MRC-based Nested Medical NER with Co-prediction and Adaptive Pre-training

Xiaojing Du[1], **Hanjie Zhao**[1]*, **Danyan Xing**[2], **Yuxiang Jia**[1]†, **Hongying Zan**[1]
[1]School of Computer and Artificial Intelligence, Zhengzhou University
[2]School of Management, Zhengzhou University
{zzu_dxj,hjzhao_zzu,xdy_0423}@163.com
{ieyxjia,iehyzan}@zzu.edu.cn

Abstract

In medical information extraction, medical Named Entity Recognition (NER) is indispensable, playing a crucial role in developing medical knowledge graphs, enhancing medical question-answering systems, and analyzing electronic medical records. The challenge in medical NER arises from the complex nested structures and sophisticated medical terminologies, distinguishing it from its counterparts in traditional domains. In response to these complexities, we propose a medical NER model based on Machine Reading Comprehension (MRC), which uses a task-adaptive pre-training strategy to improve the model's capability in the medical field. Meanwhile, our model introduces multiple word-pair embeddings and multi-granularity dilated convolution to enhance the model's representation ability and uses a combined predictor of Biaffine and MLP to improve the model's recognition performance. Experimental evaluations conducted on the CMeEE, a benchmark for Chinese nested medical NER, demonstrate that our proposed model outperforms the compared state-of-the-art (SOTA) models.

Keywords: Medical NER, MRC, Co-prediction, Adaptive pre-training, Multi-granularity dilated convolution

1. Introduction

With the rapid advancement of medical digitalization, an abundance of medical documentation is being generated, encompassing electronic medical records, medical reports, and various other forms. The extraction of medical information, notably medical named entity recognition (NER), garners increasing significance in applications such as knowledge graph construction, question-answering systems, and automated analysis of electronic medical records. Medical NER aims to automatically identify medical entities, including but not limited to body (bod), disease (dis), clinical symptom (sym), medical procedure (pro), medical equipment (equ), drug (dru), and medical examination item (ite), from medical texts.

These entities often exhibit lengthy, nested structured, and polysemous, thus presenting considerable challenges to the task of medical NER. For example, as illustrated in Figure 1, the three entities "迷走神经" (vagus nerve), "舌咽神经核" (glossopharyngeal nucleus) and "舌下神经核" (hypoglossal nucleus), denoted as "bod", are nested within the entity "迷走神经、舌咽神经核及舌下神经核受损伤"(the injury of vagus nerve, glossopharyngeal nucleus and hypoglossal nucleus), denoted as "sym".

To address the challenge of nested NER, we adopt a strategy similar to Li et al. (2020b) and Du et al. (2022), by framing NER as a machine reading comprehension (MRC) task. Like Li et al. (2022), we employ an approach that combines the strengths of both Biaffine and Multi-Layer Perceptron (MLP) predictors through joint prediction. Additionally, we introduce a task-adaptive pre-training strategy to fine-tune the original pre-trained model specifically for medical NER. Our model incorporates several techniques, including Conditional Layer Normalization (CLN), weighted layer fusion, word-pair embeddings, and multi-granularity dilated convolution, all of which have been demonstrated to improve performance. The contributions of this paper can be summarized as follows:

- We introduce a nested medical NER model based on MRC, featuring the integration of Biaffine and MLP for joint prediction. Additionally, we introduce several enhancements, including multiple word-pair embeddings and multi-granularity dilated convolution, to improve the model's performance.
- We take a task adaptive pre-training strategy to optimize the pre-trained model for medical domain. Entity type embedding is fed into the conditional layer normalization to more effectively utilize entity type information.
- Experimental results on the nested Chinese medical NER corpus CMeEE demonstrate superior performance of our model over existing state-of-the-art (SOTA) models.

*Xiaojing Du and Hanjie Zhao contributed equally to this research.

†Yuxiang Jia is the corresponding author.

LREC-COLING 2024, pages 11669–11679

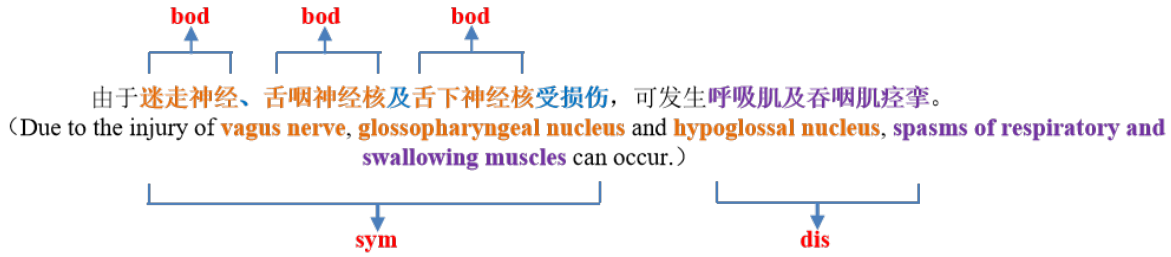

Figure 1: An example of nested entities.

2. Related Work

As the medical domain has a high level of complexity and variability among entities, the task of medical NER poses a significant challenge compared to general NER. Therefore, employing pre-trained models such as BERT (Li et al., 2020a; Qin et al., 2021) and ELMo (Li et al., 2020c; Wan et al., 2020) has become a common practice to encode input texts in the medical domain.

To effectively leverage the information in both characters and words, Ji et al. (2019) devise a method that involves constructing a drug dictionary and implementing post-processing rules to modify the entities. Furthermore, taking into account that radicals, strokes, and glyphs can offer valuable supplementary information alongside words, Zhou et al. (2021) propose a BiLSTM-CRF model that operates at the stroke and radical levels to capture the semantic nuances of Chinese characters more comprehensively. Yang et al. (2022) introduce the TSERL model, which establishes a relationship graph between radicals, characters, and words to enhance NER performance. Similarly, domain-specific data can be harnessed to augment the performance of medical NER systems. For instance, Liu et al. (2021) develop Med-BERT by pre-training it on a corpus of medical texts, resulting in significant performance enhancements. Additionally, Chen et al. (2020) propose a model that integrates domain dictionaries and rules with BiLSTM-CRF, showing improved performance in medical NER. This underscores the potential benefits of integrating domain-specific information and rules into existing NER systems.

To address the complexities inherent in nested NER and to incorporate knowledge from entity types, NER has been formulated as an MRC task (Li et al., 2020b). Expanding on this, Liu et al. (2023) adeptly integrate the interrelations among entity labels utilizing Graph Attention Networks (GATs), thereby fusing label information with textual content to refine NER strategies. To enhance the exchange of information between the initial and terminal segments of the entity, Cao et al. (2021) innovatively apply a Biaffine mechanism to MRC. Further advancements are noted by Zhu et al. (2021), who synergize sequence labeling and span boundary detection techniques through the implementation of voting strategies. Similarly, Zheng et al. (2021) achieve a harmonious ensemble of Conditional Random Fields (CRF) and MRC, showcasing the evolving landscape of NER methodologies.

Multi-task learning represents an alternative approach to enhancing performance. In this framework, the NER model benefits from parameter sharing with models designed for other tasks. Chowdhury et al. (2018) explore this approach by considering NER and POS tagging as two concurrent tasks. Additionally, Du et al. (2022) propose an integration of the MRC-CRF model for sequence labeling and the MRC-Biaffine model for span boundary detection within a multi-task learning architecture. Similarly, Luo et al. (2020) extend the multi-task learning paradigm to NER across two distinct datasets.

The advent of large language models (LLMs), exemplified by ChatGPT (OpenAI, 2022), has heralded a new paradigm in entity recognition and an increasing number of studies have been focusing on expansive medical models. Notably, ChatDoctor (Li et al., 2023) enhances its capabilities through continuous pre-training specifically within the medical domain. Likewise, MedAlpaca (Han et al., 2023) garners positive reviews from experts for its clinical response quality. In the domain of Chinese medical research, DoctorGLM (Xiong et al., 2023) leverages data processed by ChatGPT for its training, whereas BenTsao (Wang et al., 2023) employs fine-tuning with Q&A data generated by ChatGPT, sourced from CMeKG (Byambasuren et al., 2019). Furthermore, HuatuoGPT (Zhang et al., 2023) improves its performance through a combination of Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF). Additionally, Zhongjing (Yang et al., 2023) enhances its capabilities in Chinese medical consultations by incorporating feedback from medical experts and multi-turn medical dialogues from the real world.

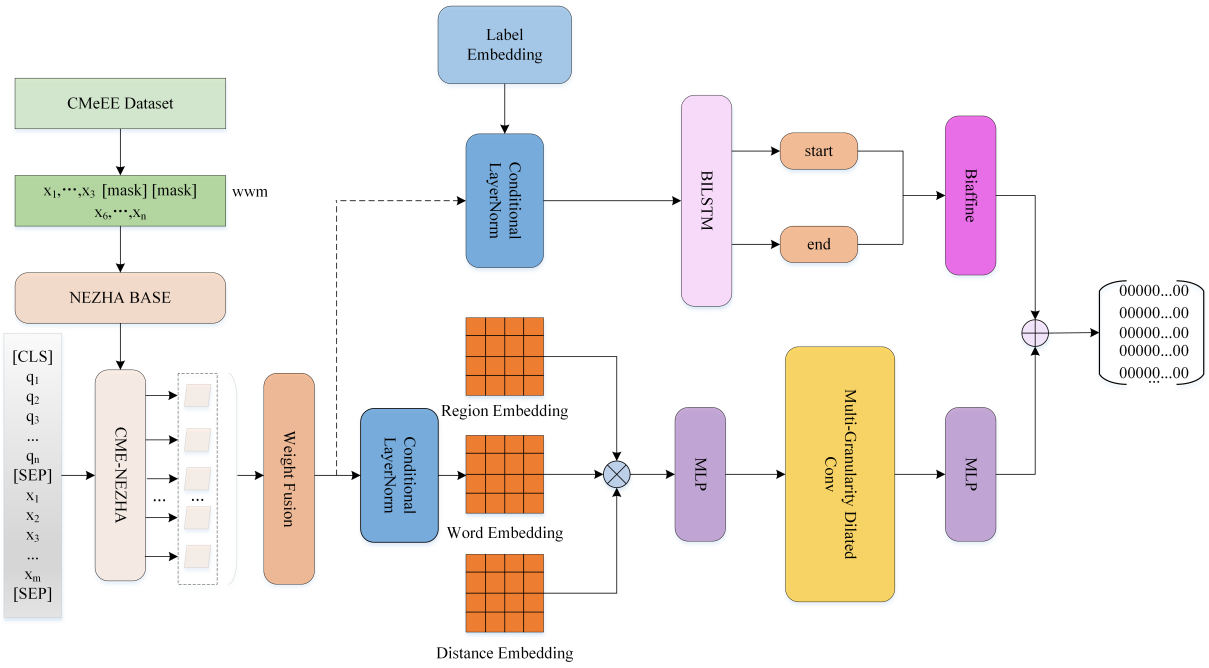

Figure 2: The architecture of the proposed NER model. The acronym 'wwm' stands for the BERT-Whole Word Masking model.

3. The MRC-CAP Model

MRC model extracts answer fragments from paragraphs by a given question. Suppose X is the input text, for each entity type y, designing a query q_y, and we can get the triple (q_y, y, X), which is exactly the $(question, answer, context)$ an MRC model needs. The model only calculates the loss of context during training, and masks the loss of query and padding.

The overall architecture of the proposed MRC-CAP (MRC with Co-prediction and Adaptive Pre-training) model is shown in Figure 2. Firstly, incremental training is performed on the existing pre-trained model NEZHA (Wei et al., 2019) through task adaptive pre-training to obtain the task pre-trained model CME-NEZHA which is more suitable for Chinese medical NER. For each entity type, the input of the model is the concatenation of context and entity description query which we will elaborate on in section 5. The input is encoded by CME-NEZHA, and then the 12 hidden layers are fused by weights.

Biaffine and MLP are used for joint prediction to enhance the decoding results. For the Biaffine predictor, conditional layer normalization with entity type embedding is used to further leverage entity type knowledge. For the MLP predictor, conditional layer normalization is used to generate word-pair grid representation, combined with other two word-pair embeddings, distance embedding and region embedding, going through a multi-granularity dilated convolution layer for capturing information exchange between distant and close words.

As for the Biaffine prediction branch, the hidden layer sequence after conditional layer normalization goes through a BiLSTM and two nonlinear activation functions to learn the representation of the start and end of the span respectively. Finally, the score of word pair (x_i, x_j) is calculated by a Biaffine classifier as follows,

$$x_i = MLP_{\text{start}}\ (h_i) \tag{1}$$

$$x_j = MLP_{\text{end}}\ (h_j) \tag{2}$$

$$y'_{ij} = x_i^T U x_j + W\ (x_i \oplus x_j) + b \tag{3}$$

where U is a tensor of $N * C * N$, W is a matrix of $2N * C$, b is a bias vector, N is the length of the sentence, and C is the number of entity categories +1 (non-entity).

As described in (Li et al., 2022), the MLP prediction branch incorporating three word-pair embeddings, including the tensor V of $N * N * d_h$ representing word information, the tensor E^d of $N * N * d_{E_d}$ representing the distance between each pair of words, and the tensor E^t of $N * N * d_{E_t}$ representing the triangle region information in the word-pair grid. The concatenation of three embeddings through an MLP is fed into a multi-granularity dilated convolution layer with different dilation rate l to capture the interactions between the words of different distances. The computation of one dilated convolution is formulated as:

Table 1: Statistics of entities in CMeEE V1 and V2.

Entity	CMeEE V1			CMeEE V2		
	#Entity	Per/%	Avg.len	#Entity	Per/%	Avg.len
bod	23580	28.72	3.38	31467	28.94	3.36
dis	20778	25.31	5.34	25699	23.64	5.36
sym	16399	19.98	**6.70**	22415	20.62	**7.42**
pro	8389	10.22	5.21	13007	11.96	5.86
dru	5370	6.54	4.68	5945	5.47	4.78
ite	3504	4.27	4.29	5749	5.28	4.97
mic	2492	3.04	4.26	2964	2.73	4.27
equ	1126	1.37	4.39	1053	0.96	4.73
dep	458	0.55	2.88	431	0.40	2.55
Total	82096	100	4.89	108730	100	5.17

$$Q^l = \sigma(\mathrm{DConv_l}(\mathrm{MLP}\left(\left[V; E^d; E^t\right]\right))) \quad (4)$$

where Q^l of $N * N * d_g$ denotes the output of the dilation convolution of dilation rate l, and σ is the GELU (Hendrycks and Gimpel, 2016) activation function. With the dilation rate 1, 2 and 3, the final word-pair grid representation is $Q = [Q^1, Q^2, Q^3]$ of $N * N * 3d_g$. Then an MLP is used to calculate score for word pair (x_i, x_j):

$$y''_{ij} = MLP\left(Q_{ij}\right) \quad (5)$$

With the combination of Biaffine and MLP, we get the co-prediction word pair score y_{ij} :

$$y_{ij} = \mathrm{Soft\,max}\left(y'_{ij} + y''_{ij}\right) \quad (6)$$

In the training stage, we optimize the following cross-entropy loss function:

$$L_{\text{Biaffine }+MLP} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{c=1}^{C}\hat{y}_{ij}^{c}\log y_{ij}^{c} \quad (7)$$

4. Datasets

The experiment uses the CMeEE V1 and V2 datasets for Chinese nested medical NER. CMeEE V2 is revised from CMeEE V1 by correcting some annotations. The texts in CMeEE are extracted from clinical pediatrics textbooks, comprising nine types of entities, including body (bod), disease (dis), clinical symptom (sym), medical procedure (pro), medical equipment (equ), drug (dru), medical examination item (ite), department (dep), and microorganism (mic).

After analyzing the entity distribution of the two datasets, we observe a similar entity distribution across both datasets with an unbalanced distribution among various types. "bod", "dis", and "sym" are the dominant types, followed by "pro". Table 1 gives a more detailed comparison of the changes in each entity type and between the two datasets.

Moreover, an analysis is conducted on the proportion of nested versus non-nested entities within the datasets. The comparative data in Table 2 reveals that CMeEE V2 enhances the annotation of nested entities threefold compared to V1, thereby enriching the dataset with more comprehensive nested information.

Table 2: Statistics of nested entities in CMeEE V1 and V2.

Entity	CMeEE V1	CMeEE V2
#Flat	73336	74160
#Nested	8760	34570
Nested/%	10.67	31.79
#Nested in sym	3808	14908
Nested in sym/%	23.22	66.51

Take the statement, “胸部X线透视和胸片可见患侧膈呼吸运动减弱肋膈角变钝。” (The chest X-ray and chest films reveal diminished diaphragmatic respiratory motion on the affected side and a blunted costophrenic angle.), as an example. In V1, the "sym" entity labeled is “患侧膈呼吸运动减弱” (diminished diaphragmatic respiratory motion on the affected side). In contrast, V2 annotates the outer "sym" entity “胸部X线透视和胸片可见患侧膈呼吸运动减弱” (The chest X-ray and chest films reveal diminished diaphragmatic respiratory motion on the affected side); similarly, while V1 identifies “肋膈角变钝” (blunted costophrenic angle) as a "sym" entity, V2 enhances this with the inner "bod" entity “肋膈角” (costophrenic angle). Overall, V2 enriches the dataset with a higher num-

ber of nested entities through the inclusion of both internally and externally nested entity annotations, thereby offering a more intricate entity structure for the task of medical nested NER.

In examining entities by average length, the entity type "sym" is found to be the longest, with lengths of 6.70 in V1 and 7.42 in V2 respectively. Notably, the nesting rate for "sym" entities, about 2/3, is double that of the dataset overall. Further examination discloses that "sym" entities with the greatest lengths exhibit intricate structural complexities, typically encompassing entities of other types.

Table 3: Statistics of entities nested inside "sym".

Entity	CMeEE V1		CMeEE V2	
	#Nested	Per/%	#Nested	Per/%
bod	4114	85.01	23720	78.30
ite	405	8.37	3856	12.73
dis	202	4.17	846	2.79
pro	56	1.16	1268	4.19
dru	27	0.56	132	0.44
mic	23	0.48	340	1.12
equ	12	0.25	128	0.42
dep	0	0.00	2	0.01
Total	4839	100	30292	100

Table 3 presents an analysis of entity types nested within clinical symptom (sym), revealing a widespread occurrence of entity nesting across almost all categories. Notably, the "bod" entities, which are the most prevalent in the dataset, as indicated in Table 1, predominate among those nested within "sym". The distribution of other nested entity types closely mirrors the overall entity distribution observed in Table 1. Exceptionally, the medical examination item (ite) category demonstrates a significant increase in its proportion of nesting within "sym", securing the second position in the ranking. As an illustration, within the entity of type "sym", which states “小脑延髓池的压力常呈负压” (The pressure in the fourth ventricle of the cerebellum is usually negative), there is a nested entity of type "ite", specifically referring to “小脑延髓池的压力” (The pressure in the fourth ventricle of the cerebellum).

5. Experiments

5.1. Experimental Settings

Below, we provide a detailed description of the experimental setup.

Query generation For this study, we incorporate query statements from (Du et al., 2022) and pair them with examples of relevant entity types such as “细胞” (cells), “皮肤” (skin) and “抗体” (antibodies), which are used as queries to represent the "bod" entity type. This is a strategic choice to ensure the accuracy of the query statements in conveying the intended entity types, which in turn provides data with greater informational value for model training. In application, these query statements are used to enrich the model’s understanding of entity types and enhance its overall performance in the NER task.

Parameter settings To begin with, we perform pre-training on the NEZHA model using the CMeEE dataset, training for 100 epochs to obtain a base model. Then, we fine-tune the base model specifically for NER. In the experiments, we set the batch size to 16, the regularization parameter dropout to 0.1, the learning rate for NEZHA to 2e-5, and the other learning rate parameters to 2.5e-3. The maximum text length for the model is set to 200. All experiments are performed on a single NVIDIA RTX3090.

Evaluation metrics For evaluation, we employ precise, recall, and F1 scores as performance metrics. In particular, we adopt Micro F1 as a comprehensive metric reflecting the overall recognition performance of the model. Micro F1, derived from the mean of precision and recall, effectively combines the performance across all categories by weighting each type’s sample count.

5.2. Comparison with Previous Models

5.2.1. Baseline Models

All the baselines we use are as follows: **(1)** Lattice-LSTM, Lattice-LSTM+Med-BERT, FLAT-Lattice and Medical-NER are from Liu et al. (2021). Lattice-LSTM, Lattice-LSTM+Med-BERT and FLAT-Lattice incorporate lexicon to decide entity boundary. Medical NER introduces a big dictionary and pre-trained domain model. **(2)** LEAR (Yang et al., 2021) independently encodes text and label descriptions and then integrates label knowledge into the text representation through a semantic fusion module. **(3)** MacBERT-large and Human are from Zhang et al. (2022), which is a variant of BERT. Human denotes the annotating result of human. **(4)** BERT-CRF, BERT-Biaffine and RICON are from Gu et al. (2022). BERT-CRF solves sequence labeling with CRF, BERT-Biaffine detects span boundary with Biaffine, and RICON learns regularity inside entities. **(5)** TsERL (Yang et al., 2022) constructs a relationship graph between radicals, characters, and words. **(6)** W2NER (Li et al., 2022) proposes a unified word relation classification model for different NER problems. **(7)** MRC-MTL (Du et al., 2022) integrates MRC-CRF model for sequence labeling and MRC-Biaffine model for span boundary detection into the multi-task learning (MTL) architecture. **(8)**

Table 4: Query for different entity types in CMeEE (Du et al., 2022).

Entity	Query
bod	在文本中找出身体部位，例如细胞、皮肤、抗体 Find body parts in the text, for example, cells, skin and antibodies
dep	在文本中找出科室，例如科、室 Find departments in the text, for example, department and room
dis	在文本中找出疾病，例如癌症、病变、炎症、增生、肿瘤 Find diseases in the text, for example, cancer and pathological changes
dru	在文本中找出药物，例如胶囊、疫苗、剂 Find drugs in the text, for example, capsule, vaccine and agent
equ	在文本中找出医疗设备，例如装置、器、导管 Find medical equipments in the text, for example, device and conduit
ite	在文本中找出医学检验项目，例如尿常规、血常规 Find medical examination items in the text, for example, urine routine and blood routine
mic	在文本中找出微生物，例如病毒、病原体、抗原、核糖 Find micro-organisms in the text, for example, virus and pathogen
pro	在文本中找出医疗程序，例如心电图、病理切片、检测 Find medical procedure in the text, for example, electrocardiogram and pathological section
sym	在文本中找出临床表现，例如疼痛、痉挛、异常 Find clinical manifestations in the text, for example, pain and spasm

FLR-MRC (Liu et al., 2023) fuses label information with text for NER. **(9)** FFBLEG (Cong et al., 2023) is based on feature fusion and a bidirectional lattice embedding graph. **(10)** ChatGPT (OpenAI, 2022) and GPT-4 (Achiam et al., 2023). We leverage the API provided by OpenAI [1], opting for the GPT-3.5-turbo-16k and GPT-4 model as our baseline. We conduct a zero-shot experiment and only set the task definition and output format in the prompt template. Notably, the annotation guidelines specific to the CMeEE dataset are provided.

5.2.2. Main Results

Upon conducting a thorough analysis of our model's performance on the CMeEE V1 dataset in comparison with prior models, we identify substantial advancements in terms of precision, recall, and F1 scores, as highlighted in Table 5 through bold text. When compared with the experimental results of large language models, our model still demonstrates certain advantages.

The experimental results for the CMeEE V2 dataset (refer to Table 6) are even more remarkable. The performance of our model on this dataset surpasses the achievements on V1, with precision increasing from 67.35% to 77.20%, and the overall F1 score also achieves significant growth, reaching 77.04%. This progress fully reflects the adaptability of our model, especially in dealing with more complex data. However, it should be noted that due to the scarcity of research related to CMeEE V2, the comparison results for this part are not listed in Table 5.

The experiments conducted affirm the efficacy of our proposed model in enhancing the task of medical nested NER. Beyond merely outperforming existing NER models, including ChatGPT, our work validates the effectiveness of the methodologies employed.

5.3. Ablation Study

To assess the efficacy of the used modules, our study conducts ablation experiments by omitting the utilized modules within the model across two datasets. These experiments demonstrate the effectiveness of our selected modules in enhancing medical NER, with a comprehensive display of improvement in Table 6.

Initially, the omission of the adaptive pre-training (AP) module, leds to a decrease in model performance across both datasets. We infer that the AP module is instrumental in acquainting the model

[1]The results of ChatGPT are obtained during February and March 2024 with official API.

Table 5: Comparison with previous models on CMeEE V1.

Model	Pre.	Rec.	F1
Lattice-LSTM (Liu et al., 2021)	57.10	43.60	49.44
Lattice-LSTM+Med-BERT (Liu et al., 2021)	56.84	47.58	51.80
FLAT-Lattice (Liu et al., 2021)	66.90	70.10	68.46
Medical NER (Liu et al., 2021)	66.41	70.73	68.50
LEAR (Yang et al., 2021)	65.78	65.81	65.79
MacBERT-large (Zhang et al., 2022)	-	-	62.40
Human (Zhang et al., 2022)	-	-	67.00
BERT-CRF (Gu et al., 2022)	58.34	64.08	61.07
BERT-Biaffine (Gu et al., 2022)	64.17	61.29	62.29
RICON (Gu et al., 2022)	66.25	64.89	65.57
TsERL (Yang et al., 2022)	61.82	64.78	63.27
W2NER (Li et al., 2022)	66.05	69.07	67.53
MRC-MTL (Du et al., 2022)	66.28	70.34	68.25
FLR-MRC (Liu et al., 2023)	66.79	66.25	66.52
FFBLEG (Cong et al., 2023)	64.70	64.92	64.81
ChatGPT (OpenAI, 2022)	42.02	32.40	36.59
GPT-4 (Achiam et al., 2023)	39.21	50.81	44.26
MRC-CAP (Ours)	**67.35**	**71.62**	**69.42**

Table 6: Ablation experiments on CMeEE V1 and V2.

Model	CMeEE V1/%			CMeEE V2/%		
	Pre.	Rec.	F1	Pre.	Rec.	F1
MRC-CAP	67.35	71.62	69.42	77.20	76.88	77.04
-AP	67.89	69.54	68.70	76.97	76.02	76.49
-(AP+MLP)	70.71	64.09	67.24	75.65	74.91	75.28
-(AP+Biaffine)	67.64	68.68	68.16	75.08	76.42	75.74
-(AP+Biaffine+MLP)	67.98	65.87	**66.91**	75.39	73.76	**74.56**
-(AP+DConv)	69.75	65.76	67.69	76.79	75.07	75.92
-(AP+Region Emb)	68.56	67.14	67.84	76.01	76.34	76.17
-(AP+Distance Emb)	67.99	67.28	67.63	76.44	75.59	76.01

with an extensive range of medical data, which amplifies its capability in medical NER. Compared with the single predictors, the joint predictor is observed to enhance the model's recognition ability. Notably, the improvement attributed to the MLP is more pronounced compared to Biaffine. The exclusion of the joint predictor results in a significant decline in accuracy, 2.51% on CMeEE V1 and 2.48% on V2, which we particularly highlight in bold text.

Furthermore, the removal of the multi-granularity dilated convolution (DConv) witnesses a steeper decline on V2, attributed to the dataset's abundance of lengthy and intricate nested entities. The experimental outcomes also indicate a decline upon the exclusion of region embeddings or distance embeddings, validating that the integration of embeddings facilitates the learning of more suitable vector representations, thereby elevating the model's performance.

5.4. Experiments on Various Entity Types

To analyze the model's recognition performance on different types of medical entities, we conduct experiments on CMeEE V1. Table 7 illustrates that the entity type "dru" exhibits the highest recognition performance, with an impressive F1 score of 80.50%. This suggests that most medical drugs are standardized terms with high recognizability. Conversely, the entity type "ite" demonstrates the lowest recognition accuracy at 46.49%, possibly due to limited data and the majority of "ite" entities being nested within "sym" entities.

We devise a confusion matrix to evaluate the model's ability to distinguish between different

Table 7: Results of different types of NEs on CMeEE V1.

Entity	Pre.	Rec.	F1
bod	67.19	65.51	66.34
dis	79.75	77.93	78.83
dru	76.77	84.60	**80.50**
dep	65.96	86.11	74.70
equ	78.33	77.44	77.88
ite	53.30	41.23	46.49
mic	81.42	78.18	79.77
pro	64.41	67.16	65.76
sym	66.85	49.22	56.70
Mac-Avg	70.44	69.71	70.07

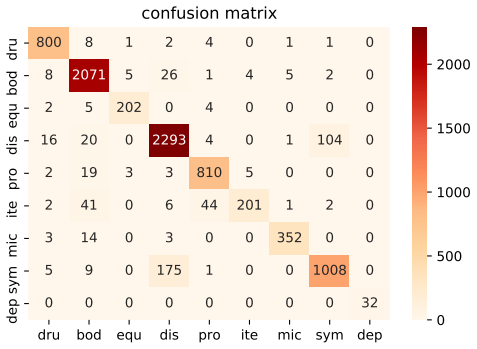

Figure 3: Confusion Matrix of NER on CMeEE V1.

types of entities. As illustrated in Figure 3, the horizontal axis denotes the correct entity types, referred to as ground truth, while the vertical axis represents the entity types identified by the model. The confusion matrix reveals that the model tends to misclassify "dis" as "sym" most frequently, with 175 out of 215 mistaken instances, and similarly, misclassifying "sym" as "dis" is also a prevalent error (104 out of 109 misclassified "sym" are erroneously labeled as "dis").

A quantitative analysis of the confusion matrix indicates that the model struggles particularly with recognizing complex and lengthy entities, such as "sym" and "dis", which rank among the top two in terms of average length. This difficulty can be attributed, at least in part, to the intricate structures inherent to these entity types.

5.5. Results of Nested and Flat NER

Experimental results for nested entity and flat entity recognition on the CMeEE V1 are presented in Table 8. It is evident from the table that the MRC-CAP model achieves higher entity recognition accuracy for both nested inner entities and flat entities compared to traditional MRC-based approaches.

Table 8: NER on CMeEE V1(Accuracy/%). "Inner" and "Outer" denote nested inner entities and nested outer entities, respectively.

Named Entity	MRC	MRC-CAP
All	65.87	70.43
Flat	67.54	71.75
Nested	52.03	59.48
Inner	46.43	63.08
Outer	56.64	55.10

Specifically, the MRC-CAP model demonstrates a 16.65% and 4.21% increase in entity recognition accuracy for nested inner entities and flat entities respectively, suggesting that enhancing embedded representations enables the model to better capture internal entity information and inter-entity relationships. However, the accuracy of nested entity recognition remains considerably lower than that of flat entity recognition, highlighting the need for further advancements in handling nested entity recognition challenges.

5.6. Case Study

We select two instances, which are wrongly recognized by MRC but correctly recognized by MRC-CAP. Detailed recognition results are provided in Table 9. In the first instance, the basic MRC model fails to identify the long entity “结核菌素皮试阳性” (tuberculin skin test positive) of type "sym", and erroneously recognizes the two nested entities within it. Conversely, MRC-CAP successfully identifies the long "sym" entity and accurately discerns the two inner-nested entities, the "mic" entity “结核菌素” (tuberculin) and the "pro" entity “皮试” (skin test). This suggests that our proposed model has enhanced the ability of MRC to recognize lengthy entities to a certain degree and provides assistance in identifying entities with nested structures.

In the second instance, while the traditional MRC model accurately identifies the boundaries of the entities “慢性排异” (chronic rejection) and “高血压” (hypertension), it misclassifies "sym" as "dis". Conversely, MRC-CAP correctly identifies the two "sym" entities. Although we acknowledge in section 5.4 that MRC-CAP occasionally misclassifies these two entity types, there is no denying that our proposed model, utilizing a joint predictor, has significantly improved in predicting entity categories compared to traditional MRC models.

6. Conclusion

This paper proposes an MRC-based medical NER model for both flat and nested NEs with Biaffine and

Table 9: Two cases

Case1	结核菌素皮试阳性结核的高危人群，应予以治疗。 High risk populations with positive skin test results for tuberculosis should be treated.
Golden Entity	[结核菌素]mic、[皮试]pro、[结核菌素皮试阳性]sym、[结核]dis
MRC	[结核菌素皮试]pro阳性[结核]dis的高危人群，应予以治疗。
MRC-CAP	{[结核菌素]mic[皮试]pro阳性}sym[结核]dis的高危人群，应予以治疗。
Case2	患儿情况好，只1例发生慢性排异及高血压。 The condition of the child is good, and only one develops chronic rejection and hypertension.
Golden Entity	[慢性排异]sym、[高血压]sym
MRC	患儿情况好，只1例发生[慢性排异]dis及[高血压]dis。
MRC-CAP	患儿情况好，只1例发生[慢性排异]sym及[高血压]sym。

MLP for joint prediction of NE span, introducing multiple word-pair embeddings and multi-granularity dilated convolution. To improve domain adaptation of the pre-trained model, we incrementally retrain it with a task-adaptive pre-training strategy. In addition, entity type embedding, conditional layer normalization, weighted layer fusion and other techniques are employed and show effectiveness. Experiments on the nested Chinese medical NER benchmark CMeEE V1 and V2 show that the proposed model outperforms comparative SOTA models. In the future, we will incorporate more domain knowledge to improve the performance of the medical NER model and explore potential of LLMs on medical NER task.

7. Ethics Statement

There are no ethics-related issues in this paper. We conduct experiments on publicly available datasets. These datasets do not share personal information and do not contain sensitive content that can be harmful to any individual or community.

8. Acknowledgments

The authors thank the anonymous reviewers for their insightful comments. This work is mainly supported by the Key Program of the National Natural Science Foundation of China (NSFC) (Grant No.U23A20316), the Key R&D Project of Hubei Province (Grant No.2021BAA029), and the Major Science and Technology Project of Yunnan Province (Grant No.202102AA100021). The authors are grateful to Zhengzhou Zoneyet Technology Co., Ltd. and Kunming Children's Hospital for their support and cooperation.

9. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Odma Byambasuren, Yunfei Yang, Zhifang Sui, Damai Dai, Baobao Chang, Sujian Li, and Hongying Zan. 2019. Preliminary study on the construction of chinese medical knowledge graph. *Journal of Chinese Information Processing*, 33(10):1–9.

Jun Cao, Xian Zhou, Wangping Xiong, Ming Yang, Jianqiang Du, Yanyun Yang, and Tianci Li. 2021. Electronic medical record entity recognition via machine reading comprehension and biaffine. *Discrete Dynamics in Nature and Society*, 2021.

Xianglong Chen, Chunping Ouyang, Yongbin Liu, and Yi Bu. 2020. 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.

Shanta Chowdhury, Xishuang Dong, Lijun Qian, Xiangfang Li, Yi Guan, Jinfeng Yang, and Qiubin Yu. 2018. A multitask bi-directional rnn model for named entity recognition on chinese electronic medical records. *BMC bioinformatics*, 19(17):75–84.

Qing Cong, Zhiyong Feng, Guozheng Rao, and Li Zhang. 2023. Chinese medical nested named

entity recognition model based on feature fusion and bidirectional lattice embedding graph. In *Database Systems for Advanced Applications: 28th International Conference, DASFAA 2023, Tianjin, China, April 17–20, 2023, Proceedings, Part IV*, pages 314–324. Springer.

Xiaojing Du, Yuxiang Jia, and Hongying Zan. 2022. Mrc-based medical ner with multi-task learning and multi-strategies. In *Chinese Computational Linguistics: 21st China National Conference, CCL 2022, Nanchang, China, October 14–16, 2022, Proceedings*, pages 149–162. Springer.

Yingjie Gu, Xiaoye Qu, Zhefeng Wang, Yi Zheng, Baoxing Huai, and Nicholas Jing Yuan. 2022. Delving deep into regularity: A simple but effective method for chinese named entity recognition. *arXiv preprint arXiv:2204.05544*.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *ArXiv preprint*, abs/2304.08247.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Bin Ji, Rui Liu, Shasha Li, Jie Yu, Qingbo Wu, Yusong Tan, and Jiaju Wu. 2019. A hybrid approach for named entity recognition in chinese electronic medical record. *BMC medical informatics and decision making*, 19(2):149–158.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.

Xiangyang Li, Huan Zhang, and Xiao-Hua Zhou. 2020a. Chinese clinical named entity recognition with variant neural structures based on bert methods. *Journal of biomedical informatics*, 107:103422.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020b. A unified mrc framework for named entity recognition. In *ACL*.

Yongbin Li, Xiaohua Wang, Linhu Hui, Liping Zou, Hongjin Li, Luo Xu, Weihai Liu, et al. 2020c. 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.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *ArXiv preprint*, abs/2303.14070.

Ning Liu, Qian Hu, Huayun Xu, Xing Xu, and Mengxin Chen. 2021. Med-bert: A pre-training framework for medical records named entity recognition. *IEEE Transactions on Industrial Informatics*.

Shuyue Liu, Junwen Duan, Feng Gong, Hailin Yue, and Jianxin Wang. 2023. Fusing label relations for chinese emr named entity recognition with machine reading comprehension. In *Bioinformatics Research and Applications: 18th International Symposium, ISBRA 2022, Haifa, Israel, November 14–17, 2022, Proceedings*, pages 41–51. Springer.

Ling Luo, Zhihao Yang, Yawen Song, Nan Li, and Hongfei Lin. 2020. Chinese clinical named entity recognition based on stroke elmo and multitask learning. *Chinese Journal of Computers*, 43(10):1943–1957.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

Qiuli Qin, Shuang Zhao, and Chunmei Liu. 2021. A bert-bigru-crf model for entity recognition of chinese electronic medical records. *Complexity*, 2021.

Qian Wan, Jie Liu, Luona Wei, and Bin Ji. 2020. A self-attention based neural architecture for chinese medical named entity recognition. *Mathematical Biosciences and Engineering*, 17(4):3498–3511.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *ArXiv preprint*, abs/2304.06975.

Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu. 2019. Nezha: Neural contextualized representation for chinese language understanding. *arXiv preprint arXiv:1909.00204*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *ArXiv preprint*, abs/2304.01097.

Donglin Yang, Huifan Yang, and Bin Wu. 2022. Tserl: Two-stage enhancement of radical and

lexicon for chinese medical named entity recognition. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2719–2726. IEEE.

Pan Yang, Xin Cong, Zhenyu Sun, and Xingwu Liu. 2021. Enhanced language representation with label knowledge for span extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4623–4635.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2023. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *arXiv preprint arXiv:2308.03549*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023. Huatuogpt, towards taming language model to be a doctor. *ArXiv preprint*, abs/2305.15075.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2022. Cblue: A chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915.

Hengyi Zheng, Bin Qin, and Ming Xu. 2021. Chinese medical named entity recognition using crf-mt-adapt and ner-mrc. In *2021 2nd International Conference on Computing and Data Science (CDS)*, pages 362–365. IEEE.

Feng Zhou, Xuming Han, Qiaoming Liu, Mingyang Li, and Yong Li. 2021. Chinese clinical named entity recognition based on stroke-level and radical-level features. In *Smart Computing and Communication: 5th International Conference, SmartCom 2020, Paris, France, December 29–31, 2020, Proceedings 5*, pages 9–18. Springer.

Qinglin Zhu, Zijie Lin, Yice Zhang, Jingyi Sun, Xiang Li, Qihui Lin, Yixue Dang, and Ruifeng Xu. 2021. Hitsz-hlt at semeval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 521–526.