中文电子病历命名实体识别方法研究*

陈婕卿[1] 竹志超[2] 张 锋[1] 曾 可[1] 姜会珍[1] 程振宁[3]

([1] 中国医学科学院北京协和医院信息中心 北京 100730 [2] 北京工业大学信息学部 北京 100124
[3] 北京安妮福克斯信息咨询有限公司 北京 100005)

〔摘要〕 **目的/意义** 探索基于中文电子病历的命名实体识别方法在构建医学知识图谱和相关应用推广方面的技术可行性。**方法/过程** 采用真实医疗电子病历数据对词嵌入表示模型进行精化，构建医学术语专有嵌入表示，并利用卷积神经网络等多模型提取局部语义特征，实现基于堆叠注意网络的中文医疗命名实体识别。**结果/结论** 堆叠注意网络模型 *F*1 值达到 91.5%，较其他模型具备更强的医疗命名实体识别性能。进一步解决中文医疗命名实体识别难点，在实现全局语义特征全面深入提取的同时降低时间成本。

〔关键词〕 电子病历；命名实体识别；堆叠注意网络
〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.04.013

Study on Named Entity Recognition of Chinese Electronic Medical Records

CHEN Jieqing[1], *ZHU Zhichao*[2], *ZHANG Feng*[1], *ZENG Ke*[1], *JIANG Huizhen*[1], *CHENG Zhenning*[3]

[1] *Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China*; [2] *Beijing University of Technology, Beijing 100124, China*; [3] *Analyzefocus Information Consultant Ltd., Beijing 100005, China*

〔**Abstract**〕 **Purpose/Significance** To explore the technical feasibility of named entity recognition (NER) method based on Chinese electronic medical records (EMR) in the construction of medical knowledge graph and related application promotion. **Method/Process** The word embedding representation model is refined by using real EMR data, and the proprietary embedding representation of medical terms is constructed. Moreover, multiple models such as convolutional neural network (CNN) are used to extract local semantic features to realize the recognition of Chinese medical named entities based on stacked attention network (SAN). **Result/Conclusion** The *F*1 value of SAN model reaches 91.5%, which has stronger performance of medical NER than other models, so as to further solve the difficulty of Chinese medical NER, achieve comprehensive and in-depth extraction of global semantic features, and reduce the time cost.

〔**Keywords**〕 electronic medical records; named entity recognition; stacked attention network

1 引言

命名实体识别[1]是从自然语言文本中发现特定目标实体，而医疗命名实体识别则是从医疗文本中识别医疗实体边界并判断医疗实体类别[2]。常见医疗实体类别包括诊断名称、查体部位、治疗信息、检查或检验项目以及症状等。医疗命名实体识别的准确性影响事件抽取、关系抽取等任务效果，是医

〔修回日期〕 2023-11-27
〔作者简介〕 陈婕卿，助理研究员，发表论文 10 余篇；通信作者：张锋，高级工程师。
〔基金项目〕 科技创新 2030——“新一代人工智能”重大项目（项目编号：2020AAA0104900）。

疗知识图谱构建的关键基础。

医疗命名实体识别方法主要包括 3 种：基于字典和规则的方法、基于统计机器学习的方法和基于深度学习的方法。前两种需要耗费大量成本构建字典、制订规则或指定模型需要学习的特征。Wagh K 等[3]利用条件随机场（conditional random field，CRF）模型实现对医学文本中的基因等生物医学术语的高精度识别；Yuan X U 等[4]建立一个结合 CRF、基于规则文本注解的医疗命名实体识别模型并取得较好效果。深度学习方法使模型真正具有自主学习能力，因此应用广泛。Gridach M[5]研究识别生物医学命名实体时的单词嵌入方式和字符级表示方法。Li J 等[6]在循环神经网络的基础上同时考虑字符级嵌入和词级嵌入，取得很好的医疗实体识别效果。李灵芳等[7]研究基于双向编码器表征（bidirectional encoder representations from transformers，BERT）模型、双向长短期记忆（bidirectional long short - term memory，BiLSTM）模型和 CRF 的组合模型，通过 BERT 字嵌入模型更好地表示文本，解决一词多义问题，最终通过实验验证所提出方法的有效性。虽然基于深度学习方法的命名实体识别模型性能显著，但尚存在缺陷，即通用领域词嵌入模型无法表示医学领域特有的医学术语，导致模型无法对特有的医学术语进行表示，专有特征出现丢失。但是，传统方法缺乏对字符间局部语义特征的提取。

为解决上述问题，本文提出专用模型并使用大规模真实医疗电子病历数据对词嵌入表示进行精化，探索基于中文电子病历的命名实体识别方法在构建医学知识图谱和相关应用推广方面的技术可行性。

2 数据处理

2.1 数据标记方法

数据来自中国医学科学院北京协和医院 2019 年度的真实电子病历文书，数据经初步清洗后，共计 8 380 条电子病历文书被纳入研究，在专业医学专家和团队的指导下手动标记。随机抽取其中的 7 380条作为训练集和验证集（训练集与验证集样本量比例为 7∶3），其余电子病历文书作为测试集。对所有数据集进行预处理，包括病历文书章节分割，即根据章节标签进行拆分，采用 BIO 标记方法，其中“B”表示医疗实体起始位置的标签，“I”表示医疗实体剩余部分的标签，“O”表示当前字符不是医疗实体。“B - X”或“I - X”中的“X”是医疗实体的类别，该数据集共有 13 个不同标记，见表 1。每条文本均标有名称、起始位置和医学类别。本文将实体定义为 6 类，包括“症状”“检查”“结果”“疾病”“治疗”和“否定”。

表 1 BIO 标记方法及其示例

类别	标记
症状	B - 症状/I - 症状
检查	B - 检查/I - 检查
结果	B - 结果/I - 结果
疾病	B - 疾病/I - 疾病
治疗	B - 药物/I - 药物
否定	B - 否定/I - 否定
非医疗实体	O

2.2 数据预处理

数据预处理阶段，前期由人工以及程序初步标注数据，然后训练命名实体识别和关系抽取模型。在命名实体识别和关系抽取过程中，将医疗本体和预训练语言模型学习到的文本特征相结合，增强特征区分能力，从而提高命名实体识别和关系抽取模型的识别精度。同时利用模型对未标注数据进行概念实体和关系的提取，过程中合理利用外部医学知识资源，互相融合扩充和优化，从而实现大规模医疗知识库的增量构建。此外，通过集成构建的实体对齐模型对图谱包含的实体进行对齐，以保证图谱质量。后期根据上述模型开发自动化标注程序，经过算法标注的数据再辅以人工审核和双人背对背校验，确保数据标注的有效性和一致性，减少标注误差的影响。总体完成的已标注数据中包含疾病类别 32 种，共 14 272 条。

3 研究方法

采用由多种模型融合而成的堆叠注意网络（stacked attention network，SAN）模型，包含 3 组模块：数据预处理模块、字符嵌入模块和语义分析模块（集成卷积神经网络（convolutional neural networks，CNN）模型、堆叠 BiLSTM 模型、注意力机制和 CRF 模型），见图 1。数据预处理模块对输入的数据进行预处理，对字符、单词和句子进行分割，并对词性进行标注。字符嵌入模块使用大规模真实电子病历未标注语料（北京协和医院 2018 年度电子病历文书）对基于 BERT 的优化版预训练模型 RoBERTa 进行预训练，以精化字符嵌入，补充医学术语特征。语义分析模块利用 CNN 模型捕捉字符间的局部依赖关系特征，通过构建堆叠 BiLSTM 全面充分捕捉文本的全局语义特征，将其与 CNN 的输出进行拼接，构建语义特征更丰富的文本表示。最后，将注意力机制分配的权重与文本表示融合送入 CRF 计算预测序列标签。

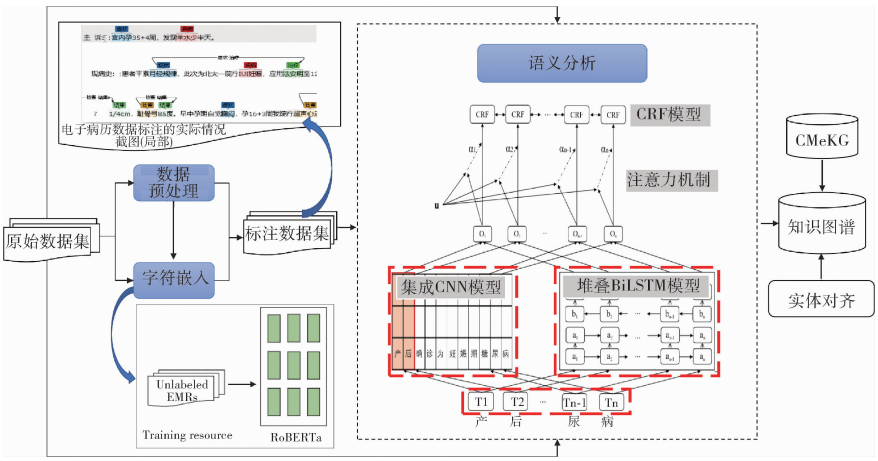


图 1 基于多模型融合的堆叠注意网络架构

RoBERTa 的输入是电子病历句子中的字符，输出是字符的特征向量。通过输入 3 个不同的特征值（文本表示向量、句子分割向量和位置向量）获得特征向量，位置向量计算方式如下，编码方法利用 sin 函数和 cos 函数，pos 是文本中的字符，i 表示维度，编码后的向量维度为 d_{model}。

$$PE(pos, 2i) = \sin(pos/10\,000^{2i/d_{\text{model}}}) \quad (1)$$

$$PE(pos, 2i+1) = \cos(pos/10\,000^{2i/d_{\text{model}}}) \quad (2)$$

3.1 卷积神经网络

CNN 模型能有效地捕捉像素点与像素点间的局部依赖关系信息，因此最初被广泛应用于计算机视觉。在自然语言处理领域，CNN 输入是以向量表示的句子或者文档。向量的每一行对应一个字符或单词，即每行代表一个嵌入向量，卷积核滑过的是向量中的一“行”（单词），见图 2。

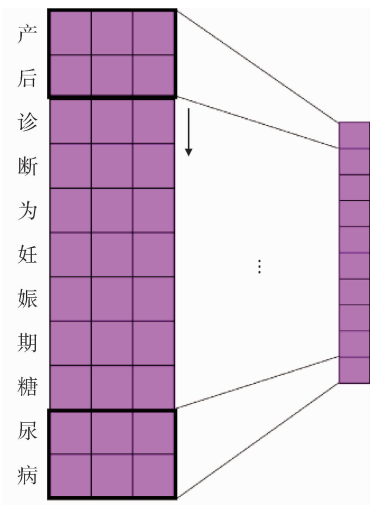
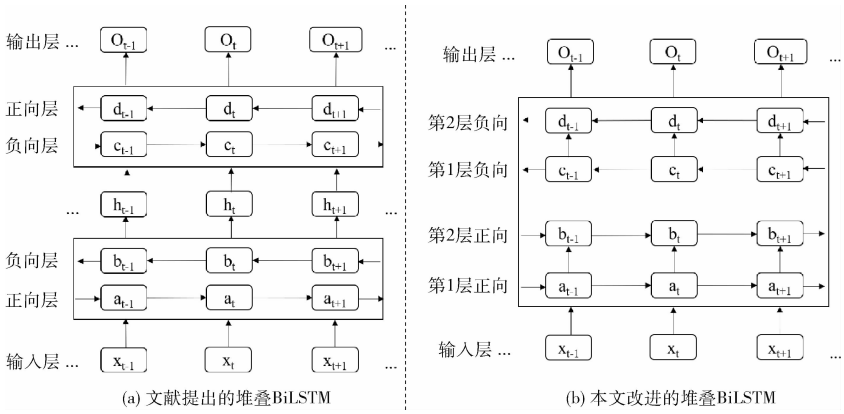


图 2 卷积神经网络模型运算方式

3.2 堆叠 BiLSTM 模型

传统 BiLSTM 模型序列标注任务的权重赋予是等价的，解决上下文依赖关系特征的捕捉问题，序列标注任务中当前状态之前和之后的状态是平权的。但现有研究中词嵌入表示呈现高维性，因此传统方法存在全局语义特征提取不充分的问题。随着技术的发展，Ding Y 等[8]利用带有多层隐藏层结构的堆叠 BiLSTM 模型对全局语义特征进行提取，并证明其具有更强的全局语义提取特征能力。受其启发，本文构建一种带有多层隐藏层结构的堆叠 BiLSTM 模型，更充分地捕捉文本全局语义特征并改进，见图 3。



(a) 文献提出的堆叠BiLSTM (b) 本文改进的堆叠BiLSTM

图 3 堆叠 BiLSTM 模型架构对比（每个方向上 2 层）

DING Y 等[8]所提出的方式仅将两个独立 BiLSTM 模型进行串联式拼接，而本文对其进行改进，将两个 BiLSTM 模型融合在一起，以实现多次全局特征提取，同时上文和下文的全局特征提取操作互不干扰，特征提取更为纯粹，省略中间层的整合计算并减少时间消耗。

3.3 注意力机制

注意力机制最重要的贡献是区分文本中的关键信息，重点关注对结果影响较大的关键特征，尽可能地忽略无关特征。因此，引入注意力机制以增强模型的实体识别性能，见图 4。运算方式如下，其中 Q、K 和 V 3 个参数分别代表 Query、Key 和 Value。

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^T/\sqrt{d_k})V \quad (3)$$

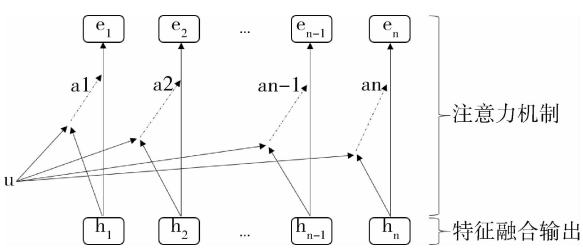


图 4 注意力机制的权重分配过程

3.4 条件随机场

CRF 模型是一种经典的判别概率无向图模型，通常用于序列标记任务。在 CRF 模型预测过程中，利用维特比算法求解全局最优序列，计算方式如下。E^* 是 Score 在函数中获得最高分数的序列。

$$E^* = \underset{x\tilde{e} \in E_X}{\text{argma}}\text{Score}(X, \tilde{e}) \quad (4)$$

3.5 多模型融合

经过微调形成的 RoBERTa 表示能力优于原始的 BERT 模型，通过强力模型和高质量电子病历数据构建文本表示模型，可以为下游任务（如命名实体识别、关系抽取等）提供更准确的文本表示。目前模型的主要改进点和创新性体现在以下 3 方面：字符嵌入优化、增加 CNN 局部特征提取和提出多层结构堆叠 BiLSTM 模型，见图 5。虚线框为 SAN 模型相比传统模型的优化改进部分，字嵌入改进即为文本表示模型的改进，CNN 改进是为增加字符间的依赖关系特征，堆叠 BiLSTM 模型能够提取全局深层特征。3 种改进方法的主要目的是为模型增加可利

用的特征，增强模型学习过程，从而实现更好的命名实体识别效果。选择“妊娠期糖尿病”数据进行训练的原因是其数据规模较大，模型训练更充分，结果更具说服力。

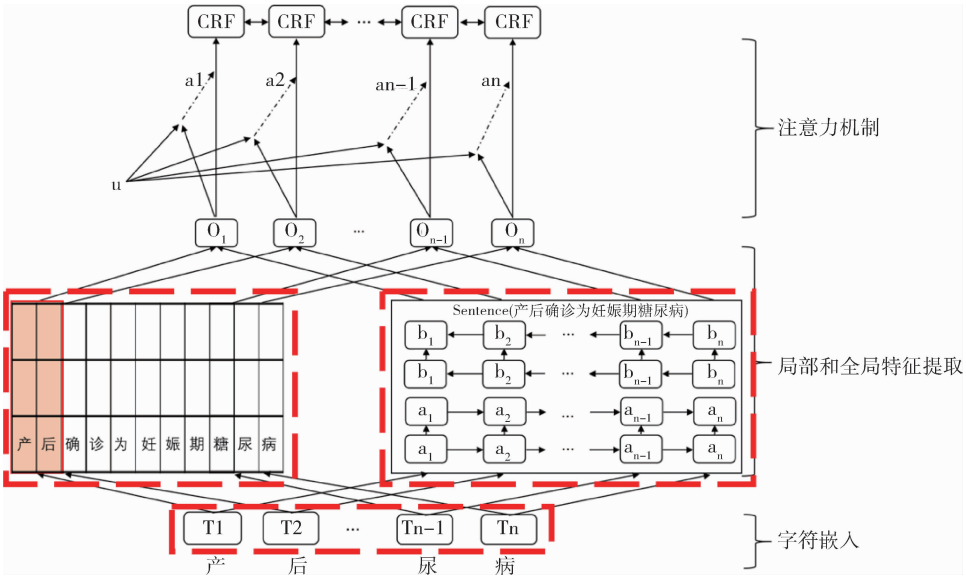


图 5 基于堆叠注意网络模型训练过程

3.6 评估指标项

评估指标采用精度、召回率和 *F*1 值。其中 TP 表示正确识别的实体数，FP 表示识别的不相关实体数，FN 表示未识别实体数。在预测过程中，判断医疗实体的预测是否完全正确的标准，即实体的边界和类别同时被正确预测。*F*1 值为精度和召回率的加权平均值，具体计算过程如下。

$$\text{精度} = \frac{TP}{TP + FP} \times 100\% \quad (5)$$

$$\text{召回率} = \frac{TP}{TP + FN} \times 100\% \quad (6)$$

$$F1 = \frac{2 \times \text{精度} \times \text{召回率}}{\text{精度} + \text{召回率}} \times 100\% \quad (7)$$

4 结果与讨论

4.1 多模型结果对比

为确保实验结果的可靠性，本研究复现一些经典和高级模型作为基线模型进行比较，拟定模型名称的简写和含义如下：DM 表示字典匹配；HMM 表示隐马尔可夫模型；BC 表示 BiLSTM + CRF；BAC 表示 BiLSTM + 注意机制 + CRF；BBC 表示 BERT 基础模型 + BiLSTM + CRF；FBBCE 表示微调 BERT + BiLSTM + CRF + 外部字典。基线和本研究所提出的 SAN 的性能结果，见表 2。BC、BAC、BBC 和 FBBCE 利用深度学习技术，优于 HMM。此外，BBC 和 FBBCE 使用基于字符嵌入的 BERT 模型学习字符级的单词表达进行实体识别，有效避免了错误的中文分词对模型的负面影响。同时，与 BBC 相比，FBBCE 得益于特定领域的知识和字典来增强模型，只进行一次全局语义特征提取，不足以充分挖掘全部特征。SAN 模型在数据集上实现最高 *F*1 值（91.5%）和较高性能，充分利用电子病历数据对字符嵌入进行精化，且 CNN 模型弥补局部依赖关系特征丢失。相比于现有 6 种典型方法，SAN 模型具有较高识别精度。使用电子病历数据验证，相比现有命名实体识别方法性能更优。

表 2 不同模型的精度、召回率和 *F*1 值对比（%）

模型	精度	召回率	*F*1
DM	42.7	67.5	53.2
HMM	74.8	74.4	74.6
BC	78.9	79.2	79.1
BAC	87.0	87.0	87.0
BBC	89.0	87.0	88.0
FBBCE	90.0	89.0	89.5
SAN	92.0	91.0	91.5

4.2 多模型融合对比

将特定领域医学知识迁移到 RoBERTa 中，对字符嵌入进行精化并使用 CNN 补充局部语义特征，提出使用多层结构的堆叠 BiLSTM 模型对文本表示进行深入的全局语义挖掘。为进一步验证这 3 种增强方法的有效性以及增加模型的可解释性，对构建模型组合进行比较，多模型融合后的名称简写和组合方式，见表 3，其中 R 代表 RoBERTa，c 代表 CNN，S 代表堆叠，B 代表 BiLSTM，A 代表注意力机制，C 代表 CRF。

表 3 多模型融合简写及其含义

模型	RoBERTa	微调 RoBERTa	RoBERTa 堆叠	微调 RoBERTa 堆叠	BiLSTM	注意力机制	CRF	CNN
RBAC 基线	√				√	√	√	
微调 RBAC		√			√	√	√	
RSBAC			√		√	√	√	
RBcAC	√				√	√	√	√
微调 RSBAC				√	√	√	√	
RSBcAC			√		√	√	√	√
微调 RBcAC		√			√	√	√	√

RBAC 未添加本文所强调的任何性能增强方法，即作为基线模型，对比结果，见表 4。微调 RBAC、微调 RSBAC 以及 SAN 的结果指标分别优于 RBAC、RSBAC 以及 RSBcAC。这一现象表明，在特定领域知识中精化后的 RoBERTa 较没有经过微调的 RoBERTa – base 生成更精确的字符嵌入，可以更好地表示医学文本，并提高模型的识别能力，这表明字符嵌入精化能增强模型的性能。观察 CNN 局部依赖关系特征提取对模型的影响可知，包含堆叠 BiLSTM 模型的 RSBAC、RSBcAC 以及 SAN 在实体识别上的性能更为出色，这也进一步证明堆叠 BiLSTM 模型强大的全局语义特征提取能力。结合 3 种增强方法的 SAN 模型得分最高、性能最好。这些实验结果增强模型可解释性的同时，也强有力地说明本文所提出的 3 种增强方法均能显著提升模型的实体识别能力。

表 4 多模型融合的精度、召回率和 *F*1 值对比（%）

模型	精度	召回率	*F*1
RBAC 基线	87.4	86.8	87.1
微调 RBAC	87.9	88.4	88.1
RSBAC	88.0	88.3	88.1
RBcAC	88.2	88.3	88.3
微调 RSBAC	89.3	90.9	90.4
RSBcAC	90.5	90.7	90.6
微调 RBcAC	90.2	90.5	90.3
SAN	92.0	91.0	91.5

4.3 堆叠 BiLSTM 模型的不同堆叠层数对比

堆叠 BiLSTM 模型提取文本的全局语义特征，其堆叠方式与其他堆叠 BiLSTM 模型存在较大区别。同时，堆叠 BiLSTM 模型在不同隐藏层数下对模型的最终实体识别性能也有影响。因此，为验证本文所提出的堆叠方式更加优越以及确定最合理的堆叠层数，在保持其他条件不变的情况下进行对比实验，采用控制变量法，在不同堆叠方式下，对堆叠的层数逐层递增。值得注意的是，在堆叠层数为 1 层时（每个方向上），本文所提出的方法与既往研究所提到的并无区别，即 BiLSTM 模型，将其设置为基线模型，与本文方式（A）和 DING Y 等[8]提出的方式（B）对比实验的结果，见表 5。

表 5 堆叠 BiLSTM 模型的不同堆叠层数对实体识别性能的影响（%）

层数	方式	精度	召回率	*F*1
1	基线模型	90.2	90.5	90.3
2	A	92.0	91.0	91.5
	B	90.9	91.4	91.1
3	A	91.8	89.9	90.8
	B	90.4	91.0	90.7
4	A	91.0	91.1	91.1
	B	90.6	91.2	90.9

随着隐藏层数的逐层增加，模型性能指标整体呈现增加趋势。从 1 层增加到 2 层时的性能跨越较明

显，*F*1 提升近 1.2 个百分点。2 层和 3 层之间，以及 3 层和 4 层之间的跨越较微弱，但随着层数的递增训练时间在成倍增加，见表 6。更多隐藏层结构的堆叠 BiLSTM 模型确实能够比基线 BiLSTM 模型在全局语义特征提取方面做得更好，但模型训练过程花费一定时间成本。因此，堆叠层数设置要根据具体任务需求进行权衡，本文认为将堆叠 BiLSTM 模型的隐藏层数设为 2 层时最好，兼顾性能和时间成本。堆叠 BiLSTM 模型的堆叠方式与既往研究中所提到的堆叠方式在实体识别性能上并无太大差别，主要原因是两种方式下的堆叠 BiLSTM 在对文本的全局语义特征提取方面的操作并无实质区别，都是在同样方向上提取相同次数，模型在性能上极为相似。

表 6 堆叠 BiLSTM 模型的不同堆叠层数对实体识别时间消耗的影响

层数	方式	时间消耗（秒）	时间消耗差值（秒）	*P*
1	基线模型	45.2	-	-
2	A	89.8	4.0	<0.05
	B	93.8		
3	A	137.5	14.6	<0.01
	B	152.1		
4	A	191.4	21.2	<0.01
	B	212.5		

从时间角度来看，堆叠 BiLSTM 模型在每个机器学习训练轮数（epoch）上的时间消耗（秒）比既往研究提及方式更短，经方差分析可以看出堆叠层数为 2 时 $P<0.05$，堆叠层数为 3 和 4 时 $P<0.01$，均具有显著性差异。关于效率提升的原因，主要是堆叠 BiLSTM 模型在结构上省去了运算中间的整合操作，并减少了运算步骤和时间。相比于深度学习模型而言，一般模型训练过程基本都要设置几百甚至上千个 epoch 才能逐渐收敛，而随着堆叠层数不断增加，时间将成倍增加。综上，堆叠 BiLSTM 模型可以显著减少模型训练时间成本，提高模型学习效率。

5 结语

本文对医疗命名实体识别任务进行研究，主要贡献在于：分析命名实体识别对医学研究的重要意义以及其特有的难点，探索基于中文电子病历的命名实体识方法在构建医学知识图谱和相关应用推广中的技术可行性，并通过真实医疗电子病历数据进行验证。本文实现基于堆叠注意网络的中文医疗命名实体识别，在数据集上的 *F*1 得分达到 91.5%，优于基线模型，具备更强医疗实体识别能力，证实该方法可以更精准地对医学文本进行表示。未来计划引入外部语言知识对模型进行优化，通过增加知识特征来进一步提高模型性能。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 NADEAU D, SEKINE S. A survey of named entity recognition and classification [J]. Lingvisticae investigationes, 2007, 30 (1): 3 - 26.

2 许思特，孙木. 基于命名实体识别与 Neo4j 的中文电子病历知识图谱构建和应用 [J]. 医学信息学杂志，2022, 43 (12): 50 - 56.

3 WAGH K S, KULKARNI A, KASHID S, et al. CRF based bio - medical named entity recognition [J]. International journal of emerging technology and computer science, 2018, 3 (2): 135 - 139.

4 YUAN X U, YAN - QIU G E, QIANG W, et al. Medical name entity recognition and application in Chinese admission record of stroke patients based on CRF and RUTA rule [J]. Journal of Sun Yat - sen University (medical sciences), 2018, 39 (3): 455 - 462.

5 GRIDACH M. Character - level neural network for biomedical named entity recognition [J]. Journal of biomedical informatics, 2017, 70 (2): 85 - 91.

6 LI J, ZHAO S, HUANG Z, et al. A novel RNN - based approach for bio - NER in Chinese EMRs [J]. Journal of supercomputing, 2020, 76 (3): 1450 - 1467.

7 李灵芳，杨佳琦，李宝山，等. 基于 BERT 的中文电子病历命名实体识别 [J]. 内蒙古科技大学学报，2020, 132 (39): 75 - 81.

8 DING Y, ZHOU X, ZHANG X. YNU_DYX at SemEval - 2019 Task 9: a stacked BiLSTM for suggestion mining classification [C]. Minneapolis: The 13th International Workshop on Semantic Evaluation, 2019.