Chinese Medical Named Entity Recognition Using Pre-Trained Language Models and an Efficient Global Pointer Mechanism

Xu Zhang[1*], Feihong Li[2], Chenlong Li[2], Xufeng Yu[2]
[1]College of Software Engineering, Xiamen University of Technology, Xiamen 361000, China
[2]College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361000, China
[*]E-mail: zeronumber@sohu.com
[*]Corresponding Author

Keywords: MNER, data augmentation, RoBERTa, Word2Vec, EGP

Received: January 14, 2025

Medical Named Entity Recognition (MNER) is a critical task in medical text mining, serving as a foundation for intelligent diagnosis, disease prediction, and related tasks. However, Chinese medical texts present unique challenges, such as diverse entity classifications, varying entity lengths, and complex entity nesting, which hinder NER performance. To address these issues, we propose a Chinese MNER method leveraging pre-trained models and the efficient global pointer (EGP). Our approach incorporates three key innovations: (1) data augmentation techniques, including easy data augmentation and remote supervision with a medical entity dictionary, to address imbalanced entity types distribution and varying entity lengths; (2) a character-word fusion strategy that integrates RoBERTa-generated character vectors and Word2Vec-generated word vectors, enhancing semantic representation; and (3) an improved decoding layer based on EGP, enabling efficient recognition of both nested and flat entities while reducing computational costs. Experimentals show that our method achieves F1 scores of 75.87% and 92.77% on the CMeEE-V2 and CCKS2020 datasets, respectively, outperforming the RoBERTa-BiLSTM-CRF baseline by 3.06% and 4.38%, respectively.

Povzetek: Članek predstavlja metodo za prepoznavanje medicinskih poimenovanj v kitajskih besedilih, ki združuje predtrenirane modele in učinkovito globalno kazalo (EGP). Poudarek je na izboljšanju prepoznavanja zapletenih, gnezdenih entitet s strategijami, kot so povečanje podatkov in združevanje besednih ter znakovnih vektorjev, kar pripomore k večji kvaliteti.

1 Introduction

The exponential growth of medical informatics has led to the accumulation of a vast amount of medical text data, encompassing a multitude of domains such as medical journals, electronic medical records, medical encyclopedias and medical textbooks [1]. These resources, primarily in narrative form, represent a valuable repository of medical knowledge. In recent years, the rapid development of artificial intelligence (AI) technology has made data mining and relational extraction a key research area at the intersection of AI and medicine. Medical Named Entity Recognition (MNER), a core component of natural language processing (NLP), plays a critical role in medical text mining [2]. It serves as the foundation for tasks, such as medical entity relationship extraction and event extraction, while also providing high-quality data for downstream applications like intelligent medical question-answering systems and disease prediction models [3, 4]. Specifically, MNER identifies and categorizes nouns or fixed expressions in medical texts, including entities such as diseases and drugs. Accurate disease entity recognition is essential for understanding disease etiology, diagnosis, and prevention research, while drug entity recognition provides valuable insights for treatment methods. Thus, further research and optimization of MNER technology are crucial for advancing medical informatics and intelligence. Raw corpora, which are medical texts annotated by experts, are not only a key data source for MNER research but also essential for training and validating related algorithms [2]. The performance of machine learning-based methods heavily relies on the quality of annotated corpora. Therefore, high-quality, diverse, and comprehensive raw corpora are critical for MNER development. However, due to the high cost and expertise required for annotation, publicly available medical text datasets remain scarce. To address this, domestic and international initiatives have organized evaluation tasks to provide research data. For example, the N2C2 (National NLP Clinical Challenges) shared task, held annually since 2006, reflects advancements in international medical NLP [5]. Similarly, CCKS (China Conference on Knowledge Graph and Semantic Computing) and CBLUE (Chinese Biomedical Language Understanding Evaluation) have introduced the Chinese Medical Named Entity Extraction Evaluation Task Corpus and the CMeEE Series Corpus, providing substantial data and benchmarks for Chinese MNER research [1, 6]. These corpora not only promote standardized datasets in the medical industry but also support comprehensive research on Chinese MNER.

Due to the delayed development of medical informatics in China, early studies on Chinese MNER were largely influenced by English MNER research [7]. However, significant differences exist between English and Chinese, particularly in linguistic features. English has clearly defined word boundaries and relatively stable grammatical-syntactic structures, whereas Chinese medical texts present unique challenges, including a wide variety of entity types with significant scale differences, varying lengths of named entities, lengthy textual contexts, and underutilization of vocabulary information. As a result, text feature extraction models designed for English often underperform when applied to Chinese medical texts. Furthermore, Chinese medical texts contain not only flat entities but also a substantial number of structurally complex nested entities. Most mainstream MNER models treat the task as a sequence labeling problem. For example, the BERT-BiLSTM-CRF model first encodes text using BERT, then processes the results with BiLSTM to capture sequence information, and finally generates labeled outputs using CRF [8]. Although sequence labeling has achieved promising results in many evaluations, it struggles to accurately identify nested entities due to the limitations of the BIO annotation scheme, which allows only one label per token. The unique characteristics of Chinese medical texts and the complexity of nested entities pose significant challenges for Chinese MNER.

To tackle these challenges, this study focuses on three key research questions (RQs):

RQ1: How does the combination of Easy Data Augmentation (EDA) and remote supervision improve the model's ability to handle imbalanced entity types distribution and varying entity lengths in Chinese MNER?

RQ2: How does the integration of RoBERTa and Word2Vec enhance the semantic representation of medical texts, and what is its impact on the performance of Chinese MNER models?

RQ3: How does the Efficient Global Pointer (EGP) strategy optimize the decoding process, and what are its effects on computational efficiency and nested entity recognition in Chinese MNER?

The main contributions of this paper are as follows:

We propose a comprehensive data enhancement strategy combining EDA and remote supervision to mitigate the impact of imbalanced entity types distribution and varying entity lengths, thereby improving model recognition performance. Experimental results show that this strategy increases the F1 score for the "ite" entity type in the CMeEE-V2 dataset by 1.7%, outperforming the baseline model.

We design a word feature fusion method that integrates RoBERTa and Word2Vec to enhance semantic representation and model performance. Compared to using BERT as the encoding layer, this method achieves a 0.91% improvement in F1 score on CMeEE-V2.

We introduce the EGP strategy to optimize the decoding process, improving computational efficiency and nested entity recognition. On the CMeEE-V2 with nested entities, compared to the SOTA model RoBERTa-BiLSTM-CRF, EGP achieves a 3.06% improvement in F1 score and an increase of more than 15% in computational efficiency.

The rest of the paper is structured as follows. Section 2 describes the work related to traditional and machine learning-based NER. Section 3 details the proposed Chinese MNER method using pre-trained models and an EGP mechanism. Section 4 presents our experimental results, while Section 5 provides a comprehensive discussion of the findings. Finally, the conclusion is presented in Section 6.

2 Related works

The study of Chinese MNER models is a critical area in NLP research. Its technological evolution has transitioned from dictionary-based and rule-based approaches to machine learning methods. Recently, the rapid advancement of machine learning and the rise of deep learning have led to significant progress in Chinese NER tasks. This chapter aims to explore machine learning-based NER methodologies in detail, providing a thorough discussion and analysis of their key aspects.

2.1 Dictionary-based methods

Dictionary-based methods use matching algorithms to identify named entities by constructing a specialized dictionary of Chinese medical entities and their categories. Given the specialized nature of Chinese medical texts and the ambiguity of entity boundaries, a high-quality dictionary is essential for accurate entity recognition. Although Yang et al. developed a large-scale, standardized medical corpus, the dictionary cannot be updated in real-time to include new entities [9]. As a result, dictionary methods are now more commonly used as feature inputs, combined with other techniques to improve named entity recognition. For example, Wang et al. integrated a lexicon with a deep neural network, significantly enhancing the recognition of rare entities [10].

2.2 Rule-based methods

Rule-based methods require analyzing Chinese medical texts and constructing rule templates, which are then applied to entities of the same type for NER through pattern matching [11]. Compared to dictionary-based approaches, rule-based methods are easier to maintain and can better handle the non-standardized nature of Chinese medical texts. However, they demand significant time and human resources, and their rule portability is limited. Recent studies have demonstrated that combining rule-based techniques with machine learning algorithms can significantly improve the model's ability to recognize specific entity types [12]. For example, Chen et al. used external rules to extract entities that the model failed to identify, enhancing overall performance [13].

2.3 Machine learning methods

The application of machine learning techniques has significantly advanced Chinese MNER. This approach

involves training a model on labeled data and using it to perform NER on unlabeled text [14]. Compared to dictionary-based and rule-based methods, machine learning not only improves recognition efficiency but also adapts better to new entity types and domains, significantly reducing the need for human intervention and time. In the following sections, we will explore the application of machine learning in Chinese MNER from three perspectives: statistical machine learning, deep learning, and hybrid models.

2.3.1 Statistical machine learning methods

In Chinese MNER, supervised machine learning models are widely used, including Support Vector Machine (SVM), Hidden Markov Model (HMM), and Conditional Random Fields (CRF) [15-17]. For instance, Ju et al. used SVM to accurately identify medical nouns in complex biomedical texts, demonstrating its strong classification capabilities [18]. Gao et al. combined HMM with weighted learning techniques to develop a transfer learning model, significantly reducing the need for labeled data in the target domain [19]. Additionally, Liu et al. explored the impact of diverse feature types in Chinese MNER, achieving improved results by integrating medical dictionaries with CRF algorithms [20]. Despite these advancements, traditional machine learning methods require extensive feature engineering, which is time-consuming and resource-intensive. Moreover, their simple structures and limited parameter sizes restrict their representation and generalization capabilities, often leading to suboptimal performance on complex and diverse medical text data.

2.3.2 Deep learning methods

Deep learning has achieved remarkable success in areas such as speech and image recognition and has become a vital tool for NLP tasks [7]. In Chinese MNER, deep learning methods have eliminated the need for time-consuming feature engineering while enabling the identification of more complex features. Early deep learning models applied to Chinese MNER included convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [21, 22]. Notably, improved RNN-based methods have achieved exceptional performance in this field. For example, Dey et al. introduced a gating mechanism based on RNNs and developed the gated recurrent unit (GRU), which effectively addressed the gradient vanishing problem common in traditional RNN training [23]. Later, Xu et al. proposed a Chinese MNER model based on bidirectional long short-term memory (BiLSTM), which captures rich bidirectional contextual information in Chinese medical texts, significantly improving recognition accuracy and efficiency [24].

In recent years, pre-trained models have gained increasing attention in Chinese MNER research. In 2018, Devlin et al. introduced BERT, which set new benchmarks in NLP tasks due to its deep bidirectional representation and Transformer architecture [25, 26]. Zhong et al. developed the BERT-Span model for recognizing rehabilitation medicine entities in Chinese medical texts. This approach uses BERT to extract feature vectors and a span model to annotate entities, enhancing recognition performance [27]. Additionally, with the rapid development of NLP technology, traditional word vector models like Word2Vec and GloVe, as well as advanced pre-trained models such as MC-BERT, ChineseBERT, and RoBERTa, have been widely adopted in Chinese MNER research [28-32]. These models not only leverage BERT's strengths but also optimize it for medical applications, bringing new insights and innovations to the field.

2.3.3 Hybrid model methods

Hybrid models have emerged as a mainstream approach in MNER research. These models combine multiple types of features, architectures, or decoding strategies to leverage their strengths and improve overall recognition performance. For example, Gridach et al. proposed a BiLSTM-CRF neural network architecture that captures both character-level and word-level representations, enabling the automatic learning of rich contextual information in biomedical texts and significantly enhancing MNER accuracy [33].

To address the unique characteristics of Chinese medical texts, researchers have proposed various hybrid model methods. For example, Kimura et al. introduced multi-task learning for text classification, using NER as an auxiliary task. This approach not only improved text classification performance but also opened new directions for NER research [34]. Similarly, Wang et al. developed a BiLSTM-CRF model with a multi-task learning mechanism, enhancing the recognition of diverse named entity categories in medical texts [35]. In contrast, Ji et al. integrated an attention mechanism between BiLSTM and CRF, enabling the model to assign varying weights to each character and better capture critical information [36]. Jiang et al. proposed a BERT-BiLSTM-CRF model for Chinese electronic medical records, leveraging BERT's pre-training capabilities to obtain richer semantic representations [8]. To simplify the complex structure of BiLSTM, Qin et al. replaced BiLSTM with BiGRU, resulting in the BERT-BiGRU-CRF model [37]. Su et al. designed a global pointer decoding strategy, which addressed nested NER challenges through a unique entity scoring matrix and loss function, achieving impressive results on Chinese medical texts [38]. Additionally, Zhang et al. introduced a lexical enhancement approach based on the RoBERTa-Global Pointer model, integrating lexical information into character-level representations to improve semantic understanding and adaptability to Chinese medical texts [39].

In summary, hybrid models show great potential and broad development prospects in Chinese MNER. As shown in Table 1, existing state-of-the-art models face limitations such as inadequate use of pre-trained models and medical lexical information, difficulties in handling nested entities efficiently, and performance degradation due to data imbalance. Our proposed model targets these challenges, aiming to improve Chinese MNER

performance and provide stronger technical support for medical information processing, clinical decision support, and related applications.

Table 1: Summary of state-of-the-art models in Chinese MNER

	Model	Datasets	F1 (%)	Key limitations
1	CRF [17]	MSRA, LDC	84.73, 76.18	Heavy reliance on feature engineering and weak contextual modeling capabilities
2	SVM [18]	GENIA	82.44	Heavy reliance on feature engineering and difficulty in capturing contextual semantics.
3	BioTrHMM [19]	GENIA V3	65.41	Reliance on handcrafted features and weak ability to capture contextual semantics
4	BERT-CRF [26]	CCKS2017, CCKS2019, CMeEE	90.05, 77.19, 62.1	Difficulty in handling medical long texts and inability to recognize nested entities
5	BERT-span [27]	CRMNER	84.75	Difficulty in handling medical long texts and a large number of negative samples
6	MC-BERT-CRF [30]	CCKS2017, CCKS2019	92.03, 84.23	Difficulty in handling long texts and inability to recognize nested entities
7	ChineseBERT [31]	Weibo, MSRA	69.02, 99.14	Limited nested entity handling and poor domain adaptability
8	RoBERTa-CRF [32]	CCKS2017, CCKS2019	92.49, 79.85	Cannot recognize nested entities and underutilizes textual lexical information
9	BiLSTM-CRF [33]	JNLPBA	72.17	Weak contextual modeling capabilities and poor nested entity handling
10	Multi-Task-BiLSTM-CRF [35]	NCBI-Dis, JNLPBA	86.14, 73.52	Potential conflicts between tasks and insufficient utilization of textual semantic information
11	Attention-BiLSTM-CRF [36]	CCKS2018	87.26	The underuse of pre-trained models, noise from attention mechanisms, and failure to detect nested entities
12	BERT-BiLSTM-CRF [8]	CCKS2017	93.25	High model complexity and limited nested entity handling
13	BERT-BiGRU-CRF [37]	AIMEMR	90.38	Lack of domain-specific focus in the medical field and inability to recognize nested entities
14	BERT-GP [38]	CLUENER, CCKS2017, CMeEE	79.44, 94.71, 66.76	Insufficient utilization of character/word information, limited domain adaptability, and unnecessary computational resource consumption
15	RoBERTa-GP-SoftLexicon [39]	CCKS2017, CCKS2019	95.56, 85.89	Fails to address the issue of data imbalance in the dataset, and the decoding layer has room for computational resource optimization

3 Research method

This paper proposes a Chinese MNER model that integrates a pre-trained model with an efficient global pointer (EGP). The model consists of two main components: an encoding layer and a decoding layer. In the encoding layer, the model takes text as input and combines token-level vectors from RoBERTa with word-level vectors from Word2Vec, producing character-word fusion vectors enriched with semantic and lexical features. In the decoding layer, the EGP receives these fusion vectors and constructs an entity matrix, which is used to score and identify entities, determining their corresponding types. The overall framework of the model is shown in Fig. 1, which illustrates the outputs for the example: the microbial entity “virus” and the body type entity “target cell.”

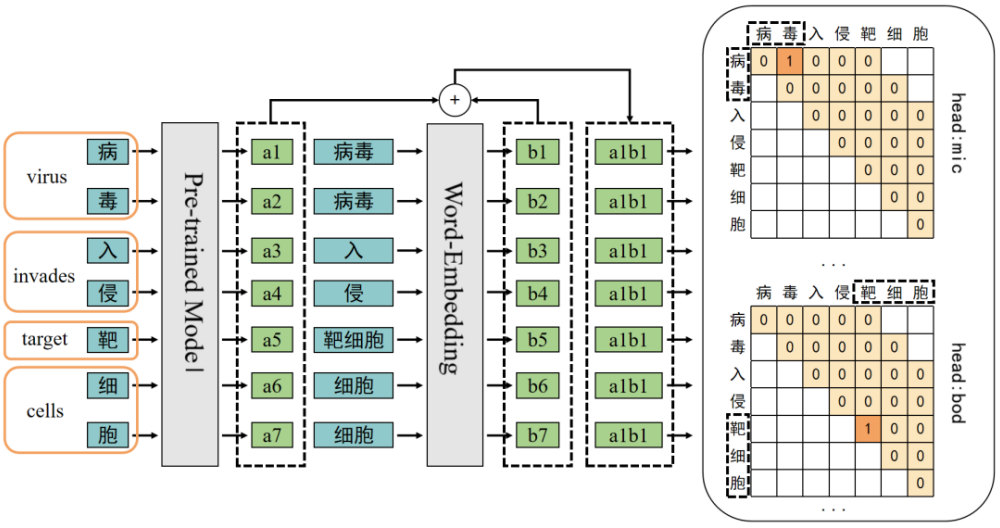

Figure 1: General framework of the model

3.1 Data augmentation

The performance of NER models heavily relies on high-quality annotated corpora, especially for Chinese medical texts. Constructing such corpora requires substantial time and effort due to the specialized nature of the domain. However, existing Chinese MNER datasets often struggle with issues like under representation of rare entity types (e.g., medical devices and laboratory tests) and significant variations in entity lengths (e.g., drug names ranging from short terms to long descriptive phrases). To address these challenges, data augmentation (DA) techniques have become a popular solution. Common DA methods for Chinese MNER include Easy Data Augmentation (EDA), remote supervision, and bootstrapping [40-42]. This paper proposes a DA strategy combining EDA and remote supervision, which aims to alleviate entity type imbalances and handle variable-length entities. By improving the diversity and quality of training data, this approach enhances the performance and generalization of the MNER model, particularly for rare and long entities.

3.1.1 Dataset's problem statement

The experiments in this paper use datasets from two authoritative Chinese MNER benchmarks, CBLUE and CCKS: the CMeEE-V2 nested dataset and the CCKS2020 flat dataset. Both datasets capture the unique characteristics of real-world Chinese medical texts but differ significantly in data volume, entity categories, and text formats. Specifically, the CMeEE-V2 dataset classifies medical entities into 9 categories: drug (dru), medical test item (ite), medical equipment (equ), microbiology class (mic), body part (bod), disease (dis), clinical symptom (sym), department (dep), and medical procedure (pro). In contrast, the CCKS2020 dataset groups entities into 6 categories: anatomical sites (ana), imaging tests (ima), drug (dru), diseases (dis), surgeries (sur), and laboratory tests (lab).

Preliminary experiments show that entity length significantly impacts recognition performance. Generally, longer entities are less frequent, making it harder for the model to learn meaningful features and leading to poorer recognition results. Based on these findings, this paper classifies Chinese medical entities into three length categories: short (1-3 characters), medium (4-7 characters), and long (8+ characters). Fig. 2 shows the proportional distribution of entities by length in the datasets. As seen in the Figure, "dis," "sym," and "pro" entities in the CMeEE-V2 dataset, as well as "dis" entities in the CCKS2020 dataset, contain a higher proportion of long entities, which are more challenging to recognize.

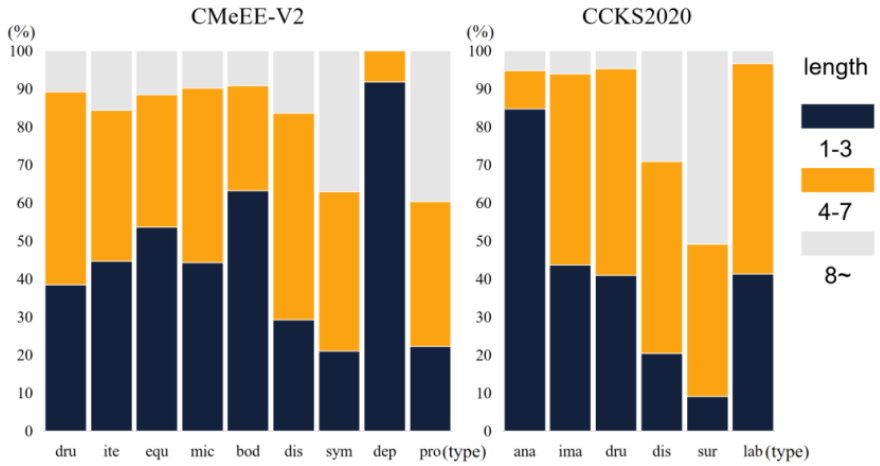

Figure 2: Percentage of different entity lengths in the datasets

Additionally, entity types with limited data are a major challenge in Chinese medical named entity recognition. The imbalanced entity types distribution and varying entity lengths stems from insufficient high-quality annotated data for these entities. Current deep learning models require large amounts of annotated data for effective training, and without it, their ability to learn and recognize these entities is severely limited, resulting in poor performance. To address this issue, this paper uses two data augmentation strategies EDA and remote supervision to expand imbalanced data and improve the model's learning capability.

3.1.2 EDA

In Chinese MNER, EDA is a widely used and efficient data augmentation technique. It introduces reasonable noise into the text to simulate the complexity and diversity of real-world samples. Compared to traditional EDA methods, this paper proposes an enhanced EDA approach with three strategies tailored to Chinese medical texts:

(1) Synonymous Entity Substitution: To address imbalanced entity type distributions, this technique replaces entities with semantically similar ones of the same type, focusing on underrepresented entities. This increases the data volume for rare entities, helping the model learn robust features and improve recognition accuracy.

(2) Entity Augmentation: To improve recognition of long entities, this technique constructs longer named entity sequences by combining similar short entities of the same type using conjunctions or punctuation. By increasing the number of long entity samples, it enables the model to better learn structural and semantic patterns of long sequences, enhancing performance.

(3) Sentence Replacement: To help the model learn contextual dependencies and adapt to the lengthy nature of medical texts, this technique replaces two sentences within or between samples with a randomly selected sentence, using punctuation or separators as the basis for exchange. By altering sentence order and combinations, it generates diverse contextual environments, improving the model's ability to learn long-distance dependencies.

These improved strategies not only retain the efficiency of EDA but also better align with the characteristics of Chinese medical texts, enhancing the performance and robustness of Chinese MNER.

3.1.3 Remote supervision

Remote supervision is an effective data augmentation technique that leverages existing knowledge bases or structured data to automatically annotate large-scale unlabeled texts. In Chinese MNER, remote supervision is commonly used by introducing a large corpus of unsupervised medical texts and automatically annotating them using script programs. This approach effectively converts unsupervised samples into richly annotated supervised data, significantly expanding the dataset for model training.

The second DA method proposed in this paper extends remote supervision with an innovative approach. First, we construct a comprehensive medical entity dictionary based on a large annotated MNER training set. To improve the dictionary's coverage and accuracy, we use ERNIE Bot, a generative pre-trained large model optimized for Chinese, to generate additional instances of rare entity types and long entities, addressing imbalances in entity type and length distributions. Next, we apply entity matching techniques to automatically annotate unsupervised medical texts using the expanded dictionary. To ensure annotation quality, the automatically labeled data is further processed and validated by ERNIE Bot, which performs contextual verification for semantic consistency, corrects mislabeled entities, and assigns confidence scores, retaining only high-confidence annotations for training.

This approach not only increases the volume of training data but also ensures its high quality, providing the MNER model with more robust and diverse samples for improved performance, especially for rare and long entities.

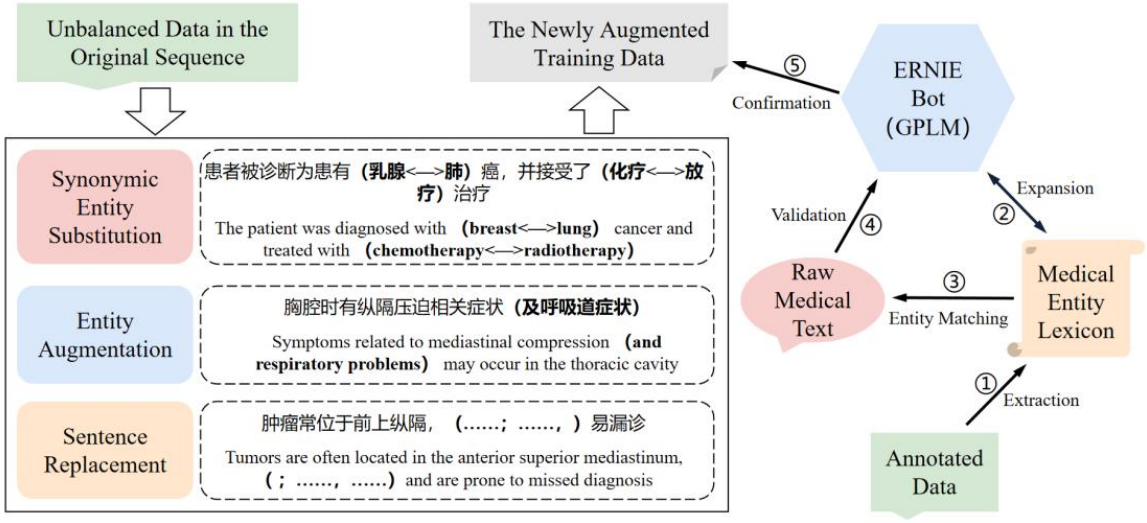

Figure 3: Data augmentation process

This paper proposes an innovative method to address poor recognition performance for specific entities caused by imbalanced entity types and varying entity lengths. The method combines two data augmentation strategies: EDA and remote supervision. The data augmentation process is illustrated in Fig. 3, with EDA on the left and remote supervision on the right. By enhancing EDA operations for Chinese medical texts—synonymous entity substitution, entity augmentation, and sentence replacement—this approach not only mitigates entity type and length imbalances but also improves the model's ability to process long medical texts. Additionally, this paper integrates a medical entity dictionary with the GPLM ERNIE Bot using remote supervision, effectively converting a large volume of unsupervised medical texts into high-quality annotated data. This significantly expands the training dataset, particularly for long entities and rare entity types [43]. Furthermore, the annotation process introduces new syntactic structures and nested entities, enriching dataset diversity. The combination of these two strategies not only enhances the performance and robustness of Chinese MNER but also provides new insights and methodologies for tackling similar imbalanced dataset challenges.

3.2 Encoding layer

In Chinese MNER, the encoding layer is the core component of the model architecture, responsible for converting raw Chinese medical text into numerical vectors that computers can process. Given the complex contextual dependencies and specialized medical domain vocabulary in medical texts, the design of the encoding layer is critical for improving NER accuracy and efficiency. This model uses a fusion encoding strategy combining RoBERTa and Word2Vec, generating character-word fusion vectors that capture both deep semantic information and word-level features.

3.2.1 RoBERTa

The core component of the encoding layer in the proposed Chinese MNER model is RoBERTa, which generates dynamic character vectors. As an optimized version of BERT, RoBERTa has shown excellent performance in NLP tasks, including NER. RoBERTa and BERT share several key similarities: both are based on the Transformer architecture, using the encoder component to capture bidirectional contextual information, allowing them to consider both preceding and following contexts when processing text and produce more accurate token vector representations. Both also use the masked language model (MLM) as their pre-training strategy, where MLM randomly masks words in the input text and trains the model to predict them, enabling the model to learn comprehensive linguistic knowledge and contextual information. Additionally, both models have similar structures, consisting of multiple stacked Transformer encoders, each incorporating a self-attention mechanism and a feed-forward neural network, which allows them to efficiently capture dependencies and semantic information in long texts. Both RoBERTa and BERT are versatile and applicable to a wide range of NLP tasks, such as text classification, NER, and question answering. In MNER, they significantly enhance model performance and improve the understanding and recognition of critical information in medical texts.

The RoBERTa encoding layer used in this paper not only retains BERT's core advantages but also introduces several optimizations to improve performance. These improvements include:

(1) Dynamic Masking: RoBERTa uses a dynamic masking strategy, generating a unique mask pattern for each input sequence. This helps the model better capture language variability and contextual dependencies.

(2) Remove the Next Sentence Prediction Task: By eliminating this task, RoBERTa focuses more on learning linguistic structures and semantics, which is particularly

useful for understanding the complex characteristics of Chinese medical texts.

(3) Increasing Training Data and Batch Size: RoBERTa is trained on 160 GB of data with a batch size of 8,000, enabling the model to learn richer medical knowledge and language patterns.

(4) Extending Training Text Length: RoBERTa processes longer sentences, enhancing its ability to handle complex text structures and dependencies, which is critical for lengthy Chinese medical texts.

In this study, we use the RoBERTa model fine-tuned by Harbin Institute of Technology (HIT) for Chinese text [44]. This version of RoBERTa has been specifically adapted to Chinese linguistic features, such as word segmentation and semantic structures. The fine-tuning process involved training on a large-scale Chinese corpus, primarily Chinese Wikipedia data, with two configurations: (1) a sequence length of 128 and a batch size of 2,560 for 100 k steps, and (2) a sequence length of 512 and a batch size of 384 for another 100 k steps. This extensive fine-tuning allows the model to better capture Chinese language nuances and improve performance on tasks like NER.

For Chinese MNER, text data is characterized by frequent domain-specific terminology, complex semantic structures, and longer sentence lengths compared to general text. These characteristics make RoBERTa, with its optimizations and Chinese-specific fine-tuning, an ideal choice for the encoding layer in our model.

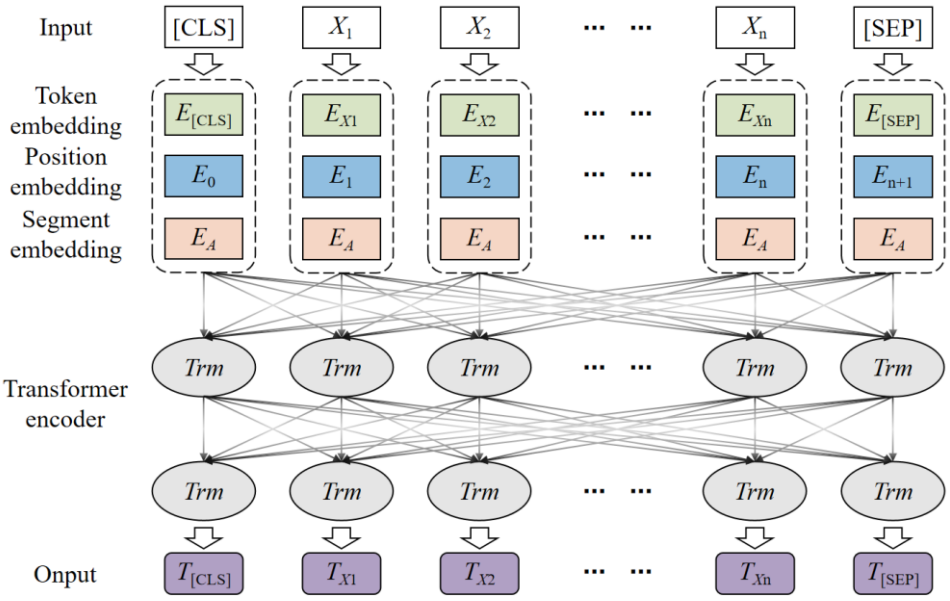

Figure 4: RoBERTa generation encoding vector process

Fig. 4 illustrates the process of encoding Chinese medical text sequences into token vectors using RoBERTa. The first step involves tokenizing the input Chinese medical text sequence using a dedicated tokenizer. Due to the unique characteristics of Chinese text, the tokenizer divides the sequence into character-level tokens, with each character serving as the basic unit. Additionally, the RoBERTa tokenizer automatically inserts special tokens [CLS] and [SEP] to mark the start and end of the text sequence. Next, the character tokens are passed through RoBERTa's embedding layer, which performs token embedding, position embedding, and segment embedding. However, in the Chinese MNER task studied here, segment embedding is typically unnecessary since the input usually consists of a single paragraph. To improve the accuracy of character sequence positional information, this study uses Rotary Position Embedding (RoPE), making positional embedding an optional step [45]. After embedding, the vectors are fed into RoBERTa's core, which consists of multiple stacked Transformer encoders. Each encoder includes self-attention mechanisms and feed-forward neural networks, enabling it to capture contextual and semantic information and learn more nuanced semantic representations of Chinese medical text. Finally, after processing by multiple Transformer encoders, the model outputs a sequence of vector representations for each character token. These vectors capture both the intrinsic information of the characters and their contextual semantics, making them suitable for downstream tasks like Chinese MNER.

Given a Chinese medical text sequence $\left[x_1, x_2, \cdots, x_n\right]$, the context vector representation of the i-th character x_i, is defined by the following Eq. (1):

$$c_i = RoBERTa\left(x_i\right) \tag{1}$$

3.2.2 Word2Vec

Word2Vec is a deep learning model designed to generate word vectors, converting natural language words into numerical vectors for use in various NLP tasks. Other word vector models, such as Global Vectors for Word Representation (GloVe) and FastText, have also shown

strong performance across tasks [29, 46]. Word2Vec employs two main methods: the Skip-gram model, which predicts context words based on a central word (ideal for low-frequency words), and the CBOW model, which predicts a central word from context words (effective for high-frequency words). The Word2Vec workflow is straightforward: text data is first converted into coded vectors, which are then fed into a neural network for training. The model optimizes its parameters through back-propagation, aiming to maximize the probability of predicting context or center words. Ultimately, it outputs continuous vector representations of words, reflecting their semantic similarities. These representations are widely used in Chinese MNER tasks.

In Chinese medical texts, words are typically composed of multiple adjoining characters, unlike English, where words are separated by spaces. This linguistic feature poses unique challenges for NER, as the model must capture both character-level contextual dependencies and word-level semantic information. To address this, we combine Word2Vec word vectors with RoBERTa character vectors, leveraging their respective strengths: Word2Vec excels at word-level semantics, while RoBERTa handles character-level contextual relationships. However, this fusion introduces a dimension inconsistency issue, as word vector sequences are shorter than character sequences. To resolve this, we modify the word vector generation process by adding special identifiers [CLS] and [SEP] at the start and end of the text, respectively, to mark sequence boundaries. Additionally, each split word is repeated n times, where n is the character length of the word. While this introduces some redundancy, it effectively simulates character-level contextual information, helping the model better understand relationships between characters within words. This is particularly important for Chinese medical NER, as many domain-specific terms are multi-character, requiring both character-level and word-level representations. The model generates word vectors according to the following Eq. (2):

$$w_i = Word2Vec(y_i) \tag{2}$$

Where is the word corresponding to the i-th character x_i of the input text, and w_i is the word vector generated by Word2Vec for the word y_i.

By combining Word2Vec and RoBERTa, our model can better handle the high frequency of domain-specific terminology and the complex semantic structures inherent in Chinese medical texts. This dual representation approach not only addresses the technical challenge of dimension inconsistency but also enhances the model's ability to comprehend and recognize medical entities in Chinese texts.

3.2.3 Character-word fusion vectors

In Chinese language processing, effectively fusing character vectors and word vectors has been shown to enhance the model's ability to capture contextual and lexical information, particularly in tasks like NER [47]. This is especially important for Chinese medical texts, where single characters often carry significant semantic meaning, particularly in multi-character domain-specific terms. By combining character-level and word-level representations, the model can better capture the nuanced relationships between characters and words, which is critical for accurate entity recognition.

Character-word fusion strategies can be broadly categorized into two approaches: (1) combining character-based word representation with the original word representation, and (2) treating characters as the fundamental semantic unit and enhancing their representations with word-level information [48]. Various techniques can be used for fusion, including vector addition, multiplication, concatenation, weighted representation via feed-forward neural networks, and tensor-based abstraction. In this study, we adopt concatenation as the fusion strategy due to its simplicity, intuitive results, and computational efficiency. Given the significant difference in dimensionality between RoBERTa-generated character vectors and Word2Vec-generated word vectors, concatenation provides an effective way to preserve complete information from both representations while maintaining manageable computational complexity.

While concatenation is our primary fusion method, future work could explore alternative strategies, such as weighted representations or tensor-based approaches, to further optimize the fusion process.

The character-word fusion vector method used in this paper can be formally expressed as Eq. (3).

$$h_i = concat(c_i, w_i) \tag{3}$$

Where c_i and w_i correspond to the word vector and word vector of the i-th character x_i, respectively, and h_i is the fusion vector obtained by concatenating the character-word vectors.

3.3 Decoding layer

In traditional Chinese MNER frameworks, the task is typically treated as a sequence labeling problem, where text is annotated using schemes like BIO or BIOES, and a CRF layer outputs label probability score. However, this approach struggles with complex medical texts containing nested or overlapping entities. Traditional sequence labeling methods assume entities are non-overlapping, making it difficult to handle nested structures directly. To address this, additional post-processing steps, such as rule-based heuristics or extra model layers, are often required, significantly increasing computational complexity and resource usage.

To overcome these limitations, we adopt a pointer network approach, which has proven effective for Chinese MNER. Pointer networks directly mark any position in the text, enabling accurate identification of nested and overlapping entities without relying on fixed label sequences like BIO or BIOES. This end-to-end approach eliminates the need for post-processing, making it more

efficient and flexible for handling complex entity structures.

Building on the strengths of pointer networks, this paper proposes a decoding layer strategy based on an EGP. The EGP incorporates a global vision component, allowing the model to comprehensively capture the text's overall information during decoding. Additionally, the EGP optimizes algorithm design to reduce the computational complexity of entity matrix scoring, providing robust support for decoding in complex Chinese MNER tasks. These features make the EGP particularly suitable for handling nested entities and other challenges in medical text analysis.

3.3.1 Entity matrices

In the Chinese MNER task, the EGP approach treats the problem as a multi-label classification by constructing a 3-dimensional matrix framework that captures all potential entities and their types in a sequence. The matrix dimensions are (Number of Entity Types, Sequence Length, Sequence Length), where each element represents the probability of a contiguous segment being an entity of a specific type. For the CMeEE-V2 dataset, which includes 9 entity types, 9 two-dimensional entity matrix slices are constructed, each focusing on a single entity type. The rows and columns of these matrices correspond to the start and end positions of entities in the sequence, respectively, as shown in Fig. 5. When the model predicts a specific segment as an entity, the corresponding matrix element transitions from 0 to 1, effectively marking the entity's boundaries.

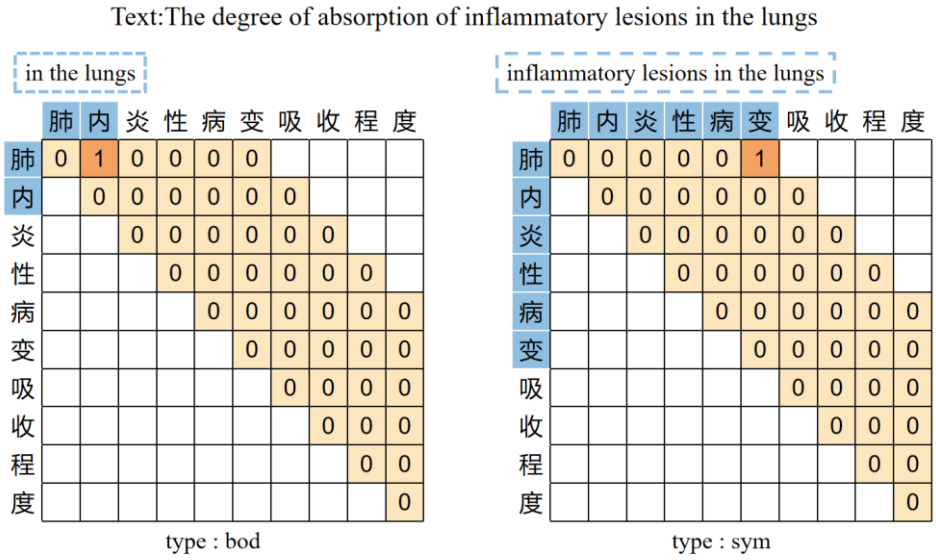

Figure 5: Entity matrices

During the construction of the entity matrix, the lower-left triangle is ignored, and a masking operation is applied to parts of the upper-right triangle, as shown in Fig. 5. This design is motivated by two key considerations:

(1) Logical Consistency: It is illogical for an entity's start position to be after its end position. Thus, calculating entity probabilities for the lower-left triangle is unnecessary and would introduce noise into the model's predictions.

(2) Entity Length Constraints: In Chinese medical texts, entity lengths often differ significantly from the overall text length. The upper-right triangle represents sequence segments that are typically longer than regular medical entities. By masking this region, we reduce unnecessary computations and focus the model on plausible entity lengths.

The masking operation not only reduces computational complexity but also improves the model's ability to identify nested entities. By eliminating invalid entity combinations (e.g., end positions earlier than start positions) and focusing on plausible entity lengths, the masking operation ensures the model can more accurately detect nested entities. For example, in the phrase “肺部炎症性病变” (inflammatory lesions in the lungs), the masking operation helps the model distinguish between the outer entity “肺部炎症性病变” (symptom type) and the nested inner entity “肺部” (body type) by restricting the search space to valid entity boundaries.

The EGP designed in this paper builds on these principles, making targeted improvements to the original global pointer approach. By optimizing computational resources and narrowing the search range for candidate entities, the masking operation enhances the accuracy and efficiency of Chinese MNER, particularly for complex nested structures.

3.3.2 Entity scoring function and loss function

The EGP is used to score entities for each element in the entity matrix. Requiring a series of transformations on the encoded character-word fusion vectors.

For an input text $[x_1, x_2, \cdots, x_n]$ of sequence length n, the sequence of word fusion vectors obtained after encoding is $[h_1, h_2, \cdots, h_n]$. The transformation equations are defined as follows:

$$q_{i,\alpha} = W_{q,\alpha} h_i + b_{q,\alpha} \tag{4}$$

$$k_{i,\alpha} = W_{k,\alpha} h_i + b_{k,\alpha} \tag{5}$$

Where h_i is the character-word fusion vector representation of the i-th character of the text, $W_{q,\alpha}$ and $W_{k,\alpha}$ are the two transformation matrices, $b_{q,\alpha}$ and $b_{k,\alpha}$ are the biases, $q_{i,\alpha}$ and $k_{i,\alpha}$ are the new vector representations obtained by transforming h_i.

To obtain the probability score that a consecutive sequence from the i-th character to the j-th character is an entity of type α, we compute the inner product of the transformed vector representations $q_{i,\alpha}$ and $k_{j,\alpha}$. The entity scoring function is defined by Eq. (6).

$$s_\alpha(i, j) = q_{i,\alpha}^{\cdot} k_{j,\alpha} \tag{6}$$

Where is the probability score of a continuous textual sequence $[x_i, \cdots x_j]$ predicted to be of entity type α, x_i is the beginning of the entity, and x_j is the end of the entity.

The EGP approach identifies entities from a global perspective but may lack sensitivity to specific entity lengths and spans, potentially misclassifying similar entities positioned head-to-tail as the target entity. Traditional position embedding strategies assign static, predefined vectors to each position, which struggle to capture relative relationships when a single character corresponds to multiple entity boundaries, as in nested entities. This limitation is particularly problematic for Chinese medical texts, where relative character positions within multi-character terms and nested entities are critical for accurate recognition.

To address this, the EGP introduces Rotary Position Embedding (RoPE), which incorporates relative positional information into the character-word fusion vector representation through a transformation matrix, as defined in Eq. (7). RoPE enables the model to flexibly capture relative character positions, which is essential for handling nested and long entities in Chinese medical texts. For example, in the term "肺部炎症性病变" (inflammatory lesions in the lungs), RoPE helps the model accurately identify nested entity boundaries by capturing the relative positions of characters within the term. Additionally, the entity scoring function is redefined to incorporate relative positional information, as shown in Eq. (8), further enhancing the model's ability to recognize complex entity structures.

$$R_i^{\cdot} R_j = R_{j-i} \tag{7}$$

$$s_\alpha(i, j) = q_{i,\alpha}^{\cdot} R_{j-i} k_{j,\alpha} \tag{8}$$

The EGP treats NER as a multi-label classification problem, aiming to identify entities of α specific type, with a total of $(2n+1-k)k/2$ candidate entities. Here, n represents the text sequence length, k denotes the maximum sequence length that can be predicted as an entity, and $n-k$ signifies the edge length of the upper-right triangular mask. However, in Chinese medical texts, the text sequence length n is often considerable, as is the number of candidate entities. Moreover, the number of entities present in each text is likely to be relatively limited. If the loss function is designed as a $(2n+1-k)k/2$ binary classification, the model may tend to predict all samples as negative due to the overwhelming number of negative samples (non-entities) compared to positive samples (entities). This would result in a significant class imbalance issue, severely impacting the model's performance.

To address this, the EGP designs a multi-label classification loss function, which generalizes the single-objective multi-classification cross-entropy. This approach is particularly suitable for multi-label classification problems with many total categories but few target categories [49]. By independently handling each category, the model can focus on the distribution of positive and negative samples within that category. Additionally, the flexibility to assign higher weights to positive samples helps alleviate the issue of limited positive sample quantities. This makes the multi-label classification loss function better suited for texts containing entities of different types, enhancing its effectiveness for Chinese MNER tasks involving nested entities. The loss function for multi-label classification is defined in Eq. (9).

$$\begin{aligned} Loss = &\log\left(1 + \sum_{(i,j)\in P_\alpha} e^{-s_\alpha(i,j)}\right) \\ &+ \log\left(1 + \sum_{(i,j)\in Q_\alpha} e^{s_\alpha(i,j)}\right) \end{aligned} \tag{9}$$

Where P_α is the first and last set of all entities of type α in the candidate entities, and Q_α is the first and last set of all non-entities and non-entities of type α in the candidate entities. In the decoding phase, we only need to consider the sequence $[x_i, \cdots x_j]$ of $1 \le i \le j \le k$, which is considered to be the output of entities of entity type α if $s_\alpha(i, j) > 0$ is satisfied.

4 Experiment

4.1 Experimental parameters and assessment indicators

The experimental environment in this study includes a computer with an Intel i5-13600KF CPU, an NVIDIA RTX 4070Ti GPU, 32 GB RAM, and a 1 TB hard disk, which are sufficient to ensure efficient training and fine-tuning of deep learning models. For the two datasets, CMeEE-V2 and CCKS2020, we first preprocessed the data by splitting the CCKS2020 dataset into training (70%), validation (15%), and test (15%) sets and converting the data into JSON format for consistency.

Next, we conducted a meticulous fine-tuning parameter optimization process. To identify the optimal model configuration, we defined a narrow range of candidate values for key hyperparameters, including the learning rate, batch size, number of training epochs, and optimizer parameters, based on existing research and preliminary trial-and-error experiments. We then used a grid search algorithm to systematically evaluate all possible combinations of these parameters and determine the best configuration. The final parameter settings are shown in Table 2.

Table 2: Experimental parameters

Hyperparameter	Value
Learning rate	2e-5
Batch size	8
Epochs	10
Max sequence length	256
Optimizer	Adam
Warmup	0.1

The experiment uses precision (P), recall (R), and F1 score as evaluation metrics. Precision represents the ratio of correctly identified entities to the total predicted entities, recall signifies the ratio of correctly identified entities to the total actual entities, and the F1 score is a harmonic mean of precision and recall, reflecting the overall performance of the model. The formulas are as follows:

$$p = \frac{TP}{TP + FP} \times 100\% \tag{10}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{11}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \times 100\% \tag{12}$$

Where *TP* represents the number of samples where the model predicted a positive category and the actual category was positive. *FP* denotes the number of samples where the model predicted a positive category but the actual category was negative. *FN* represents the number of samples where the model predicted a negative category but the actual category was positive.

4.2 Experimental results

To comprehensively evaluate the performance of the pre-trained models combined with the efficient global pointer method (DA-RW-EGP, Data Augmentation-RoBERTa with Word2Vec-Efficient Global Pointer), we conducted a detailed comparison with several popular benchmark methods on two widely recognized Chinese medical datasets, CMeEE-V2 and CCKS2020. Given that the CMeEE-V2 dataset contains nested entities, we specifically included comparison methods capable of effectively handling nested NER. The experimental results are summarized in Tables 3 and 4, which clearly illustrate the performance of the proposed methods across several key metrics.

Table 3: CMeEE-V2 comparison experiment

	Model	Pre (%)	Rec (%)	F1 (%)
1	BERT	69.65	70.51	70.08
2	RoBERTa	70.07	70.92	70.49
3	BERT-CRF	70.93	73.24	72.07
4	RoBERTa-BiLSTM-CRF	72.90	72.72	72.81
5	RoBERTa-GP	74.01	75.05	74.53
6	RoBERTa-GP-SoftLexicon	73.87	75.59	74.72
7	**DA-RW-EGP (ours)**	**74.87**	**76.91**	**75.87**

Table 4: CCKS2020 comparison experiment

	Model	Pre (%)	Rec (%)	F1 (%)
1	BERT	86.83	88.16	87.49
2	RoBERTa	87.06	88.53	87.79
3	BERT-CRF	88.21	89.11	88.66
4	RoBERTa-BiLSTM-CRF	87.85	88.93	88.39
5	RoBERTa-GP	89.99	93.29	91.61
6	RoBERTa-GP-SoftLexicon	90.76	93.16	92.08
7	**DA-RW-EGP (ours)**	**91.52**	**94.01**	**92.77**

The experimental results on the CMeEE-V2 and CCKS2020 datasets demonstrate the effectiveness of the proposed DA-RW-EGP method. As shown in Table 3, on the CMeEE-V2 dataset, our method achieves an F1 score of 75.87%, outperforming all baseline models, including RoBERTa-GP-SoftLexicon (74.72%) and RoBERTa-BiLSTM-CRF (72.81%). Similarly, on the CCKS2020 dataset (Table 4), the proposed method attains an F1 score of 92.77%, surpassing RoBERTa-GP-SoftLexicon (92.08%) and RoBERTa-BiLSTM-CRF (88.39%). These results highlight the superior performance of the DA-RW-EGP method in Chinese MNER, particularly in handling nested entities and achieving higher precision and recall.

Statistical Significance Analysis: To validate the performance improvements of the proposed DA-RW-EGP model, we conducted t-tests on both the CMeEE-V2 and CCKS2020 datasets. For each model and dataset, we ran

multiple experiments, obtaining 10 F1 scores for each configuration. These scores were then used to perform t-tests, ensuring the robustness of our results. On the CMeEE-V2 dataset, the t-test results show a t-value of 33.14 and a p-value much smaller than 0.005, indicating that the performance improvement of DA-RW-EGP is statistically significant. Consistent results were observed on the CCKS2020 dataset, as shown in Table 5, further confirming the superiority of the proposed method. These statistical tests provide strong evidence that the performance gains are not due to random variations but reflect the true effectiveness of the model.

Table 5: t-test comparison of RoBERTa-BiLSTM-CRF and DA-RW-EGP on CMeEE-V2 and CCKS2020

Dataset	DA-RW-EGP (F1%)	RoBERTa-BiLSTM-CRF (F1%)	t	p	α
CMeEE-V2	75.842±0.032	72.778±0.043	33.14	≪ α	0.05
CCKS2020	92.76±0.009	88.393±0.014	81,31	≪ α	0.05

5 Discussion

5.1 Comparison with related works

The F1-score is a core evaluation metric for measuring the performance of Chinese medical named entity recognition models. In this section, we evaluate the proposed DA-RW-EGP model on two benchmark datasets: CMeEE-V2 (with nested entities) and CCKS2020 (without nested entities). As shown in Table 3, the proposed model achieves an F1-score of 75.87% on the CMeEE-V2 dataset, demonstrating its effectiveness in handling nested entities. Additionally, Table 4 shows that the model attains an F1-score of 92.77% on the CCKS2020 dataset, highlighting its robust performance in non-nested scenarios.

To contextualize these results, we compare the proposed model with several state-of-the-art (SOTA) models, including the sequence labeling models BERT-CRF [26] and RoBERTa-BiLSTM-CRF [8], as well as the pointer network models RoBERTa-GP [38] and RoBERTa-GP-SoftLexicon [39]. As summarized in Table 3 and Table 4, the proposed model outperforms BERT-CRF by 3.80% and 4.11% on the CMeEE-V2 and CCKS2020 datasets, respectively. Compared to RoBERTa-BiLSTM-CRF, the improvements are 3.06% and 4.38%. Additionally, the proposed model surpasses RoBERTa-GP by 1.34% and 1.16%, and RoBERTa-GP-SoftLexicon by 1.15% and 0.69% on the same datasets. These improvements in F1-score clearly demonstrate the superior performance of the DA-RW-EGP model in Chinese medical named entity recognition tasks.

Furthermore, to evaluate the generalization capability and performance of the proposed model, we conducted experiments on several datasets used by SOTA models in the Related Works table (Table 1). On the CCKS2017 and CCKS2019 datasets, the proposed model achieved F1-scores of 95.88% and 86.35%, respectively. On the CCKS2017 dataset, the proposed model outperformed MC-BERT-CRF [30], RoBERTa-CRF [32], BERT-BiLSTM-CRF, and BERT-GP [38] by 3.85%, 3.39%, 2.63%, and 1.17%, respectively. On the CCKS2019 dataset, the proposed model surpassed MC-BERT-CRF and RoBERTa-CRF by 2.12% and 6.5%, respectively. These results further validate the generalization ability and superior performance of the proposed model.

5.2 Reasons for superior performance

The superior performance of the proposed DA-RW-EGP model can be attributed to several key innovations, including data augmentation, character-word fusion vectors, and the EGP module. Below, we analyze each component in detail and validate its contribution through ablation studies.

5.2.1 Data augmentation

To address the issues of imbalanced entity type distributions and varying entity lengths in Chinese MNER datasets, we adopt a comprehensive data augmentation strategy. This approach generates high-quality synthetic samples for rare entity types and long entities, significantly improving the model's ability to recognize these challenging cases. Data augmentation is essential for both sequence labeling models and pointer network models, as they require large amounts of annotated data for training and validation. By enriching the training set, our method ensures balanced learning across different entity types and lengths, ultimately enhancing the model's overall performance.

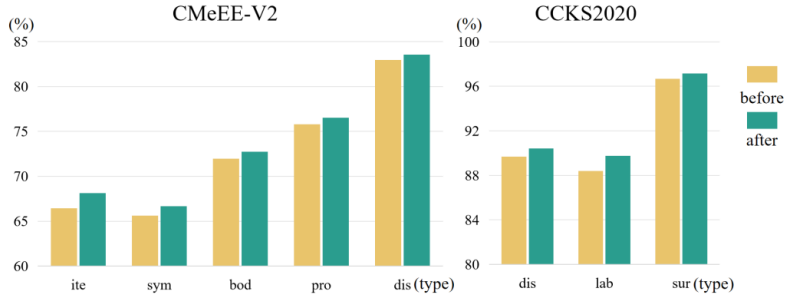

Figure 6: Comparison of F1-scores before and after data augmentation

As shown in Fig. 6, we compare the F1-scores of selected entity types before and after data augmentation on the two datasets. The results show a slight improvement in F1-scores for most entity types. For normal-scale entity types, this improvement is due to the increased number of long entities in the augmented data, enhancing the model's ability to recognize entities of varying lengths. Notably, the F1-scores for the "ite" entity type in the CMeEE-V2 dataset and the "lab" entity type in the CCKS2020 dataset increased by 1.7% and 1.3%, respectively. This improvement is attributed to the low initial data representation of these entity types. Through EDA and remote supervision, we generated a large number of high-quality pseudo-samples, improving the model's ability to recognize these rare entities. These results demonstrate that our comprehensive data augmentation strategy effectively addresses the challenges of imbalanced entity types distributions and varying entity lengths in Chinese MNER tasks.

5.2.2 Character-word fusion vectors

In Chinese MNER, medical lexical information is critical. As shown in Tables 3 and 4, the RoBERTa-GP-SoftLexicon model, which incorporates simple lexical information, outperforms RoBERTa-GP by 0.19% and 0.47% on the two datasets, respectively. In this work, we further enhance the model's encoding layer by integrating comprehensive medical lexical information using Word2Vec, generating character-word fusion vectors. This capability is absent in single pre-trained models like MC-BERT, ChineseBERT, and RoBERTa listed in Table 1. By leveraging this approach, our proposed DA-RW-EGP model captures richer semantic representations, enhances learning capacity, and significantly improves recognition performance.

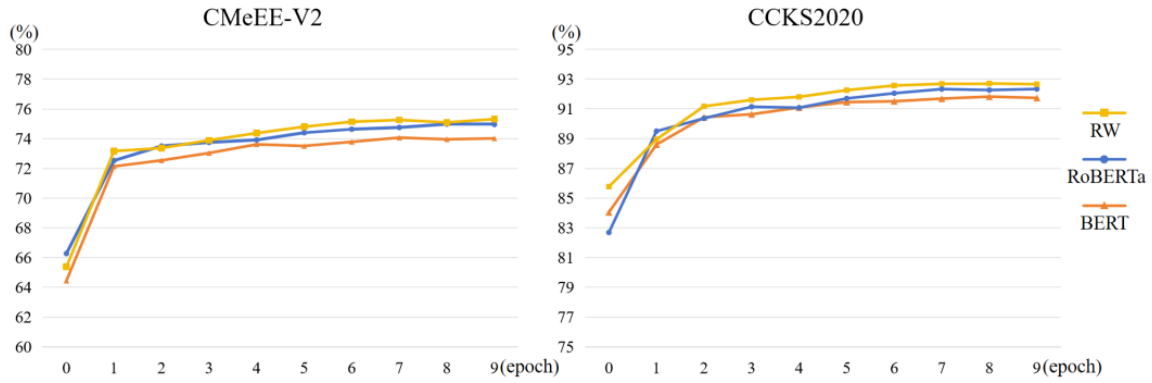

Figure 7: Comparison of F1 values for different encoding models

As illustrated in Fig. 7, we conducted comparative experiments on the model encoding layer using RW (RoBERTa with Word2Vec), RoBERTa, and BERT. On both datasets, the RW character-word fusion strategy demonstrated superior performance, outperforming RoBERTa by 0.33% and 0.37%, and BERT by 0.91% and 0.52%, respectively. These results indicate that character-word fusion vectors are more effective than traditional single vectors. Additionally, the Figure shows that RoBERTa outperforms BERT on both datasets, further validating that RoBERTa, with its larger training data size and hidden layer dimension, is more suitable for Chinese MNER tasks.

5.2.3 Efficient global pointer

GP as a novel SOTA method capable of handling both nested and non-nested NER, demonstrates significant advantages over traditional sequence labeling models. As shown in Fig. 8, on the two datasets, Models 1, 2, 3 and 4 are based on sequence labeling, while Models 5, 6, and 7 adopt GP as the decoding strategy. The latter achieves higher F1 scores, indicating superior performance. Additionally, in Table 1, Models 13 and 14 also outperform other sequence labeling models on datasets like CCKS2017 and CCKS2019, further validating the advantages of the GP approach.

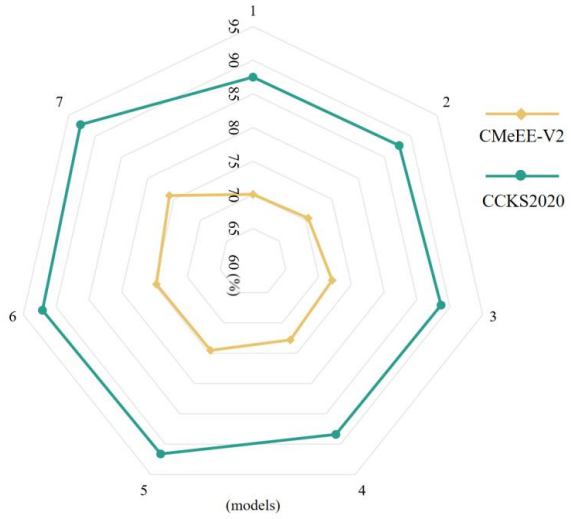

Figure 8: Comparison of F1 values for different datasets

This study further enhances GP by introducing a mask operation on the upper-right triangle of the entity matrix and setting a maximum sequence length (k) for entity

recognition. This optimization reduces computational resource consumption and minimizes the probability of identifying negative samples. As shown in Fig. 9, experiments were conducted to evaluate the impact of different k values on model performance. When k ranges from 0 to 17, the F1 score improves significantly as k increases. The model achieves optimal performance on the two datasets when k is set to 43 and 39, respectively. Compared to the original GP, the F1 scores improve by 0.16% and 0.23%, demonstrating the effectiveness of our proposed EGP design.

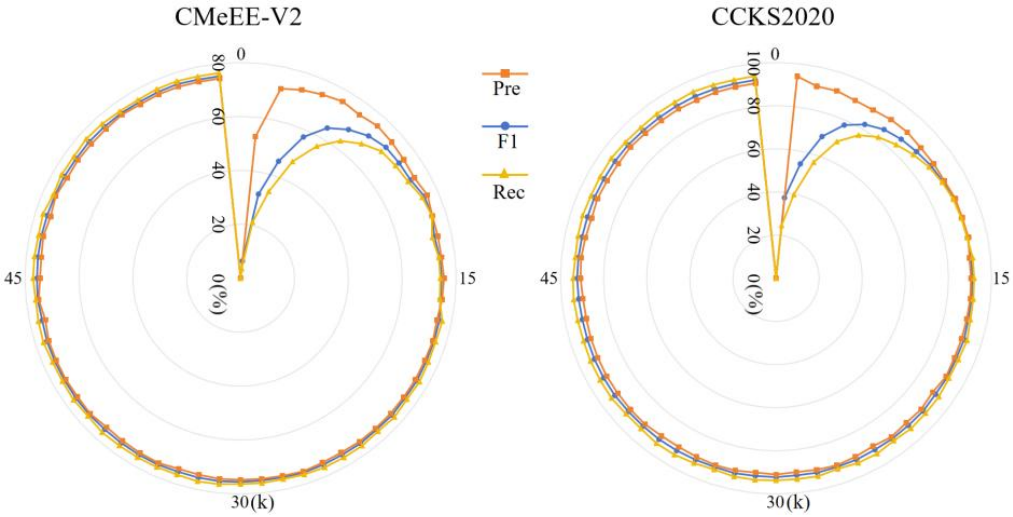

Figure 9: Comparison of the effect of different degrees of upper right triangular masking

5.2.4 Ablation control

In this section, we systematically validate the impact of individual components on the model's overall performance through controlled ablation experiments. As shown in Table 6, by sequentially removing DA, Character-Word Fusion Vectors, and the Upper-Right Triangle Mask, the model's F1 scores on both datasets exhibit varying degrees of decline, demonstrating the effectiveness of each component. Specifically:

The removal of Character-Word Fusion Vectors results in the most significant performance drop, indicating that the RW (RoBERTa with Word2Vec) encoding layer is crucial.

The removal of the Upper-Right Triangle Mask leads to the smallest performance decline, suggesting that the masking operation has a relatively minor impact but still leaves room for further optimization.

The removal of DA also causes a noticeable performance drop, particularly in Chinese MNER tasks, where DA significantly enhances the model's recognition capability and generalization performance.

In conclusion, these experiments not only validate the necessity of each component but also provide important guidance for further model optimization.

Table 6: Ablation study: Impact of individual components

Ablation Component	CMeEE-V2 (F1%)	CCKS2019 (F1%)
Full Model (DA-RW-EGP)	75.87	92.77
w/o Data Augmentation	75.21	92.31
w/o Character-Word Fusion	74.96	92.25
w/o Upper-Right Triangle Mask	75.71	92.54

5.3 Computational efficiency

Theoretically, the decoding time complexity of BiLSTM-CRF is $O(n)$, while that of GP (or EGP) is $O(n^2)$, where n is the length of the input text sequence. However, in practice, GP's loss function computation through matrix operations can be fully parallelized, resulting in an ideal time complexity of $O(1)$. In contrast, BiLSTM-CRF involves high-dimensional hidden layers, which significantly increase computational overhead. Additionally, it requires dynamic programming algorithms to decode the optimal label sequence and often needs post-processing steps to handle nested entities, further adding to the computational burden. Overall, GP demonstrates higher computational efficiency. Below, we evaluate the computational efficiency based on the training and validation time per epoch for different models on the datasets.

As shown in Table 7. On the CMeEE-V2 dataset, compared to the SOTA RoBERTa-BiLSTM-CRF model, the RW-GP model reduces training time by 14.1% and validation time by 16.4%. Further optimization with the upper-right triangle mask in the RW-EGP model achieves an additional 1.3% reduction in training time and 1.9% in validation time. Similar improvements are observed on the CCKS2020 dataset, where the RW-GP and RW-EGP models reduce training time by 9.7% and 11.4%,

respectively. Notably, the efficiency gains of the GP-based models are more pronounced for longer texts, as the Global Pointer mechanism avoids the quadratic complexity of traditional sequence labeling approaches. These results demonstrate that the proposed GP-based models significantly enhance computational efficiency across different datasets while maintaining competitive performance, making them more suitable for practical applications in Chinese MNER tasks.

Table 7: Training and validation time of different models (in seconds)

Model	CMeEE-V2 Training (s)	CMeEE-V2 Validation (s)	CCKS2020 Training (s)	CCKS2020 Validation (s)
RoBERTa-BiLSTM-CRF	873.76	42.71	94.44	5.15
RW-GP	750.87	35.71	85.32	4.23
RW-EGP	739.43	34.89	83.67	4.12

Note: Time is measured in seconds on a single GPU (NVIDIA RTX 4070Ti).

5.4 Failure cases and weaknesses

Although the proposed method achieves significant performance improvements on the experimental datasets, there are still some failure cases and limitations. For instance, in the case of the symptom entity "肺部炎症性病变" (inflammatory lesions in the lungs), the model might incorrectly classify it as "肺部炎症" (inflammation in the lungs), a disease entity, or "肺部" (lungs), a body part entity, due to insufficient learning. However, the main limitations of the model are as follows:

Limitations of Data Augmentation: While DA improves recognition performance for most entity types, it shows no significant improvement for certain types (e.g., "dru" and "ima"). We hypothesize that the high-quality annotations of these entity types make the generated pseudo-data introduce noise, which may interfere with the model's learning from the original high-quality data.

Computational Resource Consumption of Entity Matrix Scoring: When applying the upper-right triangle mask to GP during the decoding phase for entity matrix scoring, the reduction in computational resource consumption does not reach the expected 70%. This may be because the masking operation itself introduces additional computational overhead, especially when processing large-scale entity matrices. While the mask reduces the number of calculations, its design and implementation may incur extra costs that partially offset the efficiency gains.

6 Conclusion

This paper proposes a Chinese MNER method using pre-trained models and an EGP mechanism. The proposed method is evaluated for its efficiency and accuracy in processing complex Chinese medical text. To address the issues of imbalanced entity types and varying entity lengths in the original dataset, we integrate a data augmentation strategy combining EDA and remote supervision, generating high-quality pseudo-data to enrich the training set. The encoding layer combines the RoBERTa model with word vectors generated by Word2Vec, ensuring the input vectors contain rich contextual semantics and domain-specific medical word information. During decoding, the model employs the EGP mechanism, which constructs an entity matrix to identify entity start and end positions, utilizes an efficient decoding computation strategy, and narrows the search range for candidate entities. This approach enables the unified recognition of nested and flat entities while significantly improving recognition efficiency and accuracy.

Experimental results show that the proposed model achieves F1 scores of 75.87% on the CMeEE-V2 dataset and 92.77% on the CCKS2020 dataset, outperforming baseline models and demonstrating superior recognition performance. This work contributes novel optimization strategies and concepts to the field of MNER.

In the future, we plan to further explore the potential of GPLM for Chinese MNER, focusing on addressing the limitations identified in this study. First, we aim to develop DA techniques that generate higher-quality pseudo-data to avoid introducing noise into high-quality annotations. Second, we will optimize the design and implementation of the upper-right triangle mask for entity matrix scoring, aiming to reduce computational overhead and achieve the expected efficiency gains. Additionally, we will explore more sophisticated model architectures to further enhance recognition performance and extend our research to broader datasets and NLP tasks, contributing to the technological advancement of the field.

Acknowledgments

We would like to thank the funders of this project Xiamen University of Technology, Strong Digital Technology Co., Ltd., and all the teams and individuals who supported this research.

References

[1] W. Li, K. Zhang, T. Guan, H. Zhang, T. Zhu, B. Chang, et al. Overview of CHIP2020 shared Task1: Named entity recognition in Chinese medical text. Journal of Chinese Information Processing, 36(4): 66-72, 2022.

[2] J. Du, H. Yi, S. Feng. Research and development of named entity recognition in Chinese electronic medical record. Acta Electronica Sinica, 50(12): 3030-3053, 2022.

[3] W. Chen, P. Qiu, F. Cauteruccio. MedNER: A service-oriented framework for Chinese medical named-entity recognition with real-world application. Big Data and Cognitive Computing, 8(8): 86, 2024.

[4] C. Fócil-Arias, G. Sidorov, A. Gelbukh. Medical events extraction to analyze clinical records with conditional random fields. Journal of Intelligent & Fuzzy Systems, 36(5): 4633-4643, 2019.

[5] F. Shen, S. Liu, S. Fu, Y. Wang, S. Henry, O. Uzuner, et al. Family history extraction from synthetic clinical narratives using natural language processing: Overview and evaluation of a challenge data set and solutions for the 2019 national NLP clinical challenges (n2c2)/Open Health Natural Language Processing (OHNLP) Competition. JMIR Medical Informatics, 9(1): e24008, 2021.

[6] N. Zhang, M. Chen, Z. Bi, X. Liang, L. Li, X. Shang, et al. CBLUE: A Chinese biomedical language understanding evaluation benchmark. arxiv preprint arxiv:2106.08087, 2021.

[7] H. Wang, C. Wang. A Survey: Chinese medical named entity recognition. Journal of Shandong Normal University (Natural Science): 36(2): 109-117, 2021.

[8] S. Jiang, S. Zhao, K. Hou, Y. Liu, L. Zhang. A BERT-BiLSTM-CRF model for Chinese electronic medical records named entity recognition. In International Conference on Intelligent Computation Technology and Automation (ICICTA) (IEEE Press, Xiangtan, CN): 166-169, 2019.

[9] J. Yang, Y. Guan, B. He, C. Qu, Q. Yu, Y. Liu, et al. Corpus construction for named entities and entity relations on Chinese electronic medical records. Journal of Software, 27(11): 2725-2746, 2016.

[10] Q. Wang, Y. M. Zhou, T. Ruan, D. Gao, Y. Xia, P. He. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. Journal of Biomedical Informatics, 92: 103133, 2019.

[11] Z. Pan. Research on the recognition of Chinese named entity based on rules and statistics. Information Science, 30(5): 708-712, 2012.

[12] J. Jaćimović, C. Krstev, D. Jelovac. A rule-based system for automatic de-identification of medical narrative texts. Informatica, 39(1): 45-53, 2015.

[13] X. Chen, C. Ouyang, Y. Liu, Y. Bu. Improving the named entity recognition of Chinese electronic medical records by combining domain dictionary and rules. International Journal of Environmental Research and Public Health, 17(8): 2687, 2020.

[14] H. Wang, W. L. K. Yeung, Q. X. Ng, A. Tung, J. A. M. Tay, D. Ryanputra, et al. A weakly-supervised named entity recognition machine learning approach for emergency medical services clinical audit. International Journal of Environmental Research and Public Health, 18(15): 7776, 2021.

[15] A. Ekbal, S. Bandyopadhyay. Named entity recognition using appropriate unlabeled data, post-processing and voting. Informatica, 34(1): 55-76, 2010.

[16] S. Fine, Y. Singer, N. J. M. l. Tishby. The hierarchical hidden Markov model: Analysis and applications. Machine Learning, 32: 41-62, 1998.

[17] W. Chen, Y. Zhang, H. Isahara. Chinese named entity recognition with conditional random fields. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing (SIGHAN) (ACL Press, Sydney, AU): 118-121, 2006.

[18] Z. F. Ju, J. Wang, F. Zhu. Named entity recognition from biomedical text using SVM. In International Conference on Bioinformatics and Biomedical Engineering (ICBBE) (IEEE Press, Wuhan, CN), 1-4, 2011.

[19] B. T. Gao, Y. Zhang, B. Liu. BioTrHMM: named entity recognition algorithm based on transfer learning in biomedical texts. Application Research of Computers, 36(1): 45-48, 2019.

[20] K. Liu, Q. Hu, J. Liu, C. Xing. Named entity recognition in Chinese electronic medical records based on CRF. In Web Information Systems and Applications Conference (WISA) (IEEE Press, Liuzhou, CN): 105-110, 2017.

[21] T. Gui, R. Ma, Q. Zhang, L. Zhao, Y. G. Jiang, X. Huang. CNN-Based Chinese NER with Lexicon Rethinking. In International Joint Conference on Artificial Intelligence, (Macao, CN): 4982-4988, 2019.

[22] J. Chung, Ç. Gülçehre, K. Cho, Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arxiv preprint arxiv:1412.3555, 2014.

[23] R. Dey, F. M. Salem. Gate-variants of gated recurrent unit (GRU) neural networks. In International Midwest Symposium on Circuits and Systems (MWSCAS) (IEEE Press, Boston, USA), 1597-1600, 2017.

[24] K. Xu, Z. Yang, P. Kang, Q. Wang, W. Liu. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. Computers in biology and medicine, 108: 122-132, 2019.

[25] J. Devlin, M. W. Chang, K. Lee, K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arxiv preprint arxiv:1810.04805, 2018.

[26] X. Wang, C. Peng, Q. Li, Q. Yu, L. Li, P. Li, R. et al. A Chinese nested named entity recognition model for chicken disease based on multiple fine-grained feature fusion and efficient global pointer. Applied Sciences, 14(18): 8495, 2024.

[27] J. Zhong, Z. Xuan, K. Wang, Z. Cheng. A BERT-Span model for Chinese named entity recognition in rehabilitation medicine. PeerJ Computer Science, 9: e1535, 2023.

[28] S. K. Sienčnik. Adapting word2vec to named entity recognition. In Nordic Conference of Computational Linguistics (NODALIDA) (ACL Press, Vilnius, Lithuania): 239-243, 2015.

[29] J. Pennington, R. Socher, C. D. Manning. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP) (ACL Press, Doha, QA), 1532-1543, 2014.

[30] P. Chen, M. Zhang, X. Yu, S. Li. Named entity recognition of Chinese electronic medical records based on a hybrid neural network and medical MC-BERT. BMC Medical Informatics and Decision Making, 22(1): 315-327, 2022.

[31] Z. Sun, X. Li, X. Sun, Y. Meng, X. Ao, Q. He, et al. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. arXiv preprint arXiv:2106.16038, 2021.

[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al. Roberta: A robustly optimized bert pretraining approach. arxiv preprint arxiv:1907.11692, 2019.

[33] M. Gridach. Character-level neural network for biomedical named entity recognition. Journal of Biomedical Informatics, 70: 85-91, 2017.

[34] Y. Kimura, T. Komamizu, K. Hatano. An automatic labeling method for Subword-Phrase recognition in effective text classification. Informatica, 47(3): 315-326, 2023.

[35] X. Wang, Y. Zhang, X. Ren, Y. H. Zhang, M. Zitnik, J. B. Shang, et al. Cross-type biomedical named entity recognition with deep multi-task learning. Bioinformatics, 35(10): 1745-1752, 2019.

[36] B. Ji, R. Liu, S. S. Li, J. Yu, Q. B. Wu, Y. S. Tan, et al. A hybrid approach for named entity recognition in Chinese electronic medical record. BMC Medical Informatics and Decision Making, 19, 149-158, 2019.

[37] Q. Qin, S. Zhao, C. Liu. A BERT-BiGRU-CRF model for entity recognition of Chinese electronic medical records. Complexity, 2021(1): 6631837, 2021.

[38] J. Su, A. Murtadha, S. Pan, J. Hou, J. Sun, W. Huang, et al. Global pointer: Novel efficient span-based approach for named entity recognition. arxiv preprint arxiv:2208.03054, 2022.

[39] P. Zhang, W. Liang. Medical name entity recognition based on lexical enhancement and global pointer. International Journal of Advanced Computer Science Applications, 14(3): 2023.

[40] J. Wei, K. Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arxiv preprint arxiv:1901.11196, 2019.

[41] T. Yan, X. Zhang, Z. Luo. LTACL: Long-tail awareness contrastive learning for distantly supervised relation extraction. Complex & Intelligent Systems, 10(1): 1551-1563, 2023.

[42] W. Zhang, J. Jiang. Bootstrap-Based resampling methods for software reliability measurement under small sample condition. Journal of Circuits, Systems and Computers, 33(9): 2450161, 2023.

[43] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, et al. Ernie: Enhanced representation through knowledge integration. arxiv preprint arxiv:1904.09223, 2019.

[44] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang. Pre-training with whole word masking for Chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3504-3514, 2021.

[45] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568, 127063, 2024.

[46] K. Yan. Optimizing an English text reading recommendation model by integrating collaborative filtering algorithm and FastText classification method. Heliyon, 10(9): e30413, 2024.

[47] Z. Zhao, Y. Dong, J. Liu, J. Zhang and H. Cao. Medical named entity recognition incorporating word information and graph attention. Computer Engineering and Applications, 60(11): 147-155, 2014.

[48] W. K. Li, W. Li and Y. F. Wu. Combination methods of Chinese character and word embeddings in deep learning. Journal of Chinese Information Processing, 31(06): 140-146, 2017.

[49] J. Su, M. Zhu, A Murtadha, S. Pan, B. Wen and Y. Liu. Zlpr: A novel loss for multi-label classification. arxiv preprint arxiv:2208.02955, 2022.