基于深度学习的医疗命名实体识别的研究

莫大利*，吴冠锋，徐馨怡

贵州医科大学 贵州贵阳

【摘要】电子病历包含患者的病史、检查结果、诊断和治疗方案等重要医学信息，是支撑医疗智能化研究不可或缺的数据来源。然而由于电子病历的语言表达多样且复杂，医务人员往往难以快速准确地进行信息抽取和分析。基于深度学习的命名实体识别已成为电子病历自动化信息抽取的核心技术。但是医学自然语言处理的复杂性和医疗实体独特性导致电子病历的命名实体识别仍然存在挑战。本文通过构建 BERT-MNER 模型训练数据集和利用 BERT 预训练语言模型获取每个字的上下文相关嵌入表示，结合 BiLSTM 和 CRF 层完成医疗实体识别。实验结果表明，相较于基于单一深度学习网络的命名实体模型，BERT-MNER 模型能更有效地识别电子病历中的医疗实体。

【关键词】电子病历；BERT-MNER 模型；深度学习；命名实体识别

【收稿日期】2023 年 5 月 17 日 【出刊日期】2023 年 6 月 20 日 【DOI】10.12208/j.ijmd.20230095

A study on medical named entity recognition based on deep learning

Dali Mo, Guanfeng Wu, Xinyi Xu*

Guizhou Medical University, Guiyang, Guizhou

【**Abstract**】 Electronic medical record contains important medical information such as patients' medical history, examination results, diagnosis and treatment plans, and is an indispensable data source to support intelligent medical research. However, due to the diverse and complex linguistic representations of electronic medical records, it is often difficult for medical professionals to extract and analyze information quickly and accurately. Named entity recognition based on deep learning has become a core technology for automated information extraction from electronic medical records. However, the complexity of medical natural language processing and the uniqueness of medical entities lead to the challenges of named entity recognition in electronic medical records. In this paper, medical entity recognition is accomplished by constructing a BERT-MNER model training dataset and using a BERT pre-trained language model to obtain contextually relevant embedding representations of each word, combined with BiLSTM and CRF layers. The experimental results show that the BERT-MNER model can identify medical entities in electronic medical records more effectively than the named entity model based on a single deep learning network.

【**Keywords**】Electronic medical record; BERT--MNER model; Deep learning; Named entity recognition

1 引言

海量电子病历数据是支撑医疗智能化研究的重要材料，然而电子病历文本数据半结构化甚至无结构化的特点，对其后续的分析利用造成极大困难。虽然近年来基于深度学习的命名实体识别已经成为对电子病历进行自动化信息抽取的核心技术，但中文电子病历具有其独特的文本数据特征，而这些文本数据特征增加了医学自然语言处理的复杂性。现存挑战之一为病历文本存在非规范性和专业性等问题，即医务人员会使用各自的电子病历术语体系致使其电子病历的书写风格千差万别，病历文本中可能包含如专业医学术语、缩写、拼写错误和拼音等目前尚未制定具体使用规范的病例文本用语[1]。这些特征与问题使得病历文本的自动处理变得更加困难。另一个挑战是医疗实体的独特性。医学实体包括疾病、症状、药物、手术等，它们的命名和描述

*通讯作者：莫大利

方式都是独特的。这些实体的标准化和识别对于病历文本分析至关重要，但是这需要充足的语料库和算法支持[2]。还有一个挑战是标注好的语料集稀缺。由于医疗行业的特殊性质，很多数据由于涉及患者隐私不予公开，致使目前标注好的中文电子病历的语料集相对比较稀缺。因此，需要更多的数据资源才能更好构建高质量的中文电子病历标注语料库。为了应对这些挑战，研究者们正在积极探索各种方法，如基于深度学习的模型、基于规则的方法和深度强化学习等，以帮助解决中文电子病历的医学自然语言处理问题。

2 数据来源与预处理

本文使用的实验数据来自于阿里云天池数据集中的 Yidu-S4K：医渡云结构化 4K 数据集（Yidu-S4K:医渡云结构化 4K 数据集_数据集-阿里云天池 （aliyun.com），其源自 CCKS2019 评测任务一的子任务“面向中文电子病历的医疗实体识别”。该数据集包括训练集和测试集，各包含 7717 个句子和 379 个句子，每个句子标注了医疗实体的名称、开始位置、结束位置和预定义类别信息[3]。电子病历数据预处理步骤包括数据抽取、数据清洗、数据规约和数据脱敏。而在获取的数据集的中，存在一定缺失值、异常值、重复值等问题，因此为了避免这些问题对我们数据的分析结果产生误导，对于这些问题数值，在本次分析中，我们采取删除缺失值的方法，最后保证得到对数据质量有可靠性的评估结果[4]。BERT-MNER 模型是基于预训练的 BERT 模型，结合命名实体识别（NER）任务进行微调训练，从而实现对肿瘤原发部位的识别[5]。可以分为以下几个步骤：

特征提取，将文本数据转化为向量形式，再提取相关特征，如词向量、部分词性特征等；

设计模型：根据任务需求设计深度学习模型，如 BiLSTM-CRF 模型、BERT-MNER 模型等。

模型训练：使用训练集对模型进行训练，通过交叉验证等方式进行参数调优，以提高模型准确率。

模型评估：使用测试集验证模型效果，计算模型的准确度、召回率、F1 值等指标。

模型应用：将训练好的模型应用到实际应用场景中进行测试、部署。

命名实体识别的本质是对文本数据进行实体识别和分类。在本次研究中，将 NER 用于从病历中提取有关医学实体的信息，包括肿瘤原发部位、病灶大小和转移。为了标注整个文本序列，NE 标注通常采用 BIO 或 BIOES 的标签方法。BIO 标签方法将实体的开始字符标记为 B，中间字符标记为 I，其它字符标记为 O[6]。BIOES 标签方法则将实体的开始字符标记为 B，结束字符标记为 E，中间字符标记为 I，单个字符标记为 S，其它字符标记为 O。对于医疗电子病历的标注，可以使用 BIO 或 BIOES 标签方法标注每个实体类型，再从序列中抽取出有用的医疗实体信息。通过人工标注的医疗实体，可以为建立自然语言处理模型提供必要的训练数据，进而提升对医疗文本分析的效果。通过训练好的模型对电子病历中的文本数据进行实体识别，进而识别出文本中的肿瘤相关实体。

3 模型的建立

BERT-MNER 模型是结合 BERT 预训练语言模型对电子病历中的医疗实体进行识别。传统的 BiLSTM-CR F 模型是通过中文维基百科语料上训练的字向量进行嵌入表示，是上下文无关的。而 BERT-MNER 模型包括 BERT 层、BiLSTM 层和 CRF 层，并且使用 BERT 预训练语言模型得到每个字的上下文相关的嵌入表示。BERT 层利用 BERT 预训练语言模型对输入的文本序列进行特征提取，得到每个字符或词组的上下文相关的嵌入表示，BiLSTM 层通过双向 LSTM 层将 BERT 层得到的特征向量编码自动抽取上下文特征， CRF 层采用 CRF 层对 BiLSTM 编码后的特征序列进行全局标签解码，并给出最优序列作为输出。预训练语言模型通过大规模的语料预训练得到了通用的语言知识，能够更好地理解自然语言中的多义性，并为下游自然语言处理任务提供更好的输入表示，从而提高任务的表现。在 BERT 预训练语言模型中，其编码器采用了 Transformer 结构，能够很好地捕捉输入文本中的长程依赖关系，通过自注意力机制提取关键信息[7]。总特征向量是由字向量、句子切分向量和位置向量相加得到的，其中位置向量采用正弦和余弦函数进行编码，能够区分不同位置之间的词语或者字符，具备很好的位置信息表示能力。BERT 预训练语言模型的强大表现得益于这些强有力的特性，为各类自然语言处理任务提供了高度可定制的输入表示，其应用实现了自然语言处理任务中最优秀的表现，为这个领域的许多重要工作提供了极大的助力。

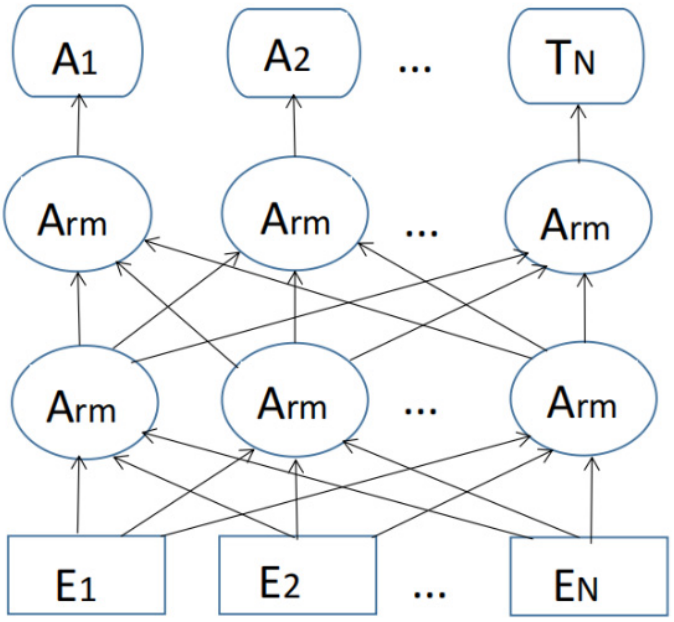

图 1 BERT 预训练语言模型

$$PE（pos，2i）=\sin（pos/100002i/dmodel）\quad（1）$$

$$PE（pos，2i+1）=\sin（pos/100002i/dmodel）\quad（2）$$

BERT 深度学习的一个基础模型，其他模型可以在此基础上进行迁移学习，通过增加双向 Transformer、以及遮蔽语言模型还有句子级别等的负采样方式，对文本中的词级、句子级、字符级、以及语句之间的特征关系进行充分描述。在中文电子病历的命名实体识别中，通过微调 BERT 模型，其具有作为特征提取器的特性，在与下游任务进行融入时，既定任务的词嵌入特征可以由通过模型提取出的特征替代。BERT-wwm 模型是一个专门为中文设计的深度学习模型，可对汉字医疗进行文本中的医疗实体特征表示进行选择，此模型由 24 层网络构成，其中隐藏层的维度为 1024，含 16 个头，总参数量达 330 M。类似于 BERT 这样的预训练语言模型，通过对大规模无标注数据集进行预训练，有效保证了模型训练效果。同时在中文电子病历的标注语料集极度稀缺的情况下，可以有效应对医疗实体命名识别在较为独特的中文电子病历上进行难题。并且目前 BERT 模型在命名实体识别存在两种研究趋势：一个是此模型对输入的文本长度有所限制，针对这一问题的其中一种解决方法是将 BERT 作为一个具有字符嵌入功能的生成器，将对较长的文本进行分割后得到的等长文本输入 BERT 模型中其对应的字符嵌入进行计算，在与字音、字形等其他特征进行融合后输入到其他模型中继续进行对中文电子病历的命名实体识别。另一种解决方式是从字粒度、词粒度和句粒度等方面进行变长与分割处理原始语料，再输入 BERT 模型，以这一解决方法可以有效保证了输入文本特征的完整性。而神经网络定义为在训练时输入随机的初始化参数后，开启循环计算输出结果，与实际结果比较，但是要想达到处理信息的目的，还需要依靠系统自身的复杂程度，对参数内部大量节点之间相互连接的关系进行调整。将各个节点想象成其自身的线性回归模型，由输入数据、权重、偏差（或阈值）和输出组成。公式如下（3）。而单层神经网络可以计算连续输出而不是阶跃函数。一种常见的选择是所谓的逻辑函数（下式 4）。这种基于单层神经网络的计算与广泛运用于统计建模的逻辑回归模型相同。

4 模型求解

本文通过建立以 BERT——MNER 模型，经过计算得到各部位数据，最终得出的预测结果如表 1 所示。

$$\sum_{i=1}^{m} w_i x_i + bias = w_1 x_1 + w_2 x_2 + w_3 x_3 + bias \quad (3)$$

$$f(x) = \frac{1}{1 + e^{-x}} \quad (4)$$

表 1 预测结果

预测测试集结果 Y	原文	肿瘤原发部位	原发病灶大小	转移部位
144.57666891190365	139	5	123	97
105.94876317879556	140	2	124	8
145.64971104505878	141	26	125	98
146.74262214258607	142	29	126	99
147.56814656427127	143	38	127	100
148.57192876985118	144	43	128	101
150.55612878685224	145	26	129	102
152.8522799256691	146	2	130	103
152.2071776302226	147	44	131	104
153.3892176196973	148	45	132	105
129.68970876069704	149	46	133	50
156.2448375917999	150	26	134	106
62.20972356973719	151	18	69	8
158.58555317659042	152	1	135	107
159.7675931660651	153	2	136	108

5 结论

本文综述了中文电子病历命名实体识别的研究和进展，通过系统梳理，介绍了中文电子病历命名实体识别，将 BERT-MNER 模型结合 BERT 预训练语言模型和 BiLSTM-CRF 模型对电子病历中的医疗实体进行识别能够更好地识别出电子病历中的医疗实体。

参考文献

[1] Zhang, R., Zhao, P., Guo, W., Wang, R., & Lu, W. (2022). Medical named entity recognition based on dilated convolutional neural network. Cognitive Robotics, 2, 13-20.

[2] Govindarajan, S., Mustafa, M. A., Kiyosov, S., Duong, N. D., Raju, M. N., & Gola, K. K. (2023). An optimization based feature extraction and machine learning techniques for named entity identification. Optik, 272, 170348.

[3] Wang, X., Liu, R., Yang, J., Chen, R., Ling, Z., Yang, P., & Zhang, K. (2022, May). Cyber threat intelligence entity extraction based on deep learning and field knowledge engineering. In 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 406-413). IEEE.

[4] An, Y., Xia, X., Chen, X., Wu, F. X., & Wang, J. (2022). Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF. Artificial Intelligence in Medicine, 127, 102282.

[5] Rezayi, S., Dai, H., Liu, Z., Wu, Z., Hebbar, A., Burns, A. H. & Li, X. (2022, September). Clinicalradiobert: Knowledge-infused few shot learning for clinical notes named entity recognition. In International Workshop on Machine Learning in Medical Imaging (pp. 269-278). Cham: Springer Nature Switzerland.

[6] Xiong, Y., Peng, H., Xiang, Y., Wong, K. C., Chen, Q.,

Yan, J., & Tang, B. (2022). Leveraging Multi-source knowledge for Chinese clinical named entity recognition via relational graph convolutional network. Journal of Biomedical Informatics, 128, 104035.

[7] Li, J., Wei, Q., Ghiasvand, O., Chen, M., Lobanov, V., Weng, C., & Xu, H. (2022). A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. BMC medical informatics and decision making, 22(3), 1-10.