LREC-COLING 2024

THE 2024 JOINT INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, LANGUAGE RESOURCES AND EVALUATION

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

Xiaojing Du, Hanjie Zhao∗, Danyan Xing, Yuxiang Jia†, Hongying Zan
Zhengzhou University, Zhengzhou, China

Reporter： Hanjie Zhao

Natural Language Processing Laboratory of Zhengzhou University

目录

1 Introduction

Introduction

□ Complex nested structures and sophisticated medical terminologies

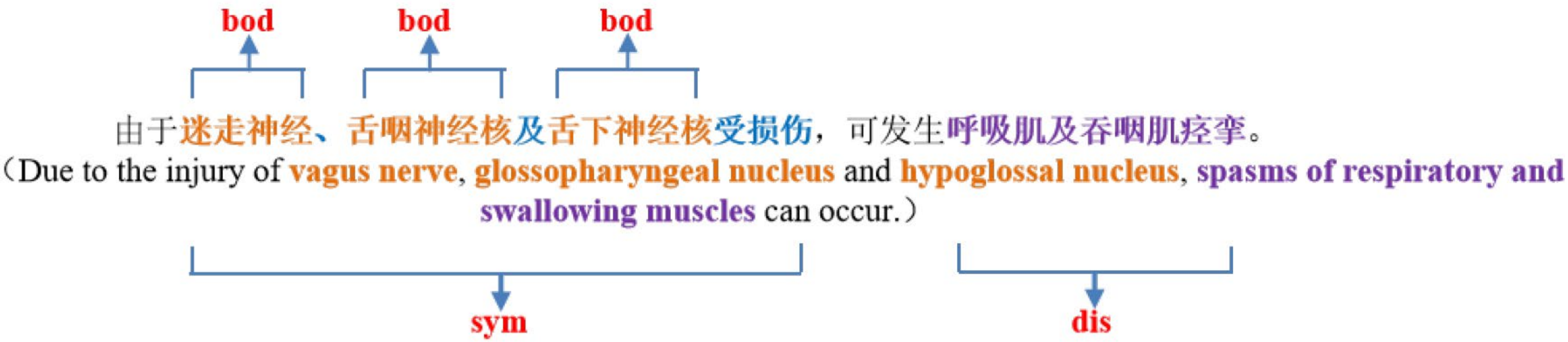


Figure 1: An example of nested entity.

Contributions

- We introduce a nested medical Named Entity Recognition model based on Machine Reading Comprehension (MRC), featuring the integration of Biaffine and Multi-Layer Perceptron (MLP) for joint prediction.
- We take a task adaptive pre-training strategy to optimize the pre-trained model for medical domain.
- Experimental results on the nested Chinese medical NER corpus CMeEE demonstrate superior performance of our model over compared models.

2

The MRC-CAP Model

The MRC-CAP Model

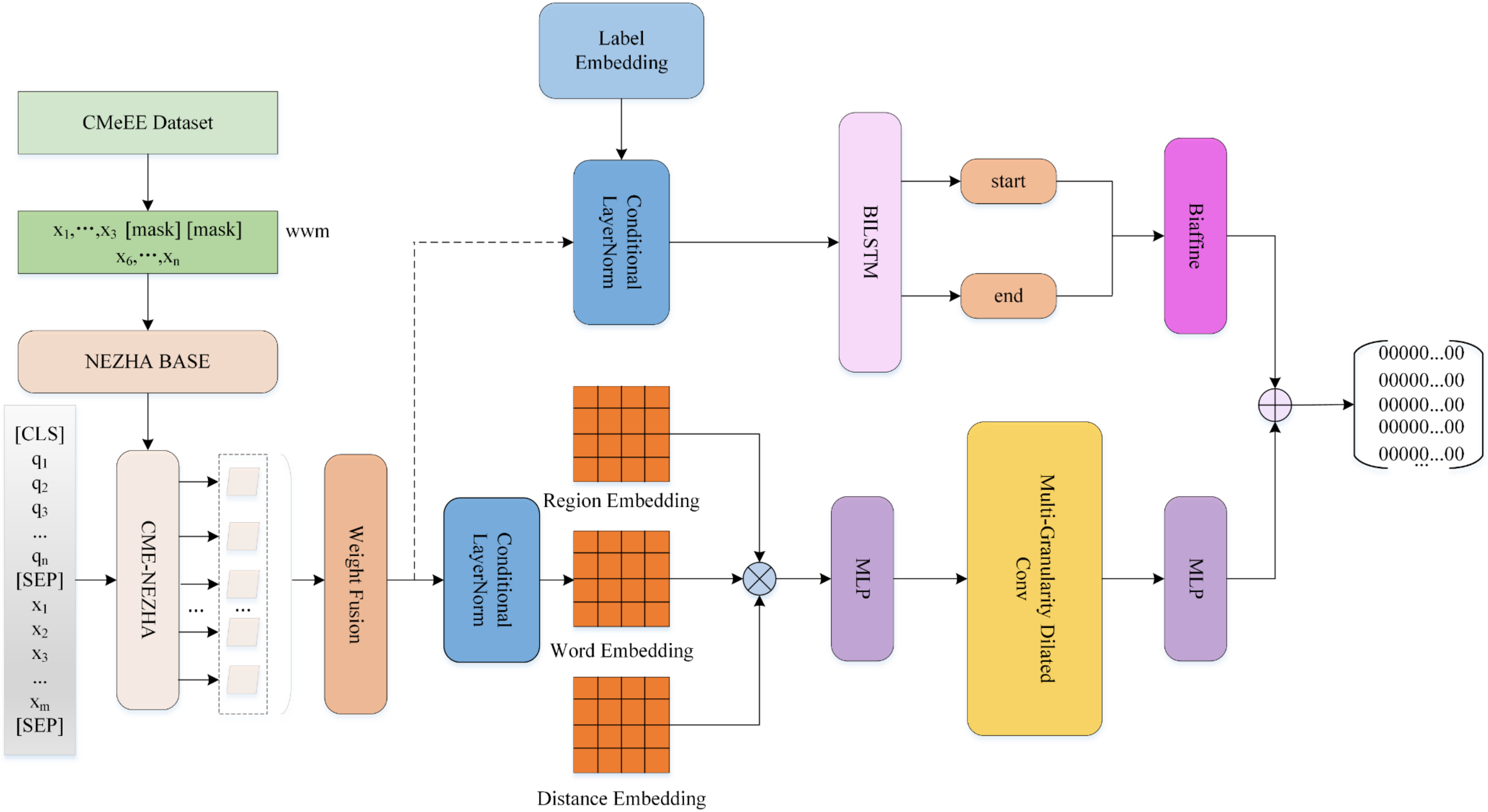
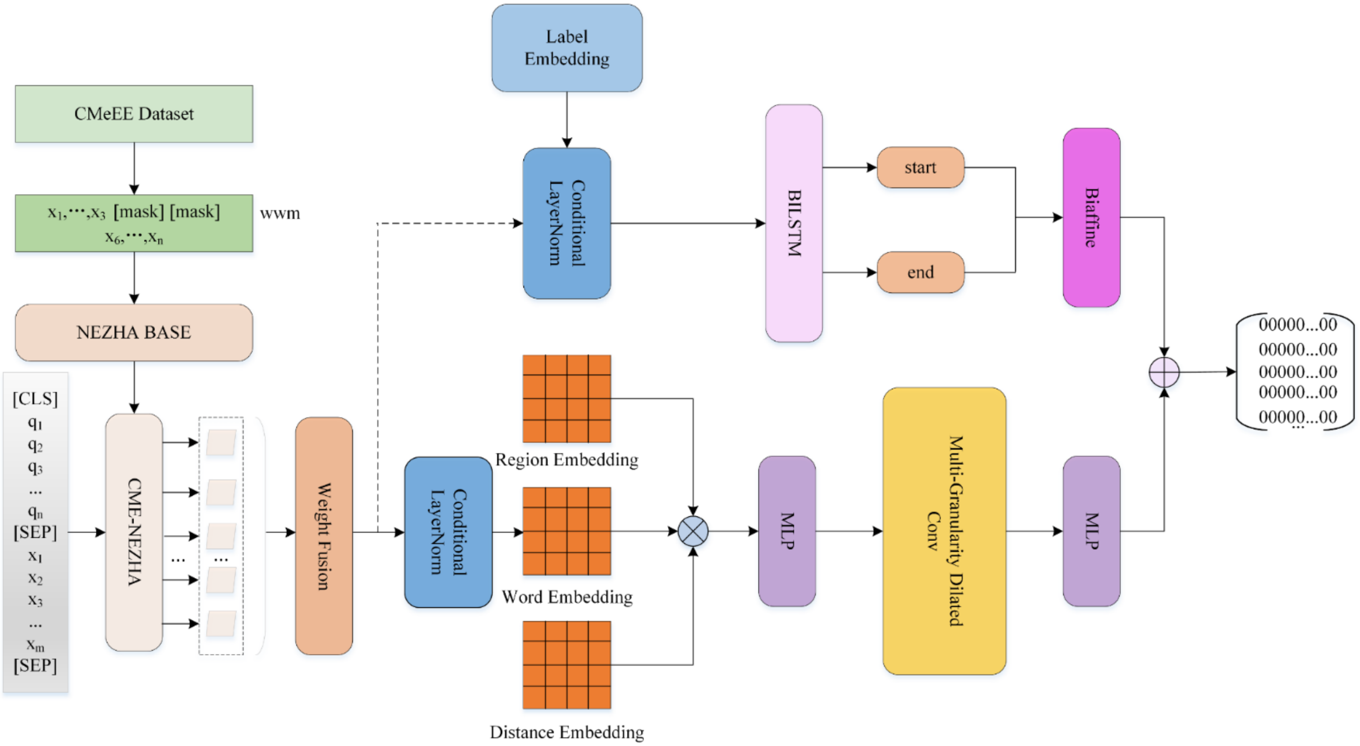


Figure 2: The architecture of the proposed NER model. The acronym 'wwm' stands for the BERT-Whole Word Masking model.

The MRC-CAP Model

◻ Model Details



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

$$x_j = MLP_{\text{end}}\left(h_i\right) \tag{2}$$

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

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

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

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

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

N is the length of the sentence, and C is the number of entity types+1

3 Datasets

More entity information

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

Datasets

☐ More nested entities

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

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

4 Experiments

Experiments

□ Query configuration

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

Comparison with previous models

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(OpenAI, 2023)	39.21	50.81	44.26
MRC-CAP(Ours)	**67.35**	**71.62**	**69.42**

Ablation study

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

☐ Recognition of various entity types

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

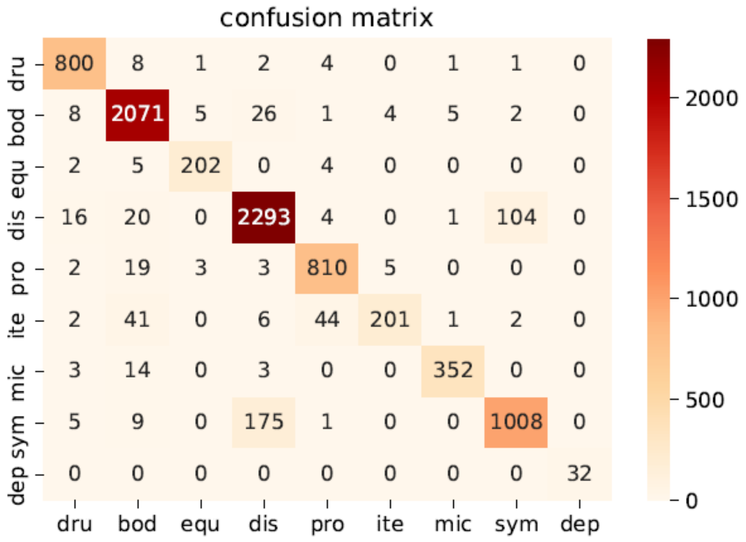


Figure 3: Confusion Matrix of NER on CMeEE V1.

Results of nested and flat NER

Table 9: 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

□ Case study

Table 8: 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。

5 Conclusion

Conclusion

- We propose an MRC-based medical NER with Biaffine and MLP for joint prediction of entity.
- Experiments on the nested Chinese medical NER benchmark CMeEE V1 and V2 show that the proposed model outperforms comparative 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.

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

Thanks