

Corresponding authors
Zhou Cheng, fromctoz@163.com
Mengting Sun, smt2908923795@163.com

Academic editor
Bilal Alatas

Additional Information and Declarations can be found on page 21

DOI **10.7717/peerj-cs.3176**

INTRODUCTION

Named entity recognition (NER) is a fundamental information extraction task and a key step in constructing knowledge graphs, playing an important role in natural language processing (NLP). NER identifies entities containing key semantic information in target texts, providing structured and normalized information for downstream tasks such as

question answering systems, machine translation, and text analysis (*Liu et al., 2021*). The Chinese language itself has distinct characteristics, and its smallest unit, the Chinese character, contains rich semantic information, which poses both requirements and challenges for Chinese NER.

The linguistic characteristics of Chinese greatly influence the effectiveness of the general NER in the field of Chinese rehabilitation medicine. Unlike English sentences, Chinese sentences do not have obvious space delimiters between words, making word segmentation an important preprocessing step in Chinese NER tasks. Additionally, the polysemy and homophony of Chinese characters, as well as the relational information between character features, increase the difficulty for models to understand Chinese text content. Data sparsity is another major challenge in NER. Current medical NER systems primarily emphasize word-level features while overlooking the inherent semantic relationships within Chinese characters. In rehabilitation medicine, this limitation is exacerbated by three domain-specific challenges: scarce annotated data, rare terminology, and intricate semantic ambiguities arising from character-level polysemy and homophony. General Chinese NER methods rarely address these concerns.

Prior knowledge can help us understand the text background more quickly and effectively. Some scholars incorporate prior knowledge into models to address issues such as small data size, complex semantics, and rare vocabulary in Chinese NER tasks, thereby improving the performance of entity recognition models. *Yin et al. (2019)* utilize the pictographic nature of Chinese characters to mine deep semantic information of characters, using convolutional neural network (CNN) to extract radical feature information of characters and concatenating it with character information as model input. *Xiong et al. (2022)* use a medical vocabulary as the basis for word segmentation and determine the boundary relationships of entity words through existing disease knowledge graphs, supplementing prior knowledge into the model input to improve model performance. *Li et al. (2019b)* integrate word embeddings and position embeddings based on original character embeddings, compensating for the limited character representation information and the lack of contextual information. *Shi et al. (2022)* integrate external feature information such as dictionaries, pinyin, and radicals from multiple perspectives including character morphology, character units, and word units, fully mining information features such as homophones and pictographs in Chinese. As the types of introduced feature information increase, the relational information between features becomes increasingly important, and establishing relationships between features can further enhance the effectiveness of entity recognition.

This article constructs a rehabilitation medicine Chinese character knowledge graph (RMCCKG) (https://github.com/AaabbB-quick/A-Named-Entity-Recognition-model-embedding-rehabilitation-medicine-Chinese-character-knowledge-graph) for rehabilitation medicine to integrate associative information among Chinese character features. Additionally, a rehabilitation medicine Chinese character knowledge graph embedding method is incorporated into the embedding layer. We propose the RMCCKG +BERT-BiLSTM-CRF model, where BERT-BiLSTM-CRF refers to Bidirectional Encoder Representations from Transformers (BERT) combined with bidirectional long short-term

memory (BiLSTM) and conditional random field (CRF). This model introduces prior knowledge into target texts to fully explore semantic information, enrich entity features, and utilize associative information among Chinese character features. This approach enhances the model's performance in Chinese NER tasks. Experimental results demonstrate that our proposed method is more suitable for NER research in the Chinese rehabilitation medicine domain.

The main contributions of this article are as follows:

(i) A rehabilitation medicine Chinese character knowledge graph (RMCCKG) is constructed by us using over 4,000 rehabilitation-specific and over 8,000 generic Chinese characters. This knowledge graph integrates rich semantic relationships among characters, features, and entities, covering seven entity types and eight interrelationships. Unlike existing medical knowledge graphs that focus on word-level entities, RMCCKG enables character-level semantic mining, addressing challenges like data sparsity and rare terms in rehabilitation texts.

(ii) A novel model that combines RMCCKG embeddings with BERT embeddings as joint input representations is proposed. This integration enhances the model's ability to capture domain-specific semantics and contextual dependencies, making it more effective in identifying named entities with limited labeled data. The experimental results demonstrate that our model outperforms existing methods in Chinese NER tasks within the field of rehabilitation medicine, thereby proving its superiority in recognizing complex rehabilitation entities. As far as we are aware, our work is one of the two studies on the recognition of Chinese rehabilitation medicine terminology.

RELATED WORK

As a crucial component of NLP tasks, NER has consistently garnered the interest of researchers since its initial proposal as a subtask at the Sixth Message Understanding Conference (MUC-6) in 1995 (*Grishman & Sundheim, 1996*). Initially, NER primarily relied on rule-based and dictionary-based methods (*Keretna, Lim & Creighton, 2014*), requiring domain experts to write rule statements and compile dictionaries. With the popularity of machine learning, scholars have begun to apply statistical methods to NER tasks, such as hidden Markov models (HMM) (*Baum et al., 1970*), support vector machines (SVM) (*Vapnik, 2000*), and conditional random fields (CRF) (*Lafferty, McCallum & Pereira, 2001*). Supported by corpora, these methods demonstrate favorable outcomes. Nevertheless, these corpora necessitate extensive manual annotation, and the features within them lack portability, thereby complicating generalization. Concurrently, with advancements in computer technology and the proliferation of vast datasets, deep learning progressively emerges as a prevalent research approach in the domain of NER. *Xu et al. (2018)* train a BiLSTM-CRF model on supervised corpora for medical NER and further propose a similar BiLSTM-CRF model specifically for disease entity recognition, achieving an F1-score of 86.20% on the NCBI disease dataset. As research advances, single-structure neural network models reveal certain limitations, including the failure to capture global feature information and a significant loss of dependency information in

lengthy sentences. The concept of weight distribution inherent in the attention mechanism effectively addresses these issues, prompting scholars to integrate the attention mechanism into the BiLSTM-CRF framework. *Li et al. (2020a)* propose a method based on dynamic attention mechanism and dynamic embedding technology, which appends several characters forming words to the character sequence at the embedding layer to reduce the impact of word segmentation, and incorporates the attention mechanism after the BiLSTM layer to effectively capture contextual encoding information. *Li et al. (2019a)* employ the attention mechanism to augment the model's proficiency in identifying crucial contextual words within sentences for entity extraction tasks. Furthermore, they introduce a segmented attention mechanism, building upon bidirectional long short-term memory networks, to enhance the model's efficacy in processing lengthy sentences. *Wu et al. (2020)* utilize the attention mechanism relying on the BiLSTM-CRF model to complete a multi-task learning network including NER and intent analysis, and design a miniature cardiovascular question-answering system to verify the model's effectiveness.

Many tasks in NER still require a large amount of data with detailed annotations. To tackle this challenge, transfer learning emerges as a potent research strategy (*Pan & Yang, 2010*). The fundamental concept of transfer learning involves conducting unsupervised or semi-supervised pre-training on extensive unlabeled text datasets, subsequently transferring the acquired model parameters to particular downstream tasks. BERT achieves outstanding results in the field of NER (*Devlin et al., 2019*). BERT stands out for its superior language representation and feature extraction abilities, meticulously incorporating contextual information in the pre-training phase, thus acquiring more nuanced and precise language representations. *Xue et al. (2019)* incorporate the BERT pre-trained language model into the original BiLSTM network architecture built upon the attention mechanism to enhance feature extraction capabilities, and propose a parameter sharing strategy at the embedding layer to achieve knowledge transfer and efficiency improvement when handling different tasks within the same model framework. *Liu et al. (2022)* propose the Med-BERT+Span-FLAT model to extract medical entities, where Med-BERT optimizes the representation of long medical entities, and Span-FLAT replaces the traditional CRF, achieving better results in nested entity annotation. *Zhang et al. (2019a)* propose the BERT+BiLSTM-CRF model to address the ambiguity of Chinese characters, and experimental results present that the BERT-enhanced BiLSTM-CRF model performs significantly better in Chinese electronic medical record NER tasks compared to traditional models. *Zhang et al. (2019b)* use the BERT+BiLSTM-CRF method to extract concepts and their attributes from cancer clinical documents. They first pre-train BERT on a large-scale unlabeled Chinese clinical text *corpus* to enrich the contextual embedding representation capabilities of the clinical language model, and then use the BERT pre-trained contextual embeddings as input features for the BiLSTM-CRF model, further fine-tuning the model with annotated breast cancer symptom data, achieving an F1-score of 93.53% in clinical data texts. Deep neural network architectures represented by BiLSTM-CRF demonstrate excellent performance in NER tasks in the medical field.

Given the distinctive features of Chinese characters, including polysemy and homophony, selecting an appropriate embedding method has been a perennial challenge

in Chinese named entity recognition. Initially, Chinese NER techniques are categorized into two main types: word embedding and character embedding. Utilizing characters, the smallest linguistic unit in Chinese, helps circumvent issues related to word boundaries in NER, whereas word embeddings capture and retain the contextual nuances inherent in characters. Word features and character features provide feature information for NER from different perspectives. Some researchers combine the two to obtain richer features, compensating for the shortcomings of single features. *Qin & Zeng (2018)* use character-based word embedding and continuous bag of words as inputs to the BiLSTM-CRF model to capture more representations. *Flórez et al. (2018)* combine character-level features and part-of-speech features to form a comprehensive feature vector and use the BiLSTM-CRF model to identify medical entities from clinical records. *Li et al. (2019b)* propose a BiLSTM-Att-CRF model that integrates an attention mechanism into the neural network architecture for NER tasks. This model combines character-level features, part-of-speech tagging information, and entity information from external knowledge bases at the input layer. Through this multi-angle feature fusion strategy, the BiLSTM-Att-CRF model more comprehensively captures semantic and syntactic information in the text, thereby improving the accuracy of entity boundary recognition. Contrary to them, *Zhao et al. (2019)* use a lattice structure to associate character information and word information, proposing an attention-based convolutional neural network (AT-lattice LSTM-CRF) to improve the shortcomings of single feature representation. *Li et al. (2020b)* also propose the ELMo-lattice-LSTM-CRF model based on the lattice structure, adding the pre-trained ELMo model as input to the text to learn specific contextual information. As exploration advanced, scholars began to focus on the features of Chinese characters themselves. The meanings of Chinese characters are mostly related to radicals, and radical features help models to understand character meanings more easily. *Yin et al. (2019)* use CNN to extract radical-level features on the basis of the BiLSTM model, capturing the internal relationships of characters. *Tang et al. (2019)* leverage characters and radicals as model inputs, engaging CNN to extract local features, LSTM to extract global features, and the attention mechanism to focus on discontinuous word features, achieving better results in the CCKS2017, CCKS2019, and CCKS2020 medical NER tasks. *Lee & Lu (2021)* also start from radicals, characters, and word features, constructing input features and adopting multi-dimensional information embedding to comprehensively and multi-angle mine the semantic features of Chinese, applying graph structures to handle the contextual relationships between characters, words, and sentences. Other Chinese characters features are also utilized to assist models in understanding characters. *Li et al. (2022)* use Glyce to extract glyph features, combining glyph, character, and word features to obtain more information features. *Li, Zhang & Zhou (2020)* add radical and dictionary feature information on the basis of characters and words to assist the model in extracting entity feature information. The position of words is also an important feature. *Zhang et al. (2022)* propose the self-attention PAG network model, which combines the self-attention mechanism and position-aware propagation embedding to solve complications such as unclear word boundaries and long-distance dependency relationships in NER. *Sui et al. (2019)* use C-graph, T-graph, and L-graph to establish the

positional relationships of words in the text, determining word boundary information. Some scholars consider implementing prior knowledge to enhance the performance of model feature extraction. *Xiong et al. (2022)* engage medical domain vocabularies and disease knowledge graphs to supplement knowledge for word segmentation operations, integrating prior knowledge into the model input to improve model performance. *Shi et al. (2022)* fully mine information features such as homophones and pictographs from multiple perspectives, including character glyphs, phonetics, semantics, and word meanings, enhancing the model's ability to learn these features.

In summary, the current mainstream methods of NER are basically based on deep learning and transfer learning. The study of Chinese NER is characterized by the embedding of Chinese character features, but the existing literature mainly focuses on the embedding of a single feature and the independent embedding of multiple features of Chinese characters. As for medical NER research, most of the literature treats a word or phrase as the smallest unit of an entity. Obviously, these studies do not make full use of the information inherent in Chinese characters, such as the relationship between multiple features, the correlation information between Chinese characters due to features, and so on.

In contrast to existing studies, our work uniquely focuses on character-level semantic mining in Chinese rehabilitation medicine. We have constructed the RMCCKG, which operates at the character level to capture intrinsic character attributes and their relational semantics. Based on this foundation, we propose the RMCCKG+BERT-BiLSTM-CRF model. This model integrates knowledge graph embeddings with BERT contextual embeddings to address the specific challenges of Chinese NER in rehabilitation medicine. The RMCCKG embeddings, generated by TransE, are concatenated with BERT contextual embeddings at the input layer, rather than injecting external knowledge at later stages. BERT is used to generate contextual embeddings for the input text, which helps the model understand the context and semantics of the characters in the rehabilitation medicine domain. This integrated approach distinguishes our method from both pure feature concatenation and standalone knowledge graph applications. This paradigm shift not only addresses the challenges of Chinese medical NER but also establishes an extensible framework for integrating domain-specific knowledge graphs into broader NLP applications. As far as we know, this work is the second study of Chinese rehabilitation terminology entity recognition, following (*Zhong et al., 2023*).

METHODS

Data preparation

The datasets utilized in this study include the publicly available CMeEE dataset (*Zan et al., 2021*) and a self-constructed rehabilitation *corpus*, Rehab (*Zhong et al., 2023*). The CMeEE dataset is part of the entity recognition task in the CBLUE evaluation benchmark. It contains 15,000 entries with nine types of medical entities, including 504 types of diseases (dis), 7,085 types of body parts (bod), 12,907 types of clinical manifestations (sym), and 4,354 types of medical procedures, among others. The Rehab dataset is a *corpus* we independently constructed during our preliminary research on NER in rehabilitation

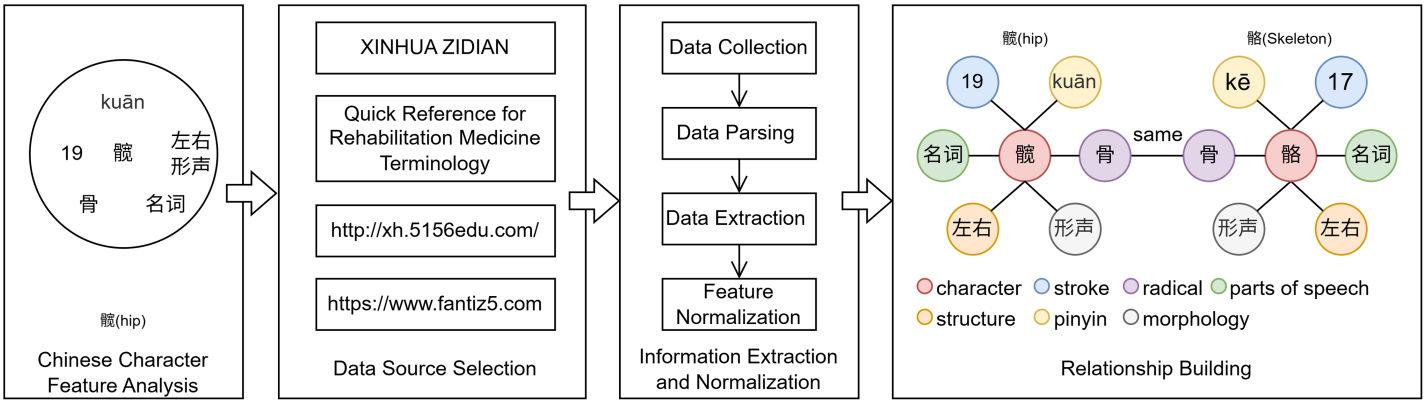

Figure 1 **The construction process of RMCCKG.** Full-size DOI: 10.7717/peerj-cs.3176/fig-1

medicine. The raw data mainly comes from rehabilitation medicine textbooks, rehabilitation medicine guidelines, and text content from CNKI literature, totaling 10,000 entries with over 610,000 characters. It covers six types of rehabilitation medicine entities: "functional disorders and manifestations," "rehabilitation assessment," "rehabilitation methods," "rehabilitation equipment," "body," and "medication," totaling 2,500 rehabilitation medicine records. While this dataset serves as a valuable resource for Chinese rehabilitation NER, two limitations should be noted. Firstly, the current version of the Rehab dataset is primarily tailored to orthopedic rehabilitation practices, as most source texts focus on this subfield. This may limit its direct applicability to other rehabilitation domains without additional adaptation, such as neurological or cardiopulmonary rehabilitation. Despite these limitations, our proposed RMCCKG +BERT-BiLSTM-CRF model is designed to be domain-agnostic. The core methodology can be extended to other rehabilitation subfields by updating the knowledge graph with domain-specific characters and relationships.

The rehabilitation medicine Chinese character knowledge graph

A knowledge graph is a graphical structure employed to represent and store knowledge. It is based on entities and the relationships between entities, forming a semantically connected network that can provide structured and semantic knowledge representation. This assists computers in understanding and processing complex real-world concepts. Existing medical knowledge graphs are mainly constructed using words or phrases as entities, thereby constructing the knowledge graph between entities while neglecting the rich semantic information inherent in Chinese characters (*Yin et al., 2019*).

This article constructs a Chinese rehabilitation knowledge graph to represent the relationships between Chinese character features, laying the groundwork for subsequent research on NER methods. The construction process is shown in Fig. 1.

Chinese characters are the most intuitive representation of Chinese text and the smallest semantic units in Chinese text. The phonetic nature of Chinese characters determines that the pronunciation of Chinese characters is an important factor in understanding their semantics. For example, in "老师" (teacher), "老鼠" (mouse), and "老虎" (tiger), the

character "老" is just a phonetic symbol and does not convey the semantic meaning of "old." The radicals of Chinese characters have the most direct impact on understanding the meaning of the characters. For instance, the radicals of "江" (river), "河" (river), "湖" (lake), and "海" (sea) are all "氵" (water), indicating that they all relate to bodies of water. The structure of Chinese characters reflects the components incorporated in the character, the number of components, the way they are combined, and their placement. Many Chinese characters are structured based on the meaning they convey, so the structure of a character can, to some extent, reflect its meaning. The strokes of Chinese characters highlight the complexity or simplicity of the character. The part of speech of a Chinese character determines its position in a sentence, its meaning, and its corresponding grammar. The morphology of a Chinese character determines its morphological structure and the way its meaning is formed. Based on relevant literature on Chinese character research, this article selects seven features—Chinese character, pronunciation, radical, structure, strokes, part of speech, and morphology—as entities for the Chinese rehabilitation knowledge graph.

"Quick Reference for Rehabilitation Medicine Terminology" (*Guo, Cai & Liang, 2020*) is a book covering professional knowledge in rehabilitation therapy, sports rehabilitation, and other fields. It systematically collects 4,000 core terms and expressions commonly employed in clinical work and study across various rehabilitation departments, including anatomy, physiology, kinesiology, functional assessment, cardiopulmonary, neurology, musculoskeletal, and pediatrics. Considering the universality and scalability of the knowledge graph, as well as improving the efficiency of model learning and understanding, we also include the "General Standard Chinese Characters Table" in our data source selection. The "General Standard Chinese Characters Table" is a character usage standard jointly developed by the Ministry of Education of the People's Republic of China and the National Language Commission. This table collects over 8,000 commonly utilized Chinese characters, meeting the character usage needs for public information exchange in the information age. It is one of the important reference books in the field of rehabilitation medicine. The online Chinese dictionary (http://xh.5156edu.com/) is a website for querying Chinese character information. It allows users to search for characters by character, radical, pinyin, and other methods, providing information on pinyin, stroke, radical, definition, stroke order, and more. The Traditional Chinese Characters website (https://www.zunxu.com/ziti/zaozifa/) offers an online query service for Chinese character morphology, allowing users to input individual characters to query their morphology.

The construction of RMCCKG involved systematic data collection, processing, and validation. First, character data were sourced from the Quick Reference for Rehabilitation Medicine Terminology and the General Standard Chinese Characters Table. For each character, seven features (character, pinyin, radical, structure, stroke, parts of speech, morphology) were programmatically extracted from two online resources: the online Chinese dictionary (http://xh.5156edu.com/) provided pinyin, stroke, and radical details, while morphology information was crawled from the Traditional Chinese Characters website (https://www.zunxu.com/ziti/zaozifa/). Next, we defined seven entity types, including character, pinyin, radical, structure, stroke, parts of speech, and morphology, and established eight relationship types, such as Character-HasRadical and Pinyin-

Table 1 The number of entities and relationships in RMCCKG.

Entity	Number	Relationship	Number
character	8,076	(character, pinyin)	8,956
pinyin	1,329	(character, radical)	8,076
radical	246	(character, stroke)	8,076
struct	8	(character, struct)	8,076
stroke	29	(character, morphology)	8,076
parts of speech	8	(character, part of speech)	8,976
morphology	6	(radical, pinyin)	285
		(radical, stroke)	246

RepresentsCharacter. These relationships were extracted using rule-based scripts that mapped features to their corresponding characters. After data cleaning, the information was converted into triplets using Neo4j's Cypher query language. For example, "膝-HasRadical- 月." Ultimately, we built a knowledge graph containing 9,702 pieces of entity information and 50,767 pieces of triplet information, including 49,867 relationship triplets. Detailed information is presented in Table 1.

To ensure data quality, this article designs three data processing solutions:

(i) Polyphonic characters processing: Pinyin is an important factor in determining the semantics of Chinese characters. For polyphonic characters, we enumerate all possible pronunciations.

(ii) Structural categories selection: There are various methods for classifying the structure of Chinese characters, each differing in their level of granularity. For example, the semi-enclosed structure can be subdivided into left-top enclosing right-bottom, left-bottom enclosing right-top, right-top enclosing left-bottom, left enclosing right, top enclosing bottom, and bottom enclosing top, but this article only considers the semi-enclosed structure, without considering the specific positions within the semi-enclosed structure.

(iii) Information completion methods: Based on the importance of the missing attributes, we use different methods to fill in the missing values, including expert evaluation, book retrieval, or mean substitution, and adding new entity classes.

After cleaning the data according to the data processing solutions, we use NEO4J to generate Chinese rehabilitation triplet information and construct the RMCCKG. The RMCCKG visualization is shown in Fig. 2.

The RMCCKG+BERT-BiLSTM-CRF model

This article proposes the RMCCKG+BERT-BiLSTM-CRF Model, which consists of an input layer, an embedding layer, and a BiLSTM-CRF layer. In terms of model design, the RMCCKG+BERT-BiLSTM-CRF model achieves complementarity between contextual

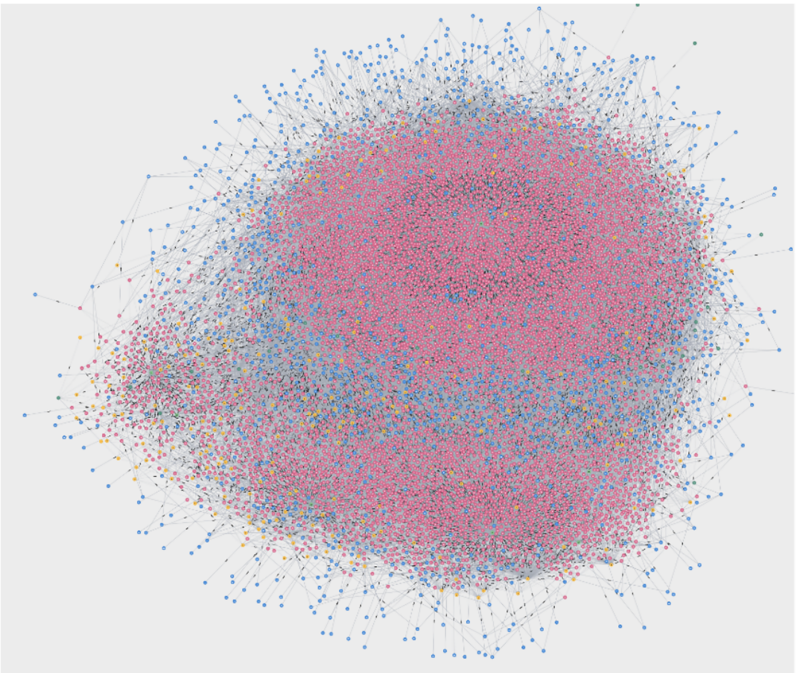

Figure 2 **The graphical illustration of RMCCKG.** Full-size DOI: 10.7717/peerj-cs.3176/fig-2

information and structured knowledge by jointly representing knowledge graph embeddings and BERT embeddings as input. According to feature fusion theory (*Liu et al., 2020*), low-dimensional fusion at the embedding layer enables the earlier integration of prior knowledge into the model, assisting subsequent contextual feature extraction and thereby optimizing the model's overall semantic representation capability. In addition, relationship information in knowledge graphs is encoded *via* graph embedding methods such as TransE (*Bordes et al., 2013*), which can capture the latent associations between character features and provide additional semantic constraints for the model. This knowledge-enhanced mechanism complements BERT's dynamic context modeling (*Zhang et al., 2019*), showing significant advantages, particularly in recognizing low-frequency entities. The input layer reads Chinese rehabilitation texts and performs word segmentation. The embedding layer uses the BERT pre-trained language model and the knowledge graph embedding model to separately receive the output from the input layer. The BERT embedding layer generates a representation sequence containing contextual information of the text, while the knowledge graph embedding layer generates a multi-feature representation sequence. By effectively integrating the representation sequences from both models, we obtain a representation sequence that contains the semantic information of the Chinese rehabilitation text, which serves as the output of the embedding layer. The BiLSTM-CRF layer receives the representation sequence from the embedding layer, extracts features, and finally generates a reasonable label sequence. The model framework is shown in Fig. 3.

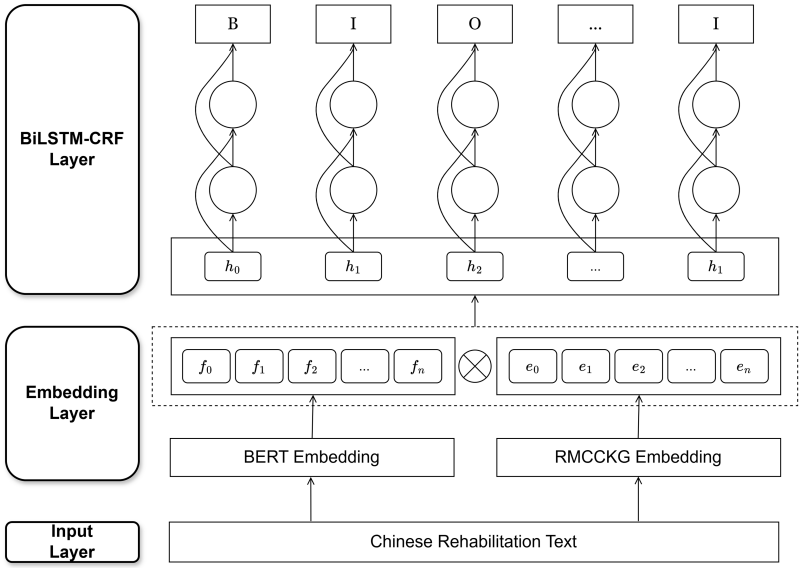

Figure 3 **The framework of the RMCCKG+BERT-BiLSTM-CRF model.**
Full-size DOI: 10.7717/peerj-cs.3176/fig-3

Table 2 **Example of BIO Tagging for Chinese named entity recognition (NER).**

Character	膝	关	节	康	复	训	练
Label	B-Body	I-Body	I-Body	B-RM	I-RM	I-RM	I-RM

Note:
RM, rehabilitation method.

Character sequence input layer

The input layer of the model performs word segmentation on the Chinese rehabilitation text sequence T. Each slice of the Chinese rehabilitation text is further subdivided into a character set $X = \{x_1, x_2, \ldots, x_n\}$, where x_i represents the i-th character in the sentence. Characters include Chinese characters, punctuation marks, numbers, or special symbols. Each character corresponds to a category label, and the BIO tagging method is used to label the character categories. BIO tagging is a labeling method used in NER to distinguish between entities and non-entities in text, where "B" indicates the beginning of an entity, "I" represents the consecutive parts within an entity, and "O" denotes parts that do not belong to any entity. After reading, the labels are stored in the character set $Y = \{y_1, y_2, \ldots, y_n\}$. The input layer finally outputs the character set X and the character category label set Y.

For example, the sentence "膝关节康复训练" (Knee joint rehabilitation training) demonstrates character segmentation and BIO annotation for entity types "身体" (body) and "康复方法" (rehabilitation method). As shown in Table 2, the annotation scheme clearly distinguishes entities at the character level.

The BERT embedding

BERT receives the character set from the character input layer, fine-tunes the character mapping rule dictionary and basic configurations in the configuration file of the BERT

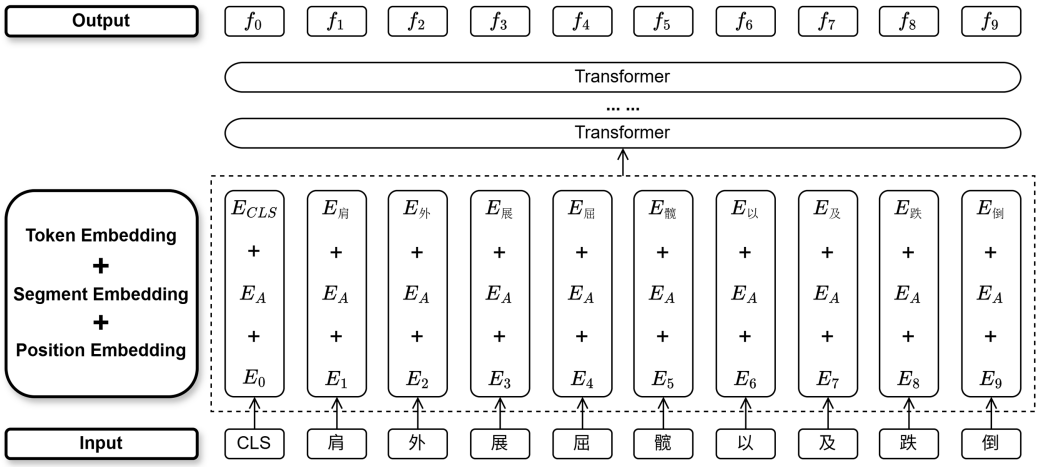

Figure 4 The BERT embedding process. Full-size DOI: 10.7717/peerj-cs.3176/fig-4

pre-trained model, and tenderizes the character set to achieve the conversion from natural language to machine language. For sensitive words in the Chinese rehabilitation text, BERT matches a suitable contextual representation vector for the sensitive words from the candidate vectors based on contextual information. We use the BERT pre-trained language model to learn embedding representations that contain contextual information about the text. The process of BERT embedding is shown in Fig. 4.

The BERT embedding layer receives the character set $X = \{x_1, x_2, \ldots, x_n\}$ output from the character input layer, where x_i represents the i-th character in the sentence, and CLS is the start marker of the sentence. The BERT embedding layer uses position embedding vectors, segment embedding vectors, and token embedding vectors to respectively store the content, segment, and position information of the character x_i. Here, we only use the masked language model task of BERT, so there is no involvement of relationship information between sentence pairs, and the segment is unique. The tensorized vectors are passed into a multi-layer Transformers network to learn contextual information, resulting in a representation sequence $F = \{f_0, f_1, \ldots, f_n\}$, where f_i represents the representation information of the i-th character. The representation sequence F is used as the input to the next layer of the neural network to further mine the semantic features of the text.

The knowledge graph embedding

Knowledge graph embedding is a special type of multi-feature embedding that focuses on handling graph-structured data. Its core idea is to incorporate experiential knowledge before model training to help the model quickly understand the text and identify key semantic features. Compared to multi-feature embedding, knowledge graph embedding retains multiple feature entities while considering the relationship information between features, using external knowledge to increase the model's learning efficiency. This greatly aids in enhancing the model's ability to learn from Chinese rehabilitation texts.

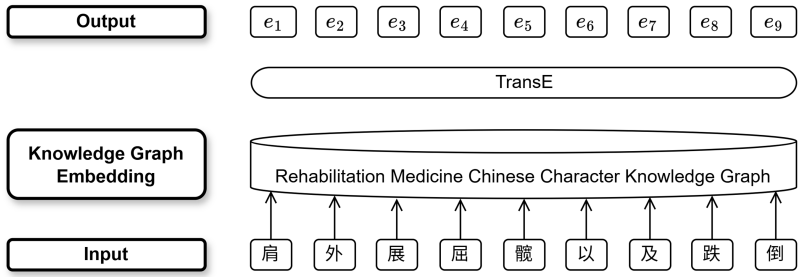

Figure 5 **The RMCCKG embedding process.** Full-size DOI: 10.7717/peerj-cs.3176/fig-5

In the knowledge graph embedding layer, we use 9,702 entity information entries and 50,767 triplet information entries, including seven types of entities (Chinese character, pronunciation, radical, structure, strokes, part of speech, and morphology) and eight types of relationships between character features from the Chinese rehabilitation knowledge graph, as external knowledge to enhance the model's ability to mine the semantic features of Chinese characters. The specific process of the Chinese rehabilitation knowledge graph embedding model is shown in Fig. 5.

The triplets $G(e_{head}, r, e_{tail})$ containing entities such as character, pronunciation, radical, structure, stroke, parts of speech, and morphology are read from the Chinese rehabilitation knowledge graph. For instance, the character "骼" (skeleton) is represented in the knowledge graph by triplets such as 骼-HasRadical-骨 (radical: "骨"), 骼-Pinyin-gé (pinyin: "gé"), and 骼-Structure-左右结构 (structure: left-right composition), which collectively encode semantic relationships between its features through TransE embeddings. Negative sample triplets of different categories are randomly generated to enhance the model's generalization ability by artificially adding noise. The TransE knowledge graph representation learning model is utilized to generate additional embedding vectors that include external knowledge. These vectors contain not only the semantic information of the entities themselves but also the semantic information of the relationships between entities, achieving information enhancement and greatly aiding the model in understanding Chinese rehabilitation texts. The model uses a squared error function with added perturbations as the loss function, calculated as follows:

$$L(e_1, r, e_2) = max\ (0, \|e_1 + r - e_2\|_2^2) - \text{margin}(m)) \tag{1}$$

where e_1 is the vector representation of the head entity, r is the vector representation of the relationship, e_2 is the vector representation of the tail entity, $\|e_1 + r - e_2\|_2^2$ represents the L2 norm used to calculate the Euclidean distance between vectors, and "margin" (m) is the boundary value of the perturbation used to ensure the model can distinguish between positive and negative samples.

Finally, in the embedding layer, we fuse the representation sequence $F = \{f_0, f_1, \ldots, f_n\}$ obtained from the BERT pre-training with the representation sequence $E = \{e_0, e_1, \ldots, e_n\}$ obtained from the Chinese rehabilitation knowledge graph embedding, resulting in a new representation sequence $H = \{h_0, h_1, \ldots, h_n\}$, which is then passed to the next layer.

BiLSTM-CRF layer

The bidirectional long short-term memory network (BiLSTM) is based on the recurrent neural network (RNN) and is a variant of the long short-term memory network (LSTM). Compared to traditional RNNs, BiLSTM is more effective in capturing contextual information. We use BiLSTM to capture the contextual representations of the text. The BiLSTM layer receives the output sequence $H = \{ h_0, h_1, \ldots, h_n\}$ from the embedding layer. The forward LSTM layer captures the preceding context features of the current token from left to right, while the backward LSTM layer captures the succeeding context features from right to left. The contextual feature sequences are then merged to form a unified feature sequence output.

The CRF layer serves as the output layer of the entire model and is employed for sequence labeling tasks. For character-based Chinese NER tasks, considering the dependency between adjacent labels is effective. The CRF layer can model the dependencies between labels and calculate the transition feature probabilities of the feature sequence from the BiLSTM layer. This helps to find the optimal label sequence combination, ensuring that the output label sequence combination is reasonable and consistent throughout the text. This improves the accuracy of NER and enhances the model's ability to handle the complex structure of Chinese text.

Inside the CRF layer, a transition score matrix A_{ij} is randomly initialized, representing the transition probabilities from label i to label j. Given the hidden layer output H, the comprehensive score calculated for the sequence is:

$$S(H, y) = \sum_{i=1}^{N+2} A_{i_{n-1}, i_n, y} + P_{i_n, y}. \tag{2}$$

Here, $A_{i_{n-1}, i_n, y}$ represents the transition value from the label at time step $n-1$ to the label at time step n, obtained from the transition matrix A. Applying the Softmax function to the possible output sequence labels yields the probability of the predicted label sequence:

$$P(y|X) = \frac{e^{s(X,y)}}{\sum_{\hat{y} \in Y} e^{s(X,\hat{y})}}. \tag{3}$$

In the formula, the numerator s represents the score of the correct label sequence, while the denominator s represents the scores of all possible label sequences. The loss function is derived by taking the negative logarithm of this value.

$$-\log P(y|X) = -\left(s(X, y) - \log\left(\sum_{\hat{y} \varepsilon Y} e^{s(X,\hat{y})} \right) \right). \tag{4}$$

Finally, the Viterbi algorithm is used for decoding to find the label sequence combination with the highest score:

$$Y^* = \underset{\hat{Y} \in Y}{\operatorname{argmax}} f\left(H, \tilde{Y}\right). \tag{5}$$

RESULTS AND DISCUSSION

In this section, we conduct comparative experiments on baseline models and feature embedding patterns on both the self-developed Rehab dataset and the public dataset CMeEE (https://github.com/CBLUEbenchmark/CBLUE). Further, we analyze the impact of information fusion at different stages and the model performance in recognizing different rehabilitation entity types. We mainly introduce the parameters of the Chinese rehabilitation knowledge graph embedding model. The number of training iterations is set to 10, the batch size is set to 4, and the learning rates for BERT and LSTM are both set to 3e−4. Adam is used as the optimizer, and knowledge graph embedding is incorporated.

We employ the most commonly used evaluation metrics in the field of NER to assess the performance of the model: precision (P), recall (R), and F1-score. True positives (TP) are the number of samples correctly predicted as positive by the model. False positives (FP) are the number of samples incorrectly predicted as positive by the model. True negatives (TN) are the number of samples correctly predicted as negative by the model. The calculation formulas are as follows:

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = \frac{2PR}{P + R}. \tag{8}$$

Comparison with baseline models

To verify the generalization ability and scalability of the model, we select different datasets and baseline models to validate the effectiveness of the Chinese rehabilitation knowledge graph embedding method in NER. For datasets, we choose the general medical *corpus* CMeEE and the self-constructed rehabilitation medical *corpus* Rehab. The former is a Chinese medical text NER dataset jointly developed by the Key Laboratory of Computational Linguistics at Peking University and Peng Cheng Laboratory, with rich entity categories and high authority. We partially sample 500 records without affecting the entity distribution. The latter is a *corpus* we independently constructed earlier, belonging to a branch of the medical field with high domain specificity. Similarly, we sample 500 records without affecting the entity distribution. For baseline models, we select the commonly used BERT-Softmax model for NER tasks. In both models, we apply three embedding methods: no feature embedding, multi-feature embedding, and Chinese rehabilitation knowledge graph embedding, to observe the models' performance in handling NER tasks. The performance of the models on the Rehab dataset is shown in Table 3.

From Table 3, we can see that the knowledge graph embedding method outperforms the other two embedding methods in all evaluation metrics for both the BERT-BiLSTM-CRF model and the BERT-Softmax model. This indicates that knowledge graph embedding can enhance information in different baseline models, and this enhancement effect is better

Table 3 Results of different baseline models on the Rehab dataset.

Model	Embedding type	P%	R%	F1%
BERT-BiLSTM-CRF	No feature	81.75	84.58	83.14
	Multi-feature	84.05	83.90	83.97
	Knowledge graph	84.50	85.77	85.13
BERT-Softmax	No feature	79.29	83.15	81.17
	Multi-feature	80.18	84.08	82.08
	Knowledge graph	81.42	86.14	83.71

than multi-feature embedding. Specifically, the Recall (*R*) value in the BERT-BiLSTM-CRF model is 1.87% higher than that of multi-feature embedding (multi-embedding). In the BERT-Softmax model, the *R* value and F1-score are 2.06 and 1.63% higher than those of multi-feature embedding (multi-embedding), respectively. In the BERT-Softmax model, the F1-score of knowledge graph embedding is even 0.26% lower than the multi-feature embedding (multi-embedding) in the BERT-BiLSTM-CRF model. Overall, the BERT-BiLSTM-CRF model is more suitable for the NER task in Chinese rehabilitation texts.

From the perspective of model architecture, the BERT-BiLSTM-CRF model has an additional BiLSTM network layer compared to the BERT-Softmax model. The BiLSTM network can handle sequential information in the text and effectively capture long-term data dependencies, thereby obtaining more comprehensive feature information and improving model performance. Comparing the enhancement effects brought by knowledge graph embedding in the two models, we can find that knowledge graph embedding has a better amplification effect on simpler models. In the BERT-BiLSTM-CRF model, the F1-score of knowledge graph embedding (RMCCKG-embedding) is 1.99% higher than that of no feature embedding (none-embedding). However, in the BERT-Softmax model, the F1-score of knowledge graph embedding (RMCCKG-embedding) is 2.54% higher than that of no feature embedding (none-embedding). The performance of the models on the CMeEE dataset is shown in Table 4.

From Table 4, we can see that the knowledge graph embedding method is still effective on the CMeEE dataset. The F1-score reaches 70.3% in the BERT-BiLSTM-CRF model and 69.59% in the BERT-Softmax model, both higher than the corresponding models with no feature embedding. For the BERT-BiLSTM-CRF model, the embedding method has a significant impact on the NER task. Both multi-feature concatenation embedding (multi-embedding) and knowledge graph embedding (RMCCKG-embedding) can improve task performance, with the latter showing a more pronounced improvement. In the BERT-Softmax model, although both multi-feature embedding (multi-embedding) and knowledge graph embedding (RMCCKG-embedding) improve the NER task, the difference between the two is not very significant. Clearly, the BERT-BiLSTM-CRF model with knowledge graph embedding (RMCCKG-embedding) has better generalization ability and can handle NER tasks in more scenarios.

Considering the characteristics of the dataset itself, we find that the effect of knowledge graph information enhancement is better for datasets with a rich variety of entities. The

Table 4 Results of different baseline models on the CMeEE dataset.

Model	Embedding type	*P* %	*R* %	F1 %
BERT-BiLSTM-CRF	No feature	65.98	70.18	68.01
	Multi-feature	68.16	69.96	69.05
	Knowledge graph	67.40	73.46	70.30
BERT-Softmax	No feature	64.75	71.71	68.05
	Multi-feature	67.42	72.15	69.70
	Knowledge graph	67.99	71.27	69.59

more types of entities there are, the more complex the semantic information becomes. The limited sample size is insufficient for the model to fully understand this semantic information. Introducing prior knowledge can help the model quickly organize semantic information, thereby improving recognition accuracy.

Comparison of feature embedding patterns

In the self-constructed dataset Rehab, we conduct NER tasks using the same BERT-BiLSTM-CRF structure with different embedding patterns: no feature embedding, single feature embedding, multi-feature embedding, and knowledge graph embedding models. For single feature embedding, we select Chinese character radicals, pinyin, structure, and strokes to analyze the performance of each model. We partially sampled 500 records without affecting the entity distribution and divided them into training and test sets in a 9:1 ratio.

Based on the same BERT-BiLSTM-CRF structure, we use different embedding patterns in the embedding layer for information enhancement: no feature embedding (none-embedding), radical embedding (radical-embedding), pinyin embedding (pinyin-embedding), structure embedding (struct-embedding), stroke embedding (stroke-embedding), multi-feature embedding (multi-embedding), and Chinese rehabilitation knowledge graph embedding (RMCCKG-embedding). We analyze the NER results of the models in the Chinese rehabilitation domain. The experimental results are shown in Table 5.

From Table 5, we can see that the pre-trained language model with no feature embedding (none-embedding) already performs well in the NER task. Single feature information enhancement can improve the performance of the NER model on the original basis, but the extent of improvement varies with different single feature embeddings. Among them, struct-embedding has the most significant impact on the model, with an F1-score reaching 84.03%, while pinyin-embedding has almost no effect on improving model performance. This indicates that different Chinese character features represent different aspects of character information and contribute differently to the extraction of semantic information. The F1-score of single feature embeddings follows the order: character structure embedding > character radical embedding > character stroke embedding > character pinyin embedding. Among the four features (structure, radical,

Table 5 Results of models with different feature embedding.

Embedding type	*P* %	*R* %	F1 %
none-embedding	81.75	84.58	83.14
radical-embedding	83.24	84.54	83.88
pinyin-embedding	80.90	84.08	82.46
struct-embedding	82.34	85.79	84.03
stroke-embedding	81.97	84.08	82.92
multi-embedding	84.05	83.90	83.97
RMCCKG-embedding	84.50	85.77	85.13

pinyin, and stroke), the model is more sensitive to structural information. Structure is the most intuitive representation of Chinese character form, and the form of Chinese characters contains rich semantic information, providing more prior knowledge to the target text. Multi-feature embedding outperforms most single feature embeddings, but there is a feature fusion problem among multiple features. An inappropriate fusion pattern may amplify the noise in the features, weaken the information intended to be expressed by the features, and interfere with the model's learning and understanding of effective information. Our proposed RMCCKG-embedding model performs best in terms of precision (P) and F1-score, only slightly lower than the struct-embedding model in recall (R). Knowledge graph embedding can better represent the semantic information of the features themselves and the relationship information between features, showing the best performance among various embedding patterns.

The impact of information fusion stage

This experiment explores whether adding external knowledge at different stages of the model has different impacts on NER and whether the concatenation method during embedding affects NER. Through this experiment, we aim to find the most suitable timing for information fusion based on the Chinese rehabilitation knowledge graph embedding. The experiment uses the Rehab dataset, and we partially sample 500 records without affecting the entity distribution. The experimental results list the effects of high-dimensional concatenation (knowledge graph embedding vectors concatenated after BERT embedding vectors) and low-dimensional concatenation (knowledge graph embedding vectors concatenated before BERT embedding vectors) at the embedding layer and BiLSTM layer, respectively. The results are shown in Table 6.

From Table 6, we can conclude that performing low-dimensional concatenation at the embedding layer and passing the concatenated result to the BiLSTM-CRF layer is the best method for information fusion, achieving an *R* value of 87.45% and an F1-score of 87.45%. Additionally, high-dimensional concatenation at the embedding layer also significantly improves model performance, with the F1-score being second only to the best method. The effect of low-dimensional concatenation at the BiLSTM layer is much worse than high-dimensional concatenation, and the F1-scores of both concatenation methods at the BiLSTM layer are lower than those at the embedding layer. We speculate that the semantic

Table 6 Performance of the model at different stages of information fusion.

Layer	Concatenation method	*P* %	*R* %	F1 %
Embedding Layer	High-dimensional	84.50	85.77	85.13
	Low-dimensional	83.39	87.45	85.69
BiLSTM Layer	High-dimensional	84.19	85.77	84.97
	Low-dimensional	82.23	85.77	83.96

Table 7 Recognition results of the model for different rehabilitation entity types.

Entity type	*P* %	*R* %	F1 %
Functional disorders and manifestations	90.91	92.90	91.89
Rehabilitation methods	77.12	79.82	78.45
Rehabilitation equipment	96.15	86.21	90.91
Rehabilitation assessment	82.50	83.90	83.19
Medication	72.73	88.89	80.00
Body	81.25	80.25	80.75

information in the Chinese rehabilitation knowledge graph can assist the BiLSTM-CRF layer in capturing the contextual dependency information of the target text. When the knowledge graph embedding vectors are added at the BiLSTM layer, the lack of prior knowledge assistance in the network may result in missing some contextual dependency information, leading to a decrease in model accuracy.

The analysis of model performance in recognizing different rehabilitation entity types

This experiment further analyzes the results of NER to identify specific factors that may constrain the NER task for Chinese rehabilitation texts, providing theoretical directions for future research. The experiment records the recognition performance of the RMCCKG +BERT-BiLSTM-CRF model on six types of entities under a small sample environment with 500 training data entries on the Rehab dataset. The results are shown in Table 7.

From Table 7, we can see that there are significant differences in the recognition performance of different entities. Among them, the "functional disorders and manifestations" entity type performs best, with an F1-score of 91.89%, while the "rehabilitation methods" entity exhibits the worst recognition performance, with an F1-score of 78.45%. The size of the dataset and the distribution of labels differently impact the model's final prediction results. Therefore, we attempt to explore the differences in recognition results between different categories of entities from these two aspects.

We enumerate the frequency of occurrence of each entity in the training and test sets. For the same entity within the same category, if it appears multiple times, we also accumulate its frequency of occurrence. The results are shown in Table 8.

From Table 8, we can see that the "functional disorders and manifestations" entity type achieves the highest frequency of occurrence in both the training and test sets. The model can more accurately extract the feature information of these entities and provide correct

Table 8 The occurrences of different entity type in the training and test sets.

Entity type	Training set	Test set
Functional disorders and manifestations	652	152
Rehabilitation methods	550	113
Rehabilitation equipment	102	29
Rehabilitation assessment	473	110
Medication	48	9
Body	336	76

prediction results. On the other hand, the "medication" entity type shows the lowest frequency of occurrence in the dataset. The medication entity may appear only once in the dataset, making it difficult for the model to accurately grasp its feature information and provide high prediction accuracy. Although the "rehabilitation methods" entity type records the second-highest frequency of occurrence after "functional disorders and manifestations," its recognition accuracy is the lowest. Further statistics reveal that the "rehabilitation methods" entity type suffers from a severe low-frequency phenomenon, meaning that there are many types of entities, most of which appear only once. Additionally, the BERT pre-trained language model is trained on general medical domain data. The "medication" and "body" entities are also applicable in clinical medicine, and some of their features can be captured by BERT in advance. However, the "rehabilitation methods" entity type is specific to the rehabilitation field, which explains why it reaches a high frequency of occurrence but the lowest accuracy.

CONCLUSIONS

This article proposes the RMCCKG+BERT-BiLSTM-CRF NER model, which considers multiple Chinese character features and their interrelationships. Experimental results indicate that our proposed model significantly enhances entity recognition performance in Chinese rehabilitation medicine. The Chinese NER method introduced in this study treats Chinese characters as the smallest semantic unit, exploring both the inherent features of characters and the relationships among these features to extract semantic information from text, thereby improving the efficiency of entity recognition in text mining. Beyond rehabilitation medicine, RMCCKG establishes a paradigm for character-centric knowledge fusion in low-resource Chinese NLP. Its graph schema can be extended to other specialized domains by incorporating domain-specific characters and relationships. Furthermore, the RMCCKG constructed in this article incorporates not only rehabilitation medical terminology but also over eight thousand commonly used Chinese characters in its data sources, ensuring the generality and scalability of the knowledge graph. This makes the knowledge graph directly applicable to diverse fields. Future work will focus on further refining our proposed model with additional datasets and expanding knowledge graphs specifically tailored to the field of rehabilitation medicine.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Rehabilitation Medicine Knowledge Graph and Consulting System Development Project (No. W2021JSKF0741). Unisound AI Co., Ltd. provided financial and material support. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Rehabilitation Medicine Knowledge Graph and Consulting System Development Project: W2021JSKF0741.
Unisound AI Co., Ltd.

Competing Interests

Shengping Liu is an employee of the Unisound AI Co., Ltd.

Author Contributions

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

Data Availability

The following information was supplied regarding data availability:

The CMeEE dataset is available at AliBaba Cloud:

- https://tianchi.aliyun.com/dataset/144495
- https://github.com/CBLUEbenchmark/CBLUE

The Rehab dataset is available at GitHub:

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

The source code is available at GitHub:

- https://github.com/AaabbB-quick/A-Named-Entity-Recognition-model-embedding-rehabilitation-medicine-Chinese-character-knowledge-graph

- Sun, Mengting (2025). A Named Entity Recognition model embedding rehabilitation medicine Chinese character knowledge graph. figshare. Dataset. https://doi.org/10.6084/m9.figshare.29466521.v1.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3176#supplemental-information.

REFERENCES

Baum LE, Petrie T, Soules GW, Weiss N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* **41(1)**:164–171 DOI 10.1214/aoms/1177697196.

Bordes A, Usunier N, García-Durán A, Weston J, Yakhnenko O. 2013. Translating embeddings for modeling multi-relational data. In: *NIPS'13: Proceedings of the 27th International Conference on Neural Information Processing Systems*. Vol. 2. Red Hook, NY: Curran Associates, Inc., 2787–2795.

Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, MN: Association for Computational Linguistics.

Flórez E, Precioso F, Riveill M, Pighetti R. 2018. Named entity recognition using neural networks for clinical notes. In: *Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection*. PMLR, 7–15.

Grishman R, Sundheim B. 1996. Message understanding conference 6: a brief history. In: *Proceedings of the 16th Conference on Computational Linguistics*. Vol. 1.

Guo Q, Cai M, Liang Z. 2020. *Quick reference for rehabilitation medicine terminology*. Shanghai: Shanghai Scientific & Technical Publishers.

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

Lafferty JD, McCallum A, Pereira F. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Lee LH, Lu Y. 2021. Multiple embeddings enhanced multi-graph neural networks for Chinese healthcare named entity recognition. *IEEE Journal of Biomedical and Health Informatics* **25(7)**:2801–2810 DOI 10.1109/JBHI.2020.3048700.

Li Y, Du G, Xiang Y, Li S, Ma L, Shao D, Wang X, Chen H. 2020a. Towards Chinese clinical named entity recognition by dynamic embedding using domain-specific knowledge. *Journal of Biomedical Informatics* **106**:103435 DOI 10.1016/j.jbi.2020.103435.

Li J, Liu R, Chen C, Zhou S, Shang X, Wang Y. 2022. An RG-FLAT-CRF model for named entity recognition of Chinese electronic clinical records. *Electronics* **11(8)**:1282 DOI 10.3390/electronics11081282.

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 Z, Yang J, Gou X, Qi X. 2019a. Recurrent neural networks with segment attention and entity description for relation extraction from clinical texts. *Artificial Intelligence in Medicine* **97(5)**:9–18 DOI 10.1016/j.artmed.2019.04.003.

Li X, Zhang H, Zhou X. 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. 2019b. An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records. *BMC Medical Informatics and Decision Making* **19(S5)**:235 DOI 10.1186/s12911-019-0933-6.

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.

Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H, Wang P. 2020. K-BERT: enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence* **34(03)**:2901–2908 DOI 10.1609/aaai.v34i03.5681.

Liu L, Ding B, Bing L, Joty S, Si L, Miao C. 2021. MulDA: a multilingual data augmentation framework for low-resource cross-lingual NER. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics*, 5832–5843 DOI 10.18653/v1/2021.acl-long.453.

Pan SJ, Yang Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* **22(10)**: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.

Shi J, Sun M, Sun Z, Li M, Gu Y, Zhang W. 2022. Multi-level semantic fusion network for Chinese medical named entity recognition. *Journal of Biomedical Informatics* **133(6)**:104144 DOI 10.1016/j.jbi.2022.104144.

Sui D, Chen Y, Liu K, Zhao J, Liu S. 2019. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Stroudsburg: Association for Computational Linguistics, 3830–3840.

Tang B, Wang X, Yan J, Chen Q. 2019. Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF. *BMC Medical Informatics and Decision Making* **19**:74 DOI 10.1186/s12911-019-0787-y.

Vapnik V. 2000. *The nature of statistical learning theory*. Cham: Springer.

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.

Xiong Y, Peng H, Xiang Y, Wong K-C, Chen Q, Yan J, Tang B. 2022. Leveraging multi-source knowledge for Chinese clinical named entity recognition via relational graph convolutional network. *Journal of Biomedical Informatics* **128(4)**:104035 DOI 10.1016/j.jbi.2022.104035.

Xu K, Zhou Z, Gong T, Hao T, Liu W. 2018. 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**:114 DOI 10.1186/s12911-018-0690-y.

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: *IEEE International Conference on Bioinformatics*. Piscataway: IEEE, 892–897 DOI 10.1109/BIBM47256.2019.8983370.

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.

Zan H, Li W, Zhang K, Ye Y, Sui Z. 2021. Building a pediatric medical corpus: word segmentation and named entity annotation. *Available at https://tianchi.aliyun.com/dataset/95414*.

Zhang W, Jiang S, Zhao S, Hou K, Liu Y, Zhang L. 2019a. 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 Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. 2019. ERNIE: enhanced language representation with informative entities. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1441–1451.

Zhang B, Liu K, Wang H, Li M, Pan J. 2022. Chinese named-entity recognition via self-attention mechanism and position-aware influence propagation embedding. *Data & Knowledge Engineering* **139**:101983 DOI 10.1016/j.datak.2022.101983.

Zhang X, Zhang Y, Zhang Q, Ren Y, Qiu T, Ma J, Sun Q. 2019b. Extracting comprehensive clinical information for breast cancer using deep learning methods. *International Journal of Medical Informatics* **132(6)**:103985 DOI 10.1016/j.ijmedinf.2019.103985.

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 Chinese named entity recognition in rehabilitation medicine. *PeerJ Computer Science* **9(2)**:e1535 DOI 10.7717/peerj-cs.1535.