中文电子病历命名实体识别的研究与进展

杜晋华[1],尹 浩[1],冯 嵩[2]
(1. 清华大学北京信息科学与技术国家研究中心,北京 100084; 2. 中南大学湘雅医院,湖南长沙 410008)

摘 要: 海量电子病历(Electronic Medical Record,EMR)数据是支撑医疗智能化研究的重要原料,然而电子病历文本数据的半结构化甚至无结构化特点,造成后续对其分析利用的极大困难. 虽然近年来基于深度学习的命名实体识别(Named Entity Recognition,NER)成为对电子病历进行自动化信息抽取的核心技术,但鉴于中文电子病历(Chinese Electronic Medical Record,CEMR)具有包括病历文本的非规范性与专业性、医疗实体的独特性和标注语料的稀缺性在内的独特文本数据特征,该研究目前仍存在诸多挑战.

本文对中文电子病历命名实体识别的研究与进展进行了综述,系统梳理了命名实体识别的概念、相关理论模型以及制约中文电子病历命名实体识别准确率和识别效率的主要原因;从技术发展角度详细分析了中文电子病历命名实体识别方法的变革历程;并对中文电子病历命名实体识别效果做了实验验证与深入分析,指出了现有模型的不足与改进方向.

鉴于国内近年来与中文信息学处理相关的测评会议CCKS持续关注中文电子病历命名实体识别,本文特别对CCKS在该领域五年来的全部代表性测评论文做了纵横对比分析,并通过在主流模型上的深入实验与研究,为后续该领域的继续推进寻求了思路.

关键词: 中文电子病历; 命名实体识别; 深度学习; 预训练模型; 自然语言处理; 医疗信息化
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2022)12-3030-24
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220485

Research and Development of Named Entity Recognition in Chinese Electronic Medical Record

DU Jin-hua[1], YIN Hao[1], FENG Song[2]
(1. *Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing* 100084, *China*;
2. *Xiangya Hospital of Central South University, Changsha, Hunan* 410008, *China*)

Abstract: Massive electronic medical record(EMR) data is an important raw material to support the research of medical intelligence, but the semi-structured or even unstructured characteristics of EMR text data make it extremely difficult to analyze and utilize them subsequently. Although named entity recognition(NER) based on deep learning has become a core technology for automated information extraction from electronic medical records in recent years, there are still many challenges in this research given the unique textual data characteristics of Chinese electronic medical record(CEMR), including the non-normative and specialized nature of medical record text, the uniqueness of medical entities and the scarcity of annotated corpus.

This paper provides an overview of the research and progress of named entity recognition in Chinese electronic medical records, systematically sorting out the concept of named entity recognition, related theoretical models and the main reasons limiting the accuracy and efficiency of named entity recognition in Chinese electronic medical records; analyzes in detail the change history of named entity recognition methods in Chinese electronic medical records from the perspective of technical development; and makes an experimental verification and in-depth analysis of the effect of named entity recognition in Chinese electronic medical records, and points out the shortcomings and improvement directions of existing models.

In view of the fact that CCKS, a domestic evaluation conference related to Chinese informatics processing, has contin-

收稿日期:2022-04-29;修回日期:2022-08-25;责任编辑:朱梅玉
基金项目:国家重点研发计划(No.2020YFC2005003);国家自然科学基金(No.92067206,No.61972222,No.62102217)

ued to focus on the recognition of named entities in Chinese electronic medical records in recent years, this paper presents a longitudinal and cross-sectional analysis of all the representative evaluation papers of CCKS in this field over the past five years, and seeks ideas for the continued advancement of this field through in-depth experiments and research on the mainstream model.

Key words: Chinese electronic medical record(CEMR); named entity recognition(NER); deep learning; pre-trained language models; natural language processing; medical information technology

1 引言

电子病历(Electronic Medical Record,EMR)是指医务人员在医疗活动过程中,使用信息系统生成的文字、符号、图表、图形、数字、影像等数字化信息,并能实现存储、管理、传输和重现的医疗记录,是病历的一种记录形式,包括门(急)诊病历和住院病历[1],是临床辅助决策[2]、专病科研数据提取[3]、医疗知识图谱构建[4]和智能预问诊[5]等应用的重要数据支撑. 然而,电子病历通常由自然语言书写而成,大多为医疗信息系统无法直接利用的半结构化甚至无结构化数据[6]. 如何利用自然语言处理技术对电子病历文本进行智能分析和信息抽取,将其组织为结构化的内容,是当前研究的重点[4].

如图1所示,命名实体识别是电子病历分析利用过程中介于数据预处理与数据应用之间的关键技术. 基于对电子病历结构化和标准化的目的,针对电子病历的命名实体识别(Named Entity Recognition,NER)是从海量电子病历数据中识别出有独立或特定意义的医疗信息实体[7],如目前公认的疾病和诊断、检查、检验、手术、药物与解剖部位在内的六类实体[8],对其进行序列标注和标准化,为进一步进行信息抽取和文本挖掘做准备. 该技术具有重要的应用前景. 截至目前,电子病历的命名实体识别方法主要经历了基于词典、规则和机器学习的三个发展阶段. 相较基于词典的方法兼容性较差和基于规则的方法可迁移性较差,基于机器学习的方法在电子病历命名实体识别上表现出较好的实用性和可移植性. 特别是在深度学习技术提出后,面向电子病历命名实体识别的深度学习模型呈井喷式增长,各个模型不断优化命名实体识别的准确性.

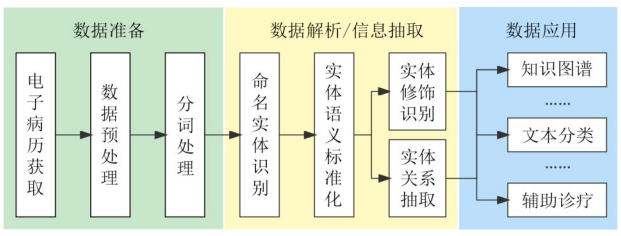

图1　电子病历分析利用流程[9]

国际上,早在1996年由NCCOSC(前NOSC)海军研究与发展小组(NRaD)的Beth Sundheim组织的MUC-6会议[10]提出命名实体识别概念就开始推动相关方面的研究. 2002年自然语言处理领域影响力最大的国际学术组织ACL下属的SIGNLL主办的计算自然语言学会议(Conference on Natural Language Learning,CoNLL)(https://www.clips.uantwerpen.be/conll2002/ner/)将跨国语言的命名实体识别作为共享任务. 2010年美国国立卫生研究院(NIH)赞助的国家生物医学计算项目"Informatics for Integrating Biology and the Bedside"(I2B2)测评任务给出电子病历命名实体识别的具体要求,聚焦推进英文电子病历命名实体识别方面的研究.

除建立词典和应用人工规则的识别方法之外,早期主要的识别方法的训练模型几乎都基于监督学习,包括采用贝叶斯模型、支持向量机[11]、条件随机场[12]等. 后续的研究中发现,半监督学习方法有别于有监督学习,只需要少量语料标注,因此也成为一段时间的研究热点,包括采用半监督协同训练[13]和多任务学习的半监督学习方法[14]等.

随着深度学习技术的发展,其由于在命名实体识别上表现优异,迅速成为研究热点. 从最初以LSTM[15]为代表的单向RNN网络到以BiLSTM[16]为代表的双向RNN网络,从基本的CNN网络[17]到其变种迭代膨胀卷积IDCNN[18],从类似CRF这样的单一模型到诸如BiLSTM+CRF[19]的多模型融合……人工参与工作量不断减少,识别精度也不断提高.

特别在将预训练模型和迁移学习方法引入后,模型对语义的理解更进一步,具体是通过自监督学习从大规模语料中获得与后续任务无关的预训练模型,并迁移到实体识别这样的下游语言任务上. 比如从Word2vec[20]到GLOVE[21],再到BiLSTM,BERT[22],以及以RoBERTa[23]为代表的BERTology系列……这些预训练模型依次出现,在优化升级过程中不断提高命名实体识别的精度.

而国内由于医疗信息化建设起步较晚,电子病历命名实体识别研究相对于英文语料环境落后. 最早杨锦锋等人[7]在2014年对国内外电子病历命名实体识别工作做了详尽总结,在2016年制定了命名实体的详细标注规范[24]. 此后国内在该领域的研究逐步展开. 比如从2017年至今每年举办的全国知识图谱与语义计算大会[25~29]均将中文电子病历命名实体识别作为测评任

务,迅速推动了该领域的研究进步.

其中,面向中文电子病历(Chinese Electronic Medical Record,CEMR)命名实体识别的主要技术路线和国外大致相同,主要在待识别文本的语言特征上两者有所差异,如英文词语边界明显、词语前后缀较易划分、词法句法结构相对固定,而中文语句没有明显的分词、偏旁部首等部分不能直接划分、词法句法结构复杂. 特别是针对医疗领域,中文医学专业词汇多,医学命名实体长,一词多义、多词一义以及词汇缩写无统一规范等问题尚未获得有效解决. 许多研究者基于国外提出的模型技术,融合中文医疗文本特征,在不断摸索提高中文电子病历命名实体识别准确性的有效方法,具体研究在CCKS历年收录的文章(详见第4节)中进行了说明.

虽然面向中文电子病历的命名实体识别目标明确,相关技术也取得了长足发展,但有别于英文或者中文通用领域的命名实体识别,中文电子病历独特的文本数据特征也给该研究带来了诸多挑战,具体包括以下几点.

(1)中文电子病历文本的非规范性和专业性. 该特征带来了三方面挑战:一是中文电子病历文本中存在大量非规范的语法、拼写错误和不完整的句子结构,如将“右心室”错误地写为“有心室”;二是中文电子病历文本包含大量专业术语、受控词汇、缩略语、符号等,如药物“Aspirin”被译作“阿司匹林”或者“阿斯匹林”中哪一种并不确定;三是中文电子病历自身特殊的文法和句法. 这些挑战均给命名实体识别造成困难.

(2)中文医疗实体的独特性. 中文电子病历文本数据中不仅有常规的实体,还有很多拥有复杂结构的实体,主要有两种情况:一是嵌套类实体存在自身复杂的结构,如“呼吸中枢受累”中存在二级实体嵌套,即“呼吸中枢受累”为症状而“呼吸中枢”为身体部位;二是跳跃类实体在文本中的位置不连续,如“尿道、膀胱、肾绞痛”中存在三个非连续实体“尿道痛”“膀胱痛”和“肾绞痛”.

(3)中文电子病历标注语料的稀缺性. 造成这一现象的原因主要是考虑到患者隐私和保密性要求,电子病历数据难以公开;此外可用于电子病历命名实体识别的数据集标注成本高,需要医疗专家的指导和参与,费时费力.

鉴于此,本文对国内外在中文电子病历医疗命名实体识别上的工作进行了详细分析;综述了近年来中文电子病历命名实体识别模型上的研究进展;同时也对当前电子病历命名实体识别的效果进行了对比检验,进而深入分析了各模型的优势与不足;在此基础上对该领域的后续研究方向进行了展望.

2 中文电子病历命名实体识别

中文电子病历命名实体识别是针对给定的一组电子病历纯文本文档,通过自然语言处理技术,识别并抽取出与医学临床相关的实体提及,并将它们归类到预定义类别[8]. 如全国知识图谱与语义计算大会(CCKS)于2021年发布的中文电子病历命名实体识别评测任务[8]中定义了6类实体,包括疾病和诊断、检查、检验、手术、药物和解剖部位. 其一般流程包括先将原始电子病历语料进行数据抽取、清洗、规约与脱敏四步预处理,获得待标记的电子病历字符序列. 之后将其输入命名实体识别模型中进行计算,获得标注好的电子病历字符序列作为最终结果. 具体到命名实体识别模型,通常由特征工程、识别方法所对应的模型识别和模型融合三部分构成,如图2所示.

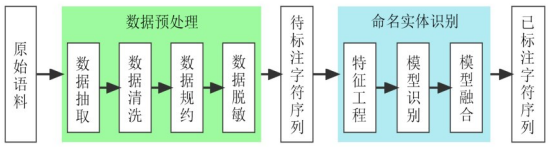

图2 中文电子病历处理流程

2.1 原始语料来源

原始语料是指经准确标注后用于训练目的的电子病历文本数据. 作为命名实体识别的数据来源,其重要性不言而喻. 特别是基于机器学习的方法非常依赖原始语料的标注质量,通常直接利用原始语料训练模型和检验模型学习效果.

但由于电子病历涉及患者隐私信息,通常公开获取原始语料的难度较大,而且对电子病历的标注需要专业人士花费大量时间完成,成本较高. 现存公开的中文电子病历标注数据十分稀缺,主要通过组织相应测评任务来促进有关方面的研究. 比较典型的测评语料库有三个.

(1)N2C2:National NLP Clinical Challenges 共享测评任务用语料库(https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/). 该任务提供了大部分最初由前美国国立卫生研究院(National Institutes of Health,NIH)资助的美国国家生物医学计算中心(National Center for Biomedical Computing,NCBC)所创建的数据集,称为整合生物学和临床的信息学(Informatics for Integrating Biology and the Bedside,I2B2). I2B2/N2C2测评从2006年到2022年已发展近20年,其任务变更和数据集的扩充修改充分反映了国际医疗自然语言处理的技术发展情况.

(2)CCKS:中国知识图谱与语义计算大会组织的中文医疗命名实体抽取测评任务CNER用语料库(http://sigkg.cn/ccks2021/?page_id=27). 自2017年以来,每年CCKS都会和医疗领域的AI企业开展合作,由企业提供

数据,并通过专业的医学团队对数据进行人工标注.

(3)CHIP:中国健康信息处理大会组织的中文临床信息处理任务用语料库(http://www.cips-chip.org.cn/).CHIP自2017年起已连续举办5届,聚集了全国顶尖的医疗信息处理学者与医疗专家,以及开展智慧医疗研究的各个单位院所.其2021年发布的中文医疗信息处理挑战榜CBLUE(https://tianchi.aliyun.com/specials/promotion/2021chinesemedicalnlpleaderboardchallenge)目前已成为业界包括中文电子病历命名实体识别在内多项任务的测评基准.

2.2　数据预处理

在获得海量电子病历数据的原始语料后,需要根据数据特性和结构化要求对数据进行预处理.电子病历数据预处理的步骤依次为以下几步.

(1)数据抽取:将不同来源的电子病历数据集成到同一个数据库中,在统一数据格式的基础上扩大数据规模.通常也需要在该步骤完成部分标准化工作,以方便模型的训练和后续算法的研究.如王正宏[30]针对目前各类医疗信息系统间数据难以共享的"数据孤岛"现象,逐一解决数据结构不统一、数据标准不统一和数据共享效率低等问题,有效实现了区域健康医疗数据的集成.

(2)数据清洗:对异常数据进行处理,包括对重复数据进行去重;对缺失数据进行删除样本或均值填补;对噪声数据(明显不正常的数值)采取平滑处理或异常值分析;对语法错误、格式错误(字母大小写、平半角等)、前后数据不一致或不统一等问题依靠人工或者算法修正.如韩丽珍[31]调研肿瘤科出院病历中存在信息不完整等问题的病历缺陷率高达5.65%,极大降低了数据质量.

(3)数据规约:在尽可能保持数据原貌的前提下,对原始语料进行选择与降维,最大限度精简数据量,剔除无关数据,以筛选出适合不同医疗研究目标的数据.如邱炎龙[32]利用命名实体识别技术从电子病历中仅识别并抽取导致心血管疾病的风险因素,用于心血管疾病的预测,模型最优性能F值达到0.958 6.

(4)数据脱敏:由于电子病历固有的隐私属性,需要在正式标注前对从医院或医疗机构搜集到的数据中敏感信息进行隐藏,以保护患者的隐私权,同时方便后续对电子病历数据的安全、有效利用.如余健等人[33]针对海量的中医药数据提出一种高效的基于属性的内积加密数据脱敏算法,方便医疗数据脱敏后的分析处理.

2.3　标注结果

通过对电子病历进行命名实体识别,本质上完成对医疗数据序列的标注,最终抽取出指定类别的医疗实体.标注结果一般是由实体所属类别、实体在序列中的起始位置、实体在序列中的结束位置构成的三元组.

标注所用标签通常有两种,分别是BIO和BIOES.各字母缩写分别代表了实体起始位置(Begin,简记为B)、实体中间位置(Intermediate,简记为I;或使用Medium表示,简记为M)、实体结束位置(End,简记为E)、单个字符(Single,简记为S)、其他无关字符(Other,简记为O).以使用BIO(Begin,Intermediate,Other)标注表示方式对数据集进行字符级别标注为例,标注结果如表1所示.

表1　标签序列举例

字符	标注结果	字符	标注结果
查	O	两	B-疾病和诊断
胸	B-检查	肺	I-疾病和诊断
部	I-检查	感	I-疾病和诊断
C	I-检查	染	I-疾病和诊断
T	I-检查	性	I-疾病和诊断
、	O	病	I-疾病和诊断
考	O	变	I-疾病和诊断
虑	O	,	O

2.4　评价指标

命名实体识别的可量化评价指标有3个,分别是准确率(Precision,简记为Prec)、召回率(Recall,简记为Rec)和F_1-Measure(简记为F_1)值.其中,准确率衡量命名实体识别模型正确识别实体的能力,召回率衡量命名实体识别模型识别整个语料库中全部实体的能力,F_1取两者的调和平均值.

设模型正确识别的相关实体数为T_p,模型错误识别的不相关实体数为F_p,模型未识别的相关实体数为F_N.则

$$\mathrm{Prec}=\frac{T_p}{T_p+F_p}\times 100\% \tag{1}$$

$$\mathrm{Rec}=\frac{T_p}{T_p+F_N}\times 100\% \tag{2}$$

$$F_1=\frac{2\times \mathrm{Prec}\times \mathrm{Rec}}{\mathrm{Prec}+\mathrm{Rec}}\times 100\% \tag{3}$$

由于命名实体识别一般涉及多个待识别实体类型,通常需要评估模型对所有实体类型的识别性能.一种方法是求宏观平均F_1值,即首先独立计算模型识别每种实体类型的F_1值,然后取平均,表示对所有实体类型同等看待;另一种方法是求微观平均F_1值,即直接将所有类型的识别结果统一求F_1值,表示对所有实体同等看待.相较而言,微观方法更容易受到语料库中实体数量较多的实体类型识别质量的影响[34].

同时,根据对识别精度要求的不同,评价指标又具体可分为严格匹配指标和宽松匹配指标两种.严格匹配指标要求命名实体识别模型兼顾实体边界和类型的识别,两者都正确时才记录到T_p中;宽松匹配指标[35]仅

要求命名实体识别模型正确识别实体类型，无论实体边界是否都准确记录到T_p中．更为复杂的评价方式还有ACE[36]等，但没有获得广泛应用．

值得一提的是，除2.1节中提到的相关测评用语料库支持通过手动计算严格匹配指标和宽松匹配指标对模型效果进行测评外，也有一些在线测评平台支持采用这些指标自动对模型进行测评．如在通用命名实体识别领域，中文语言理解测评基准CLUE（https://www.cluebenchmarks.com/ner.html）提供了对命名实体识别模型的测评[37]，截至2022年11月3日最优的模型获得的F_1值为83.351；在医疗命名实体识别领域，中文医疗信息处理挑战榜CBLUE（https://tianchi.aliyun.com/specials/promotion/2021chinesemedicalnlpleaderboardchallenge）提供了对医学文本实体识别模型的专用测评，截至2022年1月10日最优的模型在CBLUE 1.0榜单上获得的F_1值为70.617，人类标注获得的F_1值为67.0，模型效果已超越了人类水平[38]．

3 中文电子病历命名实体识别模型

电子病历命名实体识别模型的研究，主要有基于词典、规则和机器学习三种方法，各方法的优缺点如表2所示．

表2 电子病历命名实体识别方法分类

一级分类	二级分类	优点		缺点	模型示例
基于词典		实现简单		·词典的规模和质量对识别结果有重要影响 ·难以保证及时更新对新增或者补充实体信息的兼容和覆盖	DLAM[39]
基于规则		便于维护		·需要大量人力和时间成本投入 ·规则的可移植性较差	EdIE-R[40]
基于机器学习	统计机器学习	·实用 ·可移植	—	·需要有大规模高质量的标注数据集作为训练原料	SVM[11] HMM[41] ME[42] CRF[43]
	深度学习		准确		RNN[15] CNN[44] BERT[45]

3.1 基于词典的方法

基于词典的方法需要构建全面覆盖医疗领域的中文医学术语大全或医学专用词典，并结合相应的匹配算法完成对电子病历命名实体的识别．其中，词典（有标注语料）的规模和质量对识别的结果有重要影响．该方法在处理中文电子病历中所包含的大量专业术语、受控词汇、缩略语和符号类实体的识别时效果良好．

曲春燕等人[46]针对中文电子病历命名实体语料标注空白的现状，提出了标注语料库的构建方案．杨锦锋等人[24]在医生的参与和指导下，构建了规模较大、质量较高的标注语料库．考虑到Brat、Jieba分词模块和SnowNLP这样的文本标注工具不能有效支持电子病历的标注，刘一斌[47]课题组开发了电子病历命名实体手工标注工具，在标注自动化方面进行了初步探索，而标注质量核检方面仍有待研究．

虽然目前提出的方法都让词典规模和质量的可靠性有所提高，但由于很多实体对应的缩写、同义词（不同表达方式）等补充内容难以全部一次性和实体同时加入词典，而且词典无法实时包含医学领域不断增加的新实体，因此基于词典的方法难以保证及时更新对新增或者补充实体信息的兼容和覆盖．

随着基于规则和机器学习方法的提出，基于词典的方法更多作为其他方法的特征输入与之融合，借以提高相应方法在电子病历命名实体识别上的效果．如陈曙东等人[48]在序列建模前通过动态词典匹配的语义来增强命名实体识别效果；Wang等人[49]将字典合并到深度神经网络中，有效改善了单独神经网络模型通常无法处理稀有实体和未见实体的情况．

在与其他方法融合过程中，鉴于医疗实体名称的特殊性，词典的构造对目前的研究至关重要，有必要借助大量外部词典资源．这些资源通常来自医院和医疗机构的清单和医学文献，如《人体解剖学名词（第二版）》（*Chinese Terms in Human Anatomy*［Second Edition］）[49]．如何有效治理医院和医疗机构的历史数据，并将其转化为可利用的词典信息，是目前医疗大数据治理与挖掘的研究难点．

3.2 基于规则的方法

基于规则的方法不同于基于词典的方法，需要首先对待处理的电子病历文本进行分析并构建规则模板，之后在同类型文本上使用规则模板，通过模式匹配的方式实现命名实体识别．该方法一方面可以对中文电子病历文本中大量存在的非规范语法、拼写错误和不完整的句子结构进行规则修正；另一方面也可以通过规则模板的设计应对拥有特殊文法和句法结构的中文电子病历文本上的命名实体识别．

虽然基于规则的方法直观且便于维护，一定程度上弥补了基于词典的方法对未收录词无法识别的缺陷，可以应对中文电子病历文本的非规范性，但建立统一完整的识别规则库仍需要大量人力和时间成本投入，且规则的可移植性较差，基于特定电子病历文本构建的规则模板可能无法适用于其他电子病历文本上的命名实体识别．另外，受如不同医生的语言习惯或表达

方式不同所造成的语言结构本身不确定性的影响,指定统一完整的规则难度较大.

和基于词典的方法类似,目前也少有研究者单独使用基于规则的方法完成电子病历命名实体识别,多将规则和词典结合到一起辅助机器学习方法获取电子病历文本特征,通过方法融合提升机器学习方法的效果. 如Chen等人[50]通过在机器学习模型外附加规则来提取模型无法识别的实体;Gorinski等人[40]通过对比实验证明基于规则的方法可以非常有效地进行医疗实体识别并进一步提高机器学习方法的准确率.

在未来一段时间内,如何将规则更好地与机器学习方法融合,以及提高规则方法的可迁移性、降低其成本投入,仍将是值得重点关注的研究问题.

3.3　基于机器学习的方法

基于机器学习的方法利用标注过的语料进行模型训练,再利用模型完成对命名实体的识别,相较基于词典的方法和基于规则的方法,具有更好的实用性和可移植性. 它不仅可以较好地处理中文电子病历文本的非规范性和专业性造成的命名实体识别困难,而且在特殊医疗命名实体识别上表现优异.

通常该方法所构建的模型会对原始语料进行不同粒度特征的提取,如字特征提取和上下文特征提取等,模型框架如图3所示.

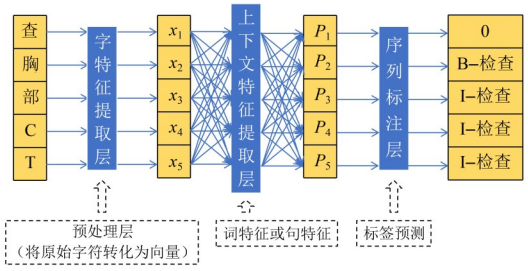

图3　基于机器学习方法的中文电子病历命名实体识别模型框架

不同研究者分别就各层提出了非常多切实可行的方法,并取得了良好的识别效果(见表3),本节将对各方法做具体介绍和分析.

表3　机器学习方法汇总

层次	方法
字特征提取层	BERT[45],BERTology[23],ELMo[51],ME[53],MEMM[42],HMM[41],SVM[11]
上下文特征提取层	LSTM[15],BiLSTM[16],Lattice-LSTM[52],RNN[54],CNN[44]
序列标注层	CRF[43]

3.3.1　传统机器学习/统计机器学习

传统机器学习包括有监督学习、半监督学习和无监督学习三类,在电子病历命名实体识别中大多采用有监督的机器学习模型:将命名实体识别看作分类任务,利用大规模带标签的训练集进行模型训练,再利用训练好的模型对未标注的原始语料进行实体识别. 其关键问题是如何从电子病历文本中提取各种有效的词法、句法和语义特征,然后利用序列标注模型进行医疗命名实体的识别. 这些模型如图4所示.

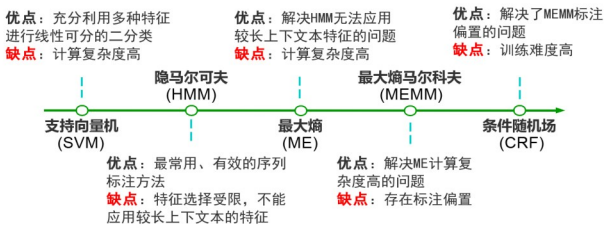

图4　传统机器学习方法发展历程

(1)支持向量机(SVM)[11]:利用高维特征空间将识别问题转化为线性可分的二分类问题. 通过构造最优分割超平面,确保两类样本数据之间的间隔最大化,以训练出可信的分类器,对解决小样本、线性不可分和高维度模式识别均有显著作用. 优点在于能够充分利用词法、句法和上下文等多种特征,缺点是识别效率低,需要依托大量数据进行训练,单独使用SVM效果不佳.

(2)隐马尔可夫(HMM)[55]:一种有向概率图模型. 它利用已知的观测序列,通过求解该序列和可能的标记序列的联合概率,来推断最优的标记序列. 高冰涛等人[41]从HMM出发,结合权值学习构建迁移学习模型,显著降低生物医学命名实体识别对目标领域标注数据的需求,较其他方法获得了更好的识别性能. 对比来看,虽然HMM是序列标注最常用和最有效的方法之一,但由于其输出依赖独立性假设,即观测元素独立于观测序列的其他元素,导致无法考虑长距离依赖(上下文的特征),不能够真实描述数据序列所包含的信息,一定程度上限制了特征的选择.

(3)最大熵(ME)[53]:基于最大熵原理,在满足约束条件的情况下,选择熵最大(不确定性最大、信息量最大)的概率分布模型. ME相较HMM不必满足独立性假设,同时可以灵活引入特征以提高模型的准确率,结构严谨,良好通用. 但ME迭代过程计算量巨大,计算时间复杂度高.

(4)最大熵马尔可夫(MEMM)[56]:为解决隐马尔科夫问题,在最大熵的基础上,提出MEMM,改变HMM中观测和隐藏状态之间的依赖关系,并在计算条件概率时采用ME直接建模. 如杨丽静等人[42]使用MEMM识别恶性肿瘤医疗文本数据取得了可供商用的识别准确度. 但MEMM并不完美,仍存在标注偏置[12]的问题.

(5)条件随机场(CRF)[12]:是遵循马尔可夫性的无

向概率图模型，也是典型的判别式模型，克服了HMM的独立性假设条件，并通过使用全局归一化函数解决了MEMM标注偏置的问题．同时，CRF通过给句子中的字符在最终预测标注上添加有效约束，解决了下文提到的BiLSTM等模型产生无效标注的问题，比如采用BIO标签方案时I一定在B后面，不同类别的I不会直接相连等．假设给定字符序列x和其对应的标签序列y，x的所有可能标签为$\omega(x)$，模型参数为θ，团势函数为$\varphi(y_i,x_i,\theta)$，则对于给定字符序列x，标签序列为y的概率为

$$P(y|x;\theta)=\frac{\prod_{i=1}^{n}\varphi(y_i,x_i,\theta)}{\sum_{y\in\omega(x)}\prod_{i=1}^{n}\varphi(y_i,x_i,\theta)} \tag{4}$$

损失函数的形式为

$$\text{Loss}=-\log\frac{P_{\text{realpath}}}{\sum_{i}^{N}P_i} \tag{5}$$

其中，N表示所有可能的路径，realpath表示其中真实的路径．

虽然CRF的收敛速度慢、训练难度高，实践中预测结果可能会出现头尾缺字或者多字等现象，也无法应对实体重叠的问题，但其优越性使其至今仍是最受欢迎的电子病历命名实体识别方法之一，其通常被用作整个机器学习模型的结束部分．如张华丽等人[43]构建的注意力Attention-BiLSTM-CRF模型最后将数据流输入CRF层中提取医疗命名实体；李博等人[57]构建的Transformer-CRF模型最后利用CRF对提取到的医疗文本特征进行分类识别．

3.3.2 深度学习

近年来，随着机器学习技术的发展，基于深度学习的命名实体识别也获得较大关注，并取得了很好的识别效果．相较统计机器学习需要依赖研究者手动设计特征工程，即用一系列工程化的方式从原始语料中筛选出更好的文本数据特征，以提升模型的训练效果，深度学习是端到端的，可以自动找到更深入、更抽象的特征．深度学习的关键在于如何在词向量的基础上设计并利用各种神经网络模型进行医疗命名实体识别．普遍采用的模型如图5所示．

图5 深度学习方法发展历程

(1)卷积神经网络(CNN)[58]：通常用于从文本中提取有用的语义特征以辅助实体边界划分．其强大的特征捕捉能力使得句子在建模过程中，经过反复组合下层邻近信息并向上传递，可以学习到相距较远的句子之间存在的联系．Santos等人[44]利用CNN来捕获词序列的语义信息，准确识别出医疗命名实体．曹春萍等人[17]通过设定CNN不同的卷积窗口大小，捕获更丰富的边界特征信息．CNN变种迭代膨胀卷积IDCNN(Iterated Dilated Convolution Neural Network)[18]在医疗命名实体识别上取得了更好的识别效果．

(2)循环神经网络(RNN)[15]：通过各神经元之间的参数共享，可以处理任意序列长度的文本信息，但存在梯度消失和梯度爆炸等问题，其变种门控循环单元GRU(Gate Recurrent Unit)和长短时记忆网络LSTM(Long Short Term Memory network)对其进行了改进．

GRU[59]在RNN基础上引入门机制，通过更新门和重置门平衡过去和当前模型输入对输出的影响，在神经网络结点间动态地建立快捷连边，避免了RNN梯度消失的问题．

LSTM主要用于文本分类[60,61]，不仅能学习序列关系，还能够避免长期依赖问题，有效缓解了RNN梯度消失的问题，解决了RNN中无法捕捉长距离依赖关系的不足．通过对细胞状态中旧信息的遗忘和新信息的记忆，使对后续时刻计算有用的信息被传递、无用的信息被丢弃，并在每个时间步输出相应的隐层状态，其中遗忘、记忆与输出由上一时刻的隐层状态和当前输入计算出来的遗忘门、记忆门以及输出门来控制．

对RNN的进一步优化提升从并行计算、信息获取、精度和中文语种四方面展开，具体如下．

并行优化方面，由于LSTM和GRU在计算速度上比较慢，考虑将其改进为GPU优化的CuDNNLSTM和CuDNNGRU．通过手动计算mask的方法，可大幅提高训练速度．

信息获取方面，鉴于GRU和LSTM只能获取单向

的信息,无法同时通过上下文来编码当前词汇的语义特征,与普通循环神经网络RNN不同的双向循环神经网络BiRNN(包括BiLSTM[16]和BiGRU[54]等)被提出.它采用正向网络从前到后计算的同时,还采用反向网络从后到前计算,因此BiRNN可以在任意时刻同时获取前后向的信息,避免人工构造大量特征,获得比单向RNN更强的表达能力,完成对上下文信息的建模,更准确地实现对句子中逐字符的分类,即命名实体的识别.

精度提升方面,虽然BiLSTM在诸如词性标注这类独立的序列标注任务中取得了成功,但由于模型忽略了标签间的依赖关系,所以在命名实体识别任务上会导致部分实体识别误差,通常将BiLSTM与CRF组合使用[62],用CRF学习标签间的关系,解决标签依赖的问题.

中文语料方面,考虑将中文词汇信息加入训练的模型中,点阵长短时记忆网络Lattice-LSTM[52]将链式结构转成图结构,同时用多出的节点记录外部词典信息.通过训练更新权重,Lattice-LSTM一方面将词汇信息加入模型,另一方面避免了分词错误造成的误差传播,通过同时使用单词本身和单词序列信息,有效提升了命名实体识别的性能.但Lattice-LSTM存在不可并行和信息损失等问题.考虑到Transformer采取全连接的自注意力机制可以很好捕捉长距离依赖,且自注意力机制对位置无偏,Li等人[63]通过引用位置向量来保持位置信息,将位置嵌入Lattice结构中,利用相对位置编码解决实体边界识别,通过词向量编码解决实体类型识别,采用Transformer结构实现并行化,从而较Lattice-LSTM进一步提高命名实体识别的速度和准确率.

(3)预训练(BERT)[22]:预训练的深度双向Transformers语言理解模型,由12层网络构成,隐藏层维度为768,含12个头,总参数量达110 M,是对海量语料进行无监督学习得到的预训练语言表征模型,由Google AI团队在2018年10月提出,应用于自然语言处理的各个领域[64~66].

BERT是可供其他模型迁移学习的一个模型,利用遮蔽语言模型、双向Transformer和句子级别的负采样,充分描述字符级、词级、句子级、语句间的关系的特征.在电子病历命名实体识别中,对BERT微调后可作为特征提取器,将提取的特征作为该任务的词嵌入特征,以融入下游任务中.专门为中文设计的BERT-wwm模型(https://github.com/ymcui/Chinese-BERT-wwm)可选择进行文本的字特征表示,由24层网络构成,隐藏层维度为1 024,含16个头,总参数量达330 M.

像BERT这样的预训练模型,在中文电子病历标注语料极度稀缺的情况下,通过大规模无标注数据的预训练,保证了模型训练效果,有效应对了在独特的中文电子病历上进行命名实体识别的难题.BERT被提出前,预处理模型采用传统的词向量方法,如Word2vec[20]和GLOVE[21]等,由于词向量与上下文无关,所以在一词多义等情况下无法建模.ELMo[51]模型利用双向LSTM进行预训练,得到与上下文依赖的词表示,解决了传统词向量存在的问题,但ELMo结构仅由两个LSTM简单拼接而成.随后出现的OpenAI GPT[67]模型和BERT模型利用Transformer取代了LSTM,获得较ELMo更好的性能.与BERT相比,OpenAI GPT只能使用从左到右受限制的Transformer,BERT模型展示出对文本上下文语义信息的更好提取.

近年来,人们以BERT模型为基础,对其进行了结构调整、性能优化、再训练等,产生了更多在专业领域表现更佳的BERTology模型,这些模型正逐渐变成医疗命名实体识别的新研究重点.如WU等人[23]使用RoBERTa进行电子病历命名实体识别的研究,其识别准确度优于传统BERT.朱岩等人[68]基于RoBERTa-WWW对中文电子病历命名实体进行识别,通过全词掩码机制,进一步提升了F_1值.刘司宇等人[69]在RoBERTa-WWW的基础上,改进其潜在的局部空间特征无法提取的缺陷,通过将膨胀卷积网络DCNN与RoBERTa-WWW集成学习,将F_1值优化到95.79%.

当前,BERT在命名实体识别问题上潜在的研究趋势有二:受到该模型对输入文本长度的限制,一种方式是将BERT视作一个字符嵌入生成器,将分割后的等长文本输入BERT计算其对应的字符嵌入,与字音、字形等其他特征融合输入其他模型中进行电子病历的命名实体识别,如乔锐等人[45]提出一种基于BERT和字符特征融合的医疗命名实体识别方法;另一种方式是从字粒度、词粒度和句粒度对原始语料进行变长分割后输入BERT模型,以保证输入文本特征的完整性.

3.3.3　特征工程

在实际使用机器学习方法前,尤其是传统机器学习方法,几乎所有研究者都会进行大量的特征定义,包括词性、拼音、词根、笔顺、偏旁部首以及词典特征等的定义,丰富有效的特征可以显著提升模型性能.无特征工程的方法在临床医疗文本的命名实体识别上并不完全适用,主要原因是目前还无法直接获得高质量、大规模的公开电子病历语料.

通常,对原始语料的特征工程设计依赖于语料自身的特性.中文电子病历的显著特征:一是中文是象形文字,相同的偏旁可能代表同一类实体,比如"月"字旁与解剖部位密切相关;二是医疗实体中有很多外来音译词,虽然读音相同但有不同的字形表示.基于此,Ma等人[70]将BERT获取的字特征和字音、字形特征相结合;晏阳天等人[71]将字音和字形特征经过charCNN深度学习网络和MaxPooling生成字embedding,与BERT结合

在一起作为BiLSTM的特征输入来预测相应的标签，这种构建特征工程的方式使F_1严格指标达到91.54%.

3.3.4 模型融合

机器学习中，利用多模型来完成学习任务被称作模型融合. 借助模型融合可在单模型基础上进一步提升模型表现. 目前在医疗命名实体识别研究领域，融合方法仍是主流，通常将多模型进行融合并搭配对人工规则的建模，如BiLSTM+CRF[19]，LSTM+CNN+CRF[70]等. 常用的模型融合策略有以下几种.

(1)堆叠：将电子病历命名实体识别任务抽象为若干个连续的子步骤，通过每一个子步骤采用不同的模型方法，实现整体提高识别准确率的目标. 如BiLSTM-CRF[19]就是使用CRF替换BiLSTM末尾的SoftMax层后将两者结合起来，既保留BiLSTM模型自动提取输入数据特征和双向非线性表示的优势，又通过CRF解决BiLSTM在预测时将相邻标注独立造成的低级识别错误，充分学习到标注之间的信息，更好地完成训练标注任务.

(2)投票：对来自不同模型的结果进行投票，如果预测结果的统计值高于某一人为设定的阈值，则保留该结果. 合理的阈值可以有效过滤掉假阳结果，确保被预测的结果具有较高的置信度，但也会丢失一些正确的结果. 如晏阳天等人[71]采用双阈值投票的方式来对不同的模型结果进行投票，通过设置高阈值来获取具有高置信度的结果，再利用该结果生成的词表来指导低阈值的投票过程，通过将低阈值结果中不属于高阈值词表的词汇滤除，保证了阈值变化后新增结果的质量，使融合后的模型效果取得较大的提升.

(3)混合：混合方式分为方法混合和模型混合两种. 方法混合是指为一定程度修正模型预测结果和实际的偏差，训练结束后在原有模型基础上混合使用一些启发式规则，即机器学习方法和规则方法的混合. 如晏阳天等人[71]提出将“连续的同类别实体合并为单一实体”“‘解剖部位’实体前面的方位词保留而后面的方位词舍去(‘右臂外侧’—>‘右臂’)”等规则融入模型中，使模型结果更符合实际. 而模型混合是指通过多模型混合，获得单一模型无法获得的准确度. 如殷章志等人[72]将基于字特征的Char-NER模型和基于词特征的Word-NER模型进行混合，可以在无人工特征参与下，提升F_1值到90以上；乔锐等人[45]将三个基本的融合模型BT(BERT-TimeDistributed_Dense)，BBT(BERT-BiLSTM-TimeDistributed_Dense)和BBC(BERT-BiLSTM-CRF)二次混合，同时将规则和词典混合到模型中，显著改善了最终识别医疗命名实体的效果.

4 中文电子病历命名实体识别效果

为实际考察目前中文电子病历命名实体识别前沿方法及其效果，为下一步研究提供方向，本节首先对CCKS近年来该领域相关论文中提及的方法进行比较，分析不同方法的特点和创新之处；再对这些方法中主流模型进行深入的实验分析，为后续研究提供切实可行、有借鉴意义的思路.

4.1 CCKS历年测评结果对比分析

CCKS测评从2017年至今历时五届，在中文电子病历命名实体识别任务上的研究获得显著成效. 对历年测评论文方法进行分析和结果比较，可以一定程度上反映近年来中文电子病历命名实体识别技术的发展和应用情况.

4.1.1 数据来源

CCKS2017，CCKS2019数据来源于北京极目云健康科技有限公司云医院平台的真实电子病历数据，CCKS2018，CCKS2020，CCKS2021数据来源于医渡云(北京)技术有限公司的专业医生团队整理和标注. 清华大学知识工程实验室、哈尔滨工业大学(深圳)和微软亚洲研究院等机构对测评的顺利开展提供了支持. 具体各年度所提供的训练和测试用数据情况见表4.

表4 2017—2021年CCKS标注数据集训练用/测试用实体数量统计

实体类型	年份				
	2017	2018	2019	2020	2021
疾病和诊断	722/553	0/0	2 116/682	4 345/1 866	4 345/1 834
影像检查	9 546/3 143	0/0	222/91	1 002/488	1 002/481
实验室检验	9 546/3 143	0/0	318/193	1 297/588	1 297/575
手术	0/0	1 329/735	765/140	923/404	923/406
药物	0/0	1 221/813	456/263	1 935/906	1 935/894
症状体征	7 831/2 311	0/0	0/0	0/0	0/0
解剖部位	10 719/3 021	9 472/6 339	1 486/447	8 811/3 849	8 811/3 861
症状描述	0/0	2 484/918	0/0	0/0	0/0
独立症状	0/0	3 712/1 327	0/0	0/0	0/0
治疗	1 048/465	0/0	0/0	0/0	0/0
总实体数	29 866/9 493	18 218/10 132	5 363/1 816	18 313/8 101	18 313/8 051
医疗记录数	300/100	600/400	1 000/379	1 200/300	1 050/450

由表4可知，在促进医疗领域实体识别初期，CCKS2017重点关注表征患者个人描述主观感受的症状、外部观察到病人的体征、医疗影像学检查、实验室检验、病人所患疾病、由症状识别出的诊断、治疗活动和

人体解剖学部位,较全面地囊括了医疗实体类型.CCKS2018考虑到症状类型实体多表现为结构化形式,将症状类型进一步细分为三类:解剖部位(复合症状的主体)、症状描述(复合症状的描述)及独立症状.同时,从CCKS2018开始不再进行治疗活动相关的实体识别,CCKS2019不再关注症状描述和独立症状两类实体的识别,并在CCKS2020和CCKS2021延续了相同的实体识别类型.

4.1.2　技术发展

五年来有近250支队伍参加了电子病历命名实体识别方面的测评,CCKS接收并发表相关论文30余篇.通过对各年度论文调研,梳理了各年度主要采用的实体识别方法和较上一年度的主要改进,对比情况见表5.在4.1.3节中,对所有文献做了详细概述与分析.

表5　CCKS 2017–2021电子病历命名实体识别任务概览

年份	任务	主要方法	改进	队伍数	论文数	最优F_1	来源
2017	电子病历命名实体识别	词典+BiLSMT-CRF	采用不同的特征向量构建方式提取中文文本的多维信息,提升模型识别效果	28	7	91.025	任务二[25]
2018	面向中文电子病历的命名实体识别	多粒度特征(词典等)+BiLSTM-CRF+后处理规则	通过利用大规模无标注数据进行模型预训练作为特征输入,提升整体模型识别效果	69	10	89.258	任务一[26]
2019	面向中文电子病历的命名实体识别	词典+预训练模型(BERT/ELMo)-BiLSTM-CRF	通过多模型融合进一步提升模型效果	44	5	85.620	任务一[27]
2020	面向中文电子病历的医疗实体及事件抽取	BERT-BiLSTM-CRF	通过对数据的不同学习利用方式提升模型效果	62	5	98.715	任务三[28]
2021	面向中文电子病历的医疗实体及事件抽取	BERT-BiLSTM-CRF	通过半监督学习、迁移学习等方式提升模型性能	51	3	76.840	任务四[29]

4.1.3　模型对比

CCKS2017所有测评论文中,基于深度学习和词典的方法被广泛使用,CRF和BiLSTM是被采用最多的模型;鉴于中文医疗实体识别较英文存在表达方式复杂和词语边界分割不够明显等问题,对中文文本数据进行不同粒度、不同编码方式、不同学习方法的特征向量构建成为研究重点.具体情况见表6.其中,PDET Feature表示与位置相关的实体类型特征(Position Dependent Entity Type Feature),SP指代私有共享多任务模型(Shared Private multitasking model),FT-BERT指代对BERT模型进行模型微调(Fine Tuning),LM是语言模型(Language Model)的简称,Att是注意力机制(Attention)的缩写,RD-CNN指代残差扩张(Residual Dilated)卷积神经网络.

CCKS2018所有测评论文中,CRF和BiLSTM模型仍旧是被采用最多的模型;基于医疗命名实体名称的特殊性,借助大量外部词典资源构造词典也成为命名实体识别的关键;所有的参赛队都进行了包括词性、拼音、词根、偏旁部首以及词典在内的大量特征定义;无特征工程的方法在临床医疗文本识别上表现不佳,潜在原因是训练过程无法获得高质量大规模的公开电子病历语料;以BiLSTM-CNN-CRF为例,搭配人工规则的融合方法仍然是模型构造的主流;部分研究开始聚焦通过ELMo或BERT等预训练模型从大规模无标注数据中提取信息,借以改进识别效果.具体情况见表7.其中,MSD表示包含字粒度和词粒度在内的不同粒度、不同信息的多粒度语义字典(Multi-granularity Semantic Dictionary),MT代表融合不同粒度语义信息的多模态树(Multimodal Tree),Re-entity表示二次利用实体信息.

CCKS2019所有测评论文中,CRF和BiLSTM模型仍然是被采用最多的模型,NER的主要实现分为序列标注和半指针半标注(即进行两次标注,分别标注实体的开始位置和终止位置)两种方式;借助大量外部词典资源进行词典构造仍然必要;预训练引入BERT模型或ELMo模型;基于特征工程与人工规则的混合模型是构造模型的主流.具体情况见表8.其中,BT模型是“BERT+时间分布密度(TimeDistributed_Dense)”的简称,BBT是BERT-BiLSTM-TimeDistributed_Dense的简称,BBC是BERT-BiLSTM-CRF的简称,UCNN是在图像语义分割任务中广泛使用的U-net[73]在文本序列分类任务上的改进模型,WaveNet是用于语音合成的自回归模型[74],self-Att是自注意力(Self Attention)的缩写,BERT-www是基于全词覆盖(Whole Word Masking)的中文BERT预训练模型[75]的简称.

CCKS2020所有测评论文中,BERT+BiLSTM+CRF是被广泛采用的模型组合;采用不同的模型融合方法进行识别效果的提升被多数研究者选用;借助外部词典资源和人工构造模型后处理规则仍然必要;针对影响中文医疗实体识别效果的标注错误、数据匮乏、模型可解释性差、数据利用不充分、召回率低等问题进行了多模型、多特征工程的尝试.具体情况见表9.其中,半监督训练简称为ST,对抗性训练简称为AT,后处理规

表6 CCKS2017模型方法对比

模型	方法	资源	改进	评价(*F*值)	文献
RNN-CRF	深度学习;规则	·百度百科和寻医问药网提供的医疗信息用以构建规则	·对中文文本数据进行了多维特征提取,包括n-gram、拼写特征、分词、词性、词头、字典特征、关系特征、词的分布表示、规则特征等 ·基于投票的方法将基于规则、基于CRF、直接基于RNN和基于特征的RNN方法的识别结果集成,充分利用大规模未标记数据,提升模型识别准确性	94.26	Hu等[76]
PDET Feature+BiLSTM-CRF	深度学习;词典	·一家上海三甲医院的11万条住院和门诊记录 ·搜狗词典 ·基于一家上海三甲医院数据自构造的分词词典	·考虑中文词边界分割不明显,将单词边界信息编码为模型输入特征 ·使用多级嵌入(字符级嵌入、词语级嵌入和字典特征级嵌入)作为BiLSTM输入	92.68	Lu等[77]
ELMo+BiLSTM-CRF-SP	深度学习;词典	·中国知网的医学文献 ·利用汉典网站(https://www.zdic.net/),获取语料库字汇表中每个字的笔画序列信息 ·利用搜狗输入法词典构建了药物词典	·在大规模无标注数据上使用双向语言模型预训练得到上下文相关且包含汉字内部结构信息的笔画ELMo向量作为输入特征 ·基于多任务学习构建神经网络模型,充分利用多个任务的相关性,进一步提升模型性能	91.75	罗凌等[78]
FT-BERT+BiLSTM-CRF	深度学习;词典	·网络爬取到中国临床文本数据 ·《药典》数据	·使用未标记的中文临床文本数据对BERT模型进行了预训练,可以利用未标记的领域特定知识 ·使用汉字偏旁部首信息作为输入特征提升预测效果	91.60	Li等[79]
LM-Att-BiGRU-CRF	深度学习	—	·融入双向语言模型提升识别效果 ·采用多头注意力机制提取文本中不同层次、不同类型的特征信息 ·以字为单位构建字向量,以应对词语边界模糊问题	91.34	唐国强等[80]
RD-CNN-CRF	深度学习;词典	·上海市曙光医院临床数据(收费项目和药物信息列表) ·一些医学文献,如(人体解剖学名词[第二版])	·提出残差扩张卷积神经网络,既能通过传统CNN捕获较短上下文信息,又能通过残差扩张CNN捕获较长上下文信息,实现与RNN类似效果,且模型结构支持异步计算,从而大大加快了训练速度	91.32	Qiu等[81]
PDET Feature+BiLSTM-CRF	深度学习;词典	·上海市曙光医院临床数据(收费项目和药物信息列表) ·一些医学文献,如(人体解剖学名词[第二版])	·设计了五种集成字典和上下文信息构建特征向量的方法 ·提出了两种扩展BiLSTM结构,在处理罕见和未见过的临床命名实体上效果明显	91.24	Wang等[49]

则简称为PR,ChiEHRBert指代使用中文电子病历文本数据预训练获得的中文电子病历BERT模型(Chinese EHR BERT),CRF-MT-adapt代表具有自适应损失加权(Adaptive Loss Weighting)的多任务(Multi-Task)序列标记模型,NER-MRC表示基于机器阅读理解的命名实体识别模型(Named Entity Recognition model based on Machine Reading Comprehension),RoBERTa-wwm-ext-large(https://github.com/ymcui/Chinese-BERT-wwm)是哈尔滨工业大学开源的中文BERT预训练模型.

CCKS2021所有测评论文中,BERT变体BERTology+BiLSTM+CRF仍是被采用最多的模型组合;利用非标注文本、半监督学习和迁移学习等方法提升模型性能是研究重点;采用模型混合的融合方法进行识别效果的提升成为必选;研究者开始考虑通过利用医疗命名实体识别任务和其他任务的关联性提升整体模型识别效果. 具体情况见表10. 其中,Medical bert wwm(https://code.ihub.org.cn/projects/1775/files)是鹏城实验室开源的中文医学BERT预训练模型,CPT指代持续预训练(Continue Pre-Training),SLE代表半监督的标签模式增强策略(Semi-supervised Label mode Enhancement strategy).

具体细化到各类医疗命名实体的识别效果上,受实体类别本身特点、数据选择、方法技术选型、训练参数设置等的影响,不同模型在对应实体类别上表现存

表7　CCKS2018模型方法对比

模型	方法	资源	改进	评价(*F*值)	文献
Att-BiLSTM-CRF	深度学习;规则;词典	·采用《中国日报》含有230万个单词的语料库 ·国家市场监督局提供的17 972种国内药品和1 361种进口药品的产品名称 ·大量的医学文献和教科书	·在BiLSTM和CRF层间加入注意力机制,利用文档级信息缓解标签不一致性问题 ·采用药物字典、后处理规则和实体自动修正算法作为辅助措施来缓解实体边界划分误差、实体识别不完全等缺陷	90.15	Li等[82]
ELMo+BiLSTM-CRF-SP	深度学习;词典	·中国知网的医学文献 ·利用汉典网站(https://www.zdic.net/),获取语料库字汇表中每个字的笔画序列信息 ·利用搜狗输入法词典构建了药物词典	·在大规模无标注数据上使用双向语言模型预训练得到上下文相关且包含汉字内部结构信息的笔画ELMo向量作为输入特征 ·基于多任务学习构建神经网络模型,充分利用多个任务的相关性,进一步提升模型性能	90.05	罗凌等[78]
MSD-MT-BERT-BiLSTM-CRF	深度学习;词典	·未公开	·构造多模态树以融合各种粒度的语义信息,减少提取过程中的信息损失,提高对复杂实体的识别能力 ·借助多粒度语义字典有效提取边界信息,减少分词错误 ·通过多粒度特征融合提高模型性能	89.88	Wang等[83]
Lattice-LSTM-Re-entity-CRF	深度学习;词典	·利用网络和书本构造词典 ·利用Gigaword(https://catalog.ldc.upenn.edu/LDC2011T13)构造词典	·提出一种利用预训练CRF预测结果构造词典并进行分词后,二次训练的CRF模型,减少了分词在医学专业文本上的错误,可以对嵌套类实体进行识别 ·在字符级神经网络模型的基础上融入单词序列信息构建Lattice词格结构,在独立于分词的情况下能利用文本中潜在的单词信息,识别准确率高于单纯基于字粒度或基于词粒度的模型	89.75	潘璨然等[84]
FT-BERT+BiL-STM-CRF	深度学习;词典	·网络爬取到中国临床文本数据 ·《药典》数据	·使用未标记的中国临床文本数据对BERT模型进行预训练,可以利用未标记的领域特定知识 ·使用汉字偏旁部首信息作为输入特征提升预测效果	89.56	Li等[79]
CRF	统计机器学习;词典;规则	·Drugbank药物数据库 ·寻医问药网	·建立了具有字、位置、偏旁部首、拼音、字典和规则特征的条件随机场(CRF)模型,提升了识别效果	89.26	Yang等[85]
Lattice LSTM/BiL-STM-CNN-CRF	深度学习;词典	·搜狗词库 ·利用汉典网站	·提出一种神经网络集成方法,结合了5个单独的神经网络模型(即CNN-CRF,BiLSTM-CRF,BiLSTM-CNN-CRF,BiLSTM+CNN-CRF和Lattite LSTM) ·采用笔顺、分词和字典作为附加特征	88.63	Luo等[86]
BiLSTM-CRF	深度学习;规则;词典	·未公开	·采用药物词典解决药品实体同义词和缩写词的识别不准确问题 ·采用后处理规则处理边界划分误差问题	87.68	Ji等[62]
CRF	规则	·MEDIC数据库	·构建句法和语义的特征工程,有效捕获疾病名称的句法结构和语义信息,从而显著提高疾病名称识别的性能	85.33	何云琪等[87]
RNN+CNN-BiL-STM-CRF	深度学习	·网络爬取和腾讯方面给出的文本数据 ·中文维基百科	·采用分级识别思路,使用RNN完成一级粗粒度识别,使用CNN-BiLSTM-CRF完成二级细粒度识别,提高了识别模型对不同数据的适应性	73.52	向政鹏等[88]

表 8 CCKS2019 模型方法对比

模型	方法	资源	改进	评价(*F*值)	文献
BT-BBT-BBC	深度学习;规则;词典	·百度文库提供常见化疗药物及其英文缩写	·将基于BERT的多个模型融合为一个模型,从而获取更高的精度 ·采用频繁模式挖掘等方法构建规则约束,应对识别实体边界模糊、合并或分裂错误等问题	85.62	乔锐等[45]
BERT-BiLSTM-CRF	深度学习;词典	·网络爬取的医疗领域内各类别实体的词典 ·基于自举策略,使用标注语料挖掘部分高频的实体上下文	·构造细粒度BiLSTM-CRF分层标签模型,引入额外的标签信息 ·借助Lattice LSTM和优化CharCNN -BiLSTM-CRF模型,引入额外的分词信息 ·通过结合拼音、字形信息的BERT模型,引入额外的语义信息 ·多个互补模型融合,进一步提升模型性能	85.59	Liu等[89]
ELMo+BiLSTM-CRF	深度学习;词典;规则	·从中国知网下载医学摘要作为未标注数据,共1 568 458条 ·利用汉典网站 ·CCKS2018语料库	·提出一种新的笔画ELMo来从语言模型中预训练上下文的字嵌入 ·通过三种不同迁移学习模型,进一步提升模型性能	85.16	Li等[90]
CNN+UCNN+WaveNet+Self-Att+BERT-www+BiLSTM-CRF	深度学习;规则	—	·在医疗命名实体识别上首次尝试多种序列标注模型(包括CNN,UCNN,self_attention, WaveNet等),都可以达到一定效果 ·通过多模型增加系统多样性,融合提升系统的整体性能	80.00	赵刚等[91]
CNN-BiLSTM-CRF	深度学习;规则	—	·提出一种基于多神经网络的方法,包括两个BiLSTM-CRF模型和一个用于句子分类的CNN模型	76.35	Ji等[92]

表 9 CCKS2020 模型方法对比

模型	方法	资源	改进	评价(*F*值)	文献
ST-AT-PR-BERT-BiLSTM-CRF	深度学习;规则	—	·构造了由基于对抗性训练的半监督噪声标签学习模型和规则后处理模块组成的混合系统,减小训练用有标记数据中实体注释标准不一致导致的识别误差 ·在模型集成中引入五重交叉投票机制来减小识别误差	91.54	Li等[93]
BERT-BiLSTM-CRF	深度学习;规则	—	·构建了一个基于BERT与字形字音特征的深度学习网络 ·模型融合策略采用双阈值的处理方式来对不同模型的结果进行投票	91.54	晏阳天等[71]
ChiENRBert-BiLSTM/IDCNN-CRF	深度学习	·大小为2G+的脱敏电子病历	·构建采用序列标注和半指针-半标注方式 的多种实体识别模型:使用电子病历对BERT模型进行预训练,同时加入了word2vec训练的字向量、词向量、bichar向量等,有效应对嵌套实体识别 ·改进Bert-Span模型的损失函数,解决样本不平衡问题 ·在中文电子病历实体识别场景下改进Simple-Lexicon模型 ·采用包括模型和实体两级融合	91.24	杨文明等[94]
CRF-MT-Adapt+NER-MRC	深度学习;词典	·从 CCKS2020 语料库提取药物词典	·构建具有自适应损失加权的多任务序列标记模型,较传统的基于CRF的模型表现出更好的性能和可解释性 ·提出基于机器阅读理解的命名实体识别模型,对长跨度实体(跳跃类实体)具有更好的提取能力	90.51	Zheng等[95]
RoBERTa-wwm-ext-large+BiLSTM-CRF	深度学习;词典	·对公开的医疗网站爬取相关的医疗实体,构造包含5万多个医疗实体的领域词典	·提出一种基于医疗领域词典与预训练模型融合的医疗命名实体识别方法,有效解决传统医疗命名实体识别方法对无标签医疗文本数据的利用不充分的问题 ·将非标注数据通过使用伪标签方法,转变成带伪标签的数据文本,扩充了训练集 ·使用无标注医疗文本数据对模型进行同领域预训练,提高模型泛化能力	88.70	温超杰等[96]

表 10　CCKS2021 模型方法对比

模型	方法	资源	改进	评价(F值)	文献
BERT-CRF	深度学习	—	·通过对抗训练优化识别效果 ·采用频繁模式挖掘方法,处理出现次数较少实体的边界识别模糊问题	76.84	Ma等[97]
Medical bert wwm/Ro-BERTa-wwm-ext-large-BiLSTM-CRF+CPT+SLE	深度学习	·从百度百科获取医疗文本数据 ·从快速问医生网站获取医疗文本数据,共计20万份	·提出一种基于预训练模型的无监督文本模式增强和标签模式增强的组合策略,充分利用小规模有标注数据集,获得更好的模型识别效果和泛化能力	—	Gan等[98]
BERT-BiLSTM-CRF	深度学习;规则	·抓取相关网站获取笔画信息	·在BERT模型基础上,通过加入字音字形特征增强了模型对中文医学词汇的表征能力 ·模型融合策略采用双阈值的处理方式,来对不同模型的结果进行投票 ·采用后处理规则校准结果	67.54	晏阳天等[99]

在较大差异. 少有研究者在具体场景下对差异原因进行分析,多从现象出发作出宏观解释,为体现逐年度在各实体类别上最优模型的识别效果,对5年全部模型评价值进行汇总,详细情况见表11.

表 11　2017—2021 年 CCKS 各实体类型最佳评价(F_1)值

实体类型	年份				
	2017	2018	2019	2020	2021
疾病和诊断	81.26	00.00	84.29	91.10	87.00
影像检查	94.94	00.00	88.01	89.96	89.30
实验室检验	94.94	00.00	76.94	85.94	88.40
手术	00.00	87.59	86.79	96.21	87.60
药物	00.00	94.58	96.02	93.75	94.20
症状体征	96.57	00.00	00.00	00.00	00.00
解剖部位	88.74	89.69	86.18	92.00	86.80
症状描述	00.00	92.57	00.00	00.00	00.00
独立症状	00.00	92.94	00.00	00.00	00.00
治疗	83.32	00.00	00.00	00.00	00.00
总体	92.68	90.82	85.62	91.56	76.84

注:本表数据仅统计历年测评论文中提及评价值,不包括未提及或其他在相应数据集上研究的论文中数据.

从时间维度看,除去症状体征、症状描述、独立症状、治疗仅在其中一年被定义为识别类型外,疾病和诊断、影像检查、实验室检验、手术、药物和解剖部位共计六类实体得到较长时间的关注. 虽然从表4可以了解到供训练用数据集质量和大小逐年提升,但表11的数据表明,各实体类型最佳评价值并没有呈现和数据集大小、质量一致而明显的正相关提升趋势,图6中更直观显示了这一问题.

从各类实体类型看,如图7所示,除症状体征、症状描述、独立症状和治疗四类实体外,虽然疾病和诊断、实验室检验、解剖部位类实体识别效果大体呈现提升趋势,影像检查、手术、药物类实体识别效果基本保持稳定. 但总体而言,除药物类实体历年评价值均超过90以外,其他实体都在85~90区间变化,且各实体间评价值相差较大,如何针对具体实体类型进一步提高其评价值仍是未来可供研究的问题.

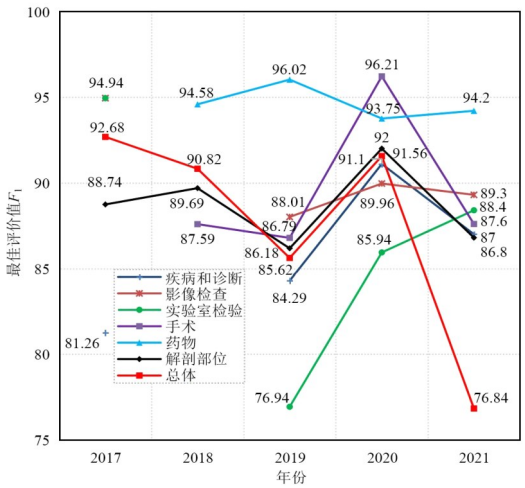

图6　2017—2021 年 CCKS 逐年度各类型命名实体识别准确度变化

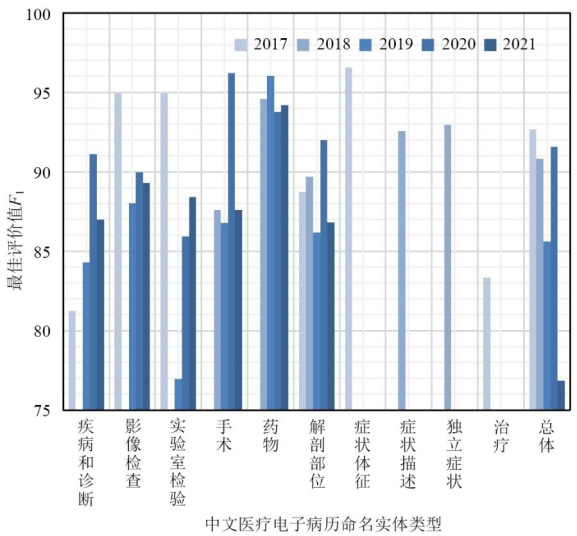

图7　CCKS 各类型命名实体识别准确度 2017—2021 年逐年度变化

通过对部分模型具体实验，可以发现细化到具体的实体类别上，影响模型识别效果的主要是嵌套类实体和非连续类实体的识别，这也是近年来研究者关注的热点.

针对嵌套类实体识别，目前主要通过换用不同的标注框架，包括序列标注、指针标注、多头注意力标注和分片段标注四种. 序列标注在上文已经介绍，由于嵌套类实体的每一个序列位置可能隶属多个类型标签，所以考虑将原来的多分类问题转化为多标签分类问题[100]，更进一步也有研究者将多个标签采用多CRF层进行合并[84]；指针标注是对每个区间的起始和结束进行标记，将嵌套实体识别转换为层叠式指针标注[94]；多头选择标注[101]对每个区间对进行标记，通过构建分类矩阵进行嵌套实体的识别；分片段标注[102]枚举所有可能的区间后进行细粒度实体分类[103].

而针对非连续实体，除采用序列标注的思路进行处理外，主要考虑将该类型实体识别转换为关系抽取问题，并结合相应的规则加以处理；也有研究者考虑结合语法树结构[83]，构造句法解析器对非连续实体进行识别.

4.2 主流模型效果验证

从表6~10的分析可知，BERT-BiLSTM-CRF(BBC)模型为近年来最受欢迎且效果相对较好的主流模型组合，本节通过实验对该模型做进一步研究说明.

4.2.1 模型构成

BBC模型具体由三部分组成：BERT预训练模型用于解决从中文电子病历中提取特征时存在的特征稀疏问题，充分获取病历文本的字粒度信息；BiLSTM在BERT基础上，双向利用上下文信息对句子进行建模；CRF考虑标签间的依赖关系，为BiLSTM预测的标签添加约束，以保证最终预测标签的合法性，进一步提升模型准确性. BBC基本结构如图8所示.

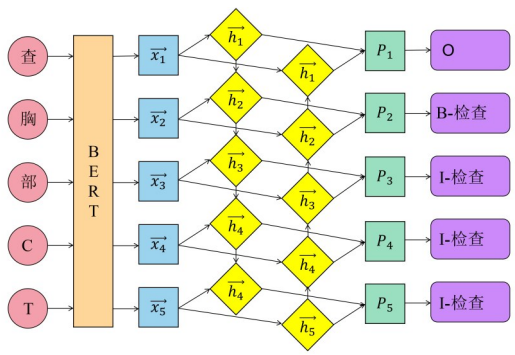

图8 基于BERT-BILSTM-CRF的电子病历命名实体识别模型

4.2.2 实验数据

本文通过将六个数据集进行集成后训练和评估该模型，包括CCKS2017(https://www.biendata.xyz/competition/CCKS2017_2/data/)，CCKS2018(https://www.biendata.xyz/competitionKS2018_1/data/)，CCKS2019(https://www.biendata.xyz/competition/ccks_2019_1/data/)，CCKS2020(https://www.biendata.xyz/competition/ccks_2020_2_1/data/)，CCKS2021(https://www.biendata.xyz/competition/ccks_2021_clinic/data/)和CBLUE(https://tianchi.aliyun.com/dataset/dataDetail?dataId=95414)数据集. 按照16:3:1划分训练集、验证集和测试集. 各原始数据集文本/医疗记录条数见表12(CCKS在表12中简称C).

表12 实验用数据集文本数量 单位：条

数据类型	数据集名称					
	C2017	C2018	C2019	C2020	C2021	CBLUE
有标注	1 200	1 200	1 379	1 500	1 500	20 000
无标注	10 420	—	—	1 300	1 000	3 000

4.2.3 参数选择

为实现深度神经网络模型，本文利用具有Pytorch后端的Keras库，每个模型都在一个NVIDIA GeForce RTX 3090 GPU上运行.

实验设定参数如表13所示. 其中max_length表示BERT模型输入的最长文本序列长度；hidden_dim表示LSTM隐层单元/节点个数(这里指单方向的隐层节点数)；dropout_rate表示在训练过程中的dropout概率，本文在BiLSTM输出部分使用Dropout技术减少过拟合；optimizer表示模型所采用的优化算法；epoch表示迭代次数；batch_size表示数据分批后逐批次大小；learning_rate表示模型的学习率.

表13 模型参数取值方案及结果

参数	方案序号			
	1	2	3	4
max_length	200	128	200	128
hidden_dim	128	256	128	64
dropout_rate	0.5	0.5	0.5	0.5
optimizer	Adam[104]	Adam	AdamW[105]	AdamW
epoch	5	10	10	5
batch_size	16	8	8	16
learning_rate	5E−5	5E−5	5E−5	5E−5
评价	90.02	92.08	91.74	93.56

表13给出调试后性能较好的4种参数取值方案，并在表尾给出了各方案的实际效果. 除dropout和learning_trate选择了定值，其他参数对比来看.

(1)max_length：虽然max_length越长越能提供给模型丰富的上下文语义环境信息，但模型可能无法较好地学习到长文本的语义信息，反而导致训练效果变差.

(2)hidden_dim：该值越大，模型结构越复杂，需要

学习的参数越多,虽然在一定范围内可以提高模型整体效果,但该参数过大后模型可能存在收敛难的问题.

(3)optimizer:机器学习模型大都需要利用损失函数检验模型效果,同时利用针对损失函数结果的学习提升模型性能. 提升的方法有很多,这里对比了四种:SGD是随机梯度下降算法;AdaGrad是自适应梯度下降算法,在SGD基础上做了改进,可以让目标函数更快收敛;RMSProp也是一种自适应梯度下降算法,缓解了AdaGrad学习率下降较快的问题;Adam是对RMSProp的升级,加入了Momentum动量机制;AdamW修正了Adam权重衰减的问题. 具体到中文电子病历实体识别场景下,这些算法对模型的训练效果有不同影响,图9明显地反映了这一问题. 其中,Adam和AdamW表现较为优异,故表13中选用了这两种优化算法.

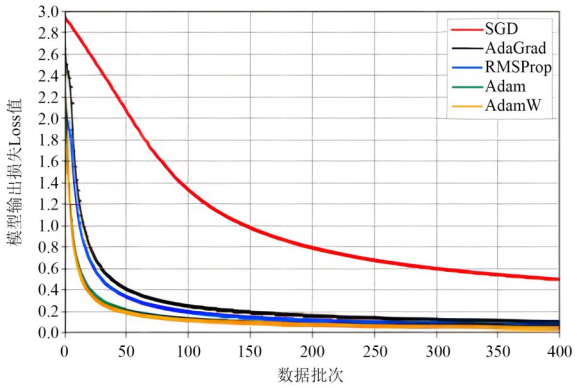

图9　不同优化算法下模型损失随训练数据集大小变化

(4)epoch:训练代数的选择需要根据实际实验结果动态确定. 随着epoch的增大,神经网络的权值也在不断更新,模型逐渐从欠拟合训练到过拟合. 从图10可以看出,在本次实验情景下,epoch在0到6区间内模型训练效果逐步提升,并在epoch=6处获得最佳,之后随着epoch的增大,模型效果在固定区间波动,不再有明显的性能提升.

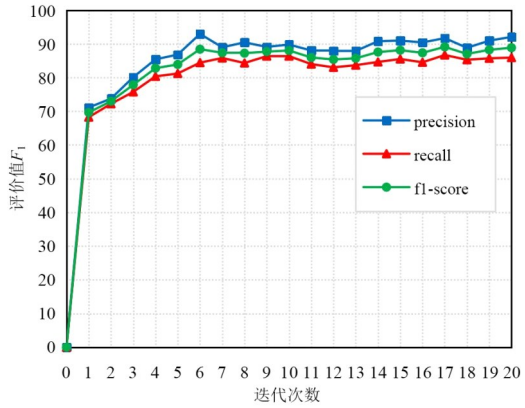

图10　不同迭代次数对应命名实体识别准确度

(5)batch_size:指征数据批次的大小. 如果该值选择过小,则每个批次内的数据重叠概率低、差异性大,需要花费更多的时间进行训练且梯度震荡、模型不易收敛;如果该值选择过大,则每个批次内的数据重叠概率高、差异性小,虽然训练需要的时间会减少但梯度容易消失,使模型陷入局部极小.

图11形象反映了中文电子病历命名实体识别场景下BBC模型在不同数据批次下模型效果和训练所耗费的时间. 结果表明:评价值随批次大小的增大先提高后维持稳定,符合上述分析;训练时间随批次大小的增大逐步减少,符合实际情况.

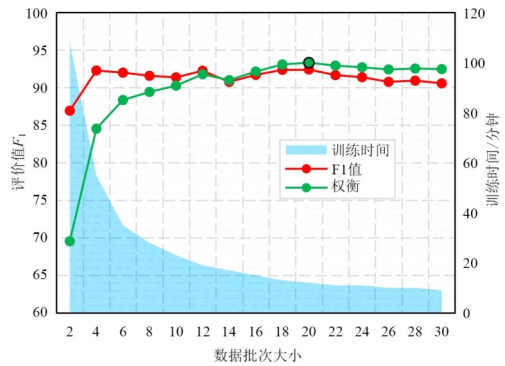

图11　不同数据批次大小下模型的评价值F_1和训练时间

为权衡模型效果和训练时间,综合评价模型,本文提出一种的权衡指标算法,其公式为

$$\mathrm{Weigh}_i=\alpha\cdot F_{1i}+100\cdot(1-\alpha)\cdot\left(1-\frac{\mathrm{time}_i-\min(\mathrm{Time})}{\max(\mathrm{Time})-\min(\mathrm{Time})}\right) \tag{6}$$

其中,Weigh表示权衡指标;α是权重,表示对评价值和训练时间两者的重视程度;Time是所有批次所用时间构成的时间列表;F_1是评价值. 图10中取α为0.9. 结果表明,数据批次选择20时,模型的识别效果和训练耗费时间最可接受.

4.2.4　实验结果

使用目前公认的Benchmark进行基准测试:数据表明,BBC分别在通用领域命名实体任务排行榜CLUE上获得80.509的得分(最高83.351),在中文医疗信息处理挑战榜CBLUE上获得69.964的得分(最高70.617),属于目前表现较好的模型之一,但仍有可提升空间.

使用4.2.2节集成并预处理好的大规模中文医疗电子病历文本数据,通过随机打乱句序进一步扩充数据集规模;结合实验室前期整理的共计325 800条词库数据;同时利用网络可爬取到的医疗无标注文本资源,对BBC模型做进一步测试. 实验结果如表14和表15所示.

表14对BBC模型各构成部分对模型整体效果的贡

表 14 不同模型命名实体识别结果 单位:%

序号	模型	Prec	Rec	F_1
1	word2vec-BiLSTM-CRF	86.74	85.12	85.92
2	cw2vec-BiLSTM-CRF	87.16	86.84	87.00
3	BERT	93.16	91.27	92.21
4	BERT-CRF	93.81	91.73	92.76
5	BERT-BiLSTM-CRF	94.49	92.64	93.56

献程度进行了实验. 结果表明,BERT 模型较传统 BiLSTM-CRF 模型获得更好的识别效果;通过将 BERT 引入 BiLSTM-CRF 模型中构成 BBC 模型,取得了最好的中文电子病历命名实体识别评价值.

表 15 具体展示了各模型结构具体在不同类别命名实体上的识别效果. 数据显示,以 BBC 为代表的模型普遍在实验室检验类实体上识别效果不佳,在其他类实体上识别效果虽然较好但程度不一. 由数据集分析可知,实验室检验类实体多中英混合,给模型识别带来较大难度;其他类实体识别效果存在差异的原因和进一步按类型提高识别精度将是下一步工作的重点.

表 15 不同类型模型命名实体识别结果 单位:%

模型	实体类型					
	疾病和诊断	影像检查	实验室检验	手术	药物	解剖部位
word2vec-BiLSTM-CRF	85.24	87.49	82.67	88.74	89.60	86.67
cw2vec-BiLSTM-CRF	85.52	88.50	83.23	89.01	89.92	86.79
BERT	92.41	93.79	90.58	94.46	94.63	93.10
BERT-CRF	93.36	94.23	91.76	94.56	95.64	93.28
BERT-BiLSTM-CRF	94.41	95.14	92.60	94.81	96.55	93.44

4.2.5 模型对比

BBC 模型结构中,BiLSTM-CRF 通常不会有较大改动,多利用参数调试进行模型优化;而 BERT 部分则可以通过替换不同的 BERT 模型,以达到进一步提高模型效果的目的.

在 2018 年 BERT 模型提出后,国内研究者首先针对中文语境提出 BERT 的汉化版 BERT-wwm,后续又有研究者针对中文医疗领域发布中文医学预训练模型 Med-BERT;2019 年 BERT 模型改良版 RoBERTa 模型提出后,其汉化版 RoBERTa_zh 和 RoBERTa-wwm 接续出现.

本文使用 BERT 变体在上述数据集上分别进行了训练,以对比显示采用不同 BERT 模型对 BBC 模型整体性能的提升,结果如表 16 所示.

4.2.6 模型优化

为进一步提升 BBC 模型的识别效果,不同研究者提出大量切实可行的方法,表 6~10 充分展示了这些研究者的思路. 结合 4.2 节实际实验效果,对改良 BBC 模型的方法做汇总梳理,供后续研究借鉴,具体情况详见表 17.

表 16 不同类型模型命名实体识别结果 单位:%

模型	F_1
BERT-base-BiLSTM-CRF	93.56
BERT-wwm-ext-BiLSTM-CRF	94.15
RoBERTa-wwm-ext-BiLSTM-CRF	94.12
RoBERTa-wwm-ext-large-BiLSTM-CRF	94.63

表 17 改良 BBC 模型的若干方法

类型	具体方法	参考
调参	对模型内部超参数做调优实验	表 13
结构调整	换用不同的 BERT 模型	文献[94,96]
	采用分级识别思路,多次利用模型	文献[88]
特征工程	先进行分词,后将词信息做嵌入	文献[76]
	将字音特征利用 CNN 等做嵌入	文献[99]
	将字形特征利用 CNN 等做嵌入	文献[80,99]
	将笔画特征利用 ELMo 等做嵌入	文献[78]
	将位置特征利用 Jieba 等做标记嵌入	文献[86]
	将词典信息做嵌入	文献[86]
多模型	采用投票思路进行多模型集成	文献[76,86]
词典	采用词典匹配对错误实体进行修正	文献[82]
后处理规则	采用人工构造的后处理规则对不完全识别实体进行修正	文献[62,82]

5 结论

海量电子病历数据是支撑医疗智能化研究的重要原料,然而电子病历文本数据的半结构化甚至无结构化特点,造成后续对其分析利用的极大困难. 虽然近年来基于深度学习的命名实体识别技术已经发展到可以有效完成电子病历的命名实体识别任务,但由于中文电子病历所具有的包括病历文本的非规范性和专业性、医疗实体的独特性和标注语料的稀缺性在内的独特文本数据特征,该研究目前仍存在诸多挑战.

本文对中文电子病历命名实体识别的研究与进展进行了综述,系统梳理了中文电子病历命名实体识别的相关理论;从技术发展角度详细叙述了中文电子病历命名实体识别方法的变革历程;并对中文电子病历命名实体识别效果做了实验验证与深入分析,指出了现有模型的不足与改进方向;鉴于国内近年来与中文信息学处理相关的测评会议 CCKS 持续关注中文电子病历命名实体识别,本文特别对 CCKS 在该领域五年来的全部代表性测评论文做了对比分析,并通过在主流模型 BBC 上的深入实验与研究,为后续该领域的继续推进寻求了思路.

虽然围绕电子病历文本数据处理的医疗命名实体

识别并非新兴研究方向,与其他通用领域文本数据上的命名实体识别技术差别不大,但中文电子病历自身所具备的专业性和隐私性等特点,让该领域到目前为止仍存在极大的研究空间,主要体现在训练语料获取难度大、现有识别方法仍存在可改进之处等. 基于本文调研,我们认为以下几个方面是未来中文电子病历命名实体识别研究中值得重点关注的方向:

(1)针对特殊实体类型研究识别率的提升方法. 上文实验结果表明,"实验室检验"类实体的F_1明显较低. 潜在原因有二:一是该类实体多有中英文混杂的情况,从而导致模型不能很好地判断实体边界;二是难以识别出长度为1的短实体以及不能完整识别出较长实体,该类实体还明显存在实体嵌套的现象,导致严格匹配指标F_1值较低. 针对不同类型实体,特别是针对中文电子病历中特殊类型的实体,包括嵌套类实体和非连续类实体,鉴于其自身结构和语义的复杂性至今仍是制约中文电子病历实体识别效果的要因,有必要对以往模型的实验结果做进一步分析,统计特殊类实体的识别情况,并对特定实体类型所存在的问题进行具体优化.

(2)寻求性能表现更佳的模型结构. 综合调研结果,我们发现基于词典和规则的实体识别方法均因自身缺陷而不再被独立研究,多结合到基于机器学习的实体识别方法中,作为提升模型性能的两种手段;而基于机器学习的实体识别方法目前仅BBC模型被广为采纳. 可以预见,在更优的模型架构提出以前,BBC模型不会被淘汰. 因此,下一步,一方面可以考虑采用4.2.6节提出的12种方法改良BBC模型,另一方面也可以考虑借鉴图像识别等其他领域思路,在中文电子病历命名实体识别情景下寻找性能更佳的模型结构.

(3)采用多元的模型学习方式. 深度学习模型大多为数据驱动,足够且高质量的数据才能让模型学到一定的知识,从而达到相较理想的模型效果. 而短时间内中文电子病历的命名实体识别仍无法获得足量的数据,这也是大部分研究者在模型识别效果提升上受到制约的潜在原因. 未来可以在模型上尝试采用不同的学习方式解决这一问题,如主动学习[106]、自学习[107]、迁移学习[108]、多任务学习[78]、元学习[109]和小样本学习等.

(4)进一步提升模型训练和测试效率. 经实际测验,在一定参数设置下一个主流的中文电子病历命名实体识别模型BBC在CPU上训练时长超过24小时,在GPU上训练时间也长达3小时. 此外,并非可并行计算模型结构中的各个部分都能采用GPU加速计算,如BERT-LSTM-CRF模型中,由于单个LSTM模型自身结构无法并行,这一部分就无法使用GPU进行加速. 因此,在算力资源不紧张的情况下采用分布式学习如联邦学习[110]等思路,在算力资源有限的情况下寻求合适的模型训练方案以提升效率,在实际应用场景下都十分必要.

(5)完善中文医疗领域语料库资源,构建开放高质量数据集. 虽然目前部分研究者如本节第(3)点所述,从小样本学习、领域迁移学习或者对医疗数据进行无监督学习等方向进行了初步尝试并取得一定进展,如高冰涛等人[41]构建的基于迁移学习的隐马尔可夫模型BioTrHMM仅需要少量的目标领域标注数据即可在医学命名实体上获得较好性能,但在大数据浪潮下,建立统一的标注标准和公共数据集,降低数据集标注的人工成本和时间成本,以及利用自动化方式获得较为完善而高质量的中文医疗领域语料库,仍然是较为紧迫的研究问题,需要政府、医院和相关研究者共同出力.

(6)与其他研究方向做联合研究. 自CCKS2019以来,中文电子病历命名实体识别任务就开始和其他任务做联合测评. 鉴于中文电子病历命名实体识别最终为电子病历文本数据结构化和标准化、医疗知识图谱的构建等服务,联合研究既降低了研究成本、减少了分开研究潜在的信息丢失和误差传递现象,同时还能通过研究方向之间的关联性,为彼此提供更丰富的扩展信息,进一步提升方法的整体性能,目前也吸引了较多研究者关注.

参考文献

[1] 国家卫健委. 关于印发电子病历应用管理规范(试行)的通知[EB/OL]. (2017-02-22)[2022-01-02]. http://www.nhc.gov.cn/yzygj/s3593/201702/22bb2525318f496f846e8566754876a1.shtml.

[2] 马欢欢, 孔繁之, 高建强. 中文电子病历命名实体识别方法研究[J]. 医学信息学杂志, 2020, 41(4): 24-29.
MA H H, KONG F Z, GAO J Q. Study on named entity recognition method of Chinese electronic medical records [J]. Journal of Medical Informatics, 2020, 41(4): 24-29. (in Chinese)

[3] 辛海燕, 李鹏, 张国庆. 医院医疗科研大数据平台的建设与应用[J]. 中国卫生信息管理杂志, 2019, 16(2): 206-209.
XIN H Y, LI P, ZHANG G Q. Construction and application of medical research big data platform in hospital[J]. Chinese Journal of Health Informatics and Management, 2019, 16(2): 206-209. (in Chinese)

[4] 崔博文, 金涛, 王建民. 自由文本电子病历信息抽取综述[J]. 计算机应用, 2021, 41(4): 1055-1063.
CUI B W, JIN T, WANG J M. Overview of information extraction of free-text electronic medical records[J]. Journal of Computer Applications, 2021, 41(4): 1055-1063. (in

Chinese)

[5] 付秀, 陈麒麟, 李杰, 等. 基于智能预问诊的全景多学科会诊平台的设计与应用[J]. 中国数字医学, 2021, 16(10): 79-82.
FU X, CHEN Q L, LI J, et al. Design and application of the panoramic multi-disciplinary treatment platform based on intelligent pre-consultation[J]. China Digital Medicine, 2021, 16(10): 79-82. (in Chinese)

[6] 吴宗友, 白昆龙, 杨林蕊, 等. 电子病历文本挖掘研究综述[J]. 计算机研究与发展, 2021, 58(3): 513-527.
WU Z Y, BAI K L, YANG L R, et al. Review on text mining of electronic medical record[J]. Journal of Computer Research and Development, 2021, 58(3): 513-527. (in Chinese)

[7] 杨锦锋, 于秋滨, 关毅, 等. 电子病历命名实体识别和实体关系抽取研究综述[J]. 自动化学报, 2014, 40(8): 1537-1562.
YANG J F, YU Q B, GUAN Y, et al. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction[J]. Acta Automatica Sinica, 2014, 40(8): 1537-1562. (in Chinese)

[8] 全国知识图谱与语义计算大会. CCKS 2021 评测二: 电子病历命名实体识别[EB/OL]. (2021-5-31)[2022-01-02]. https://www.biendata.xyz/competition/ccks_2021_clinic/.

[9] 程楠, 侯豪, 牛亚军, 等. 基于NLP技术后结构化处理的电子病历应用[J]. 河南医学研究, 2021, 30(24): 4510-4513.
CHENG N, HOU H, NIU Y J, et al. Application of post-structured electronic medical record based on NLP technology[J]. Henan Medical Research, 2021, 30(24): 4510-4513. (in Chinese)

[10] NADEAU D, SEKINE S. A survey of named entity recognition and classification[J]. Lingvisticæ Investigationes, 2007, 30: 3-26.

[11] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[12] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 2001: 282-289.

[13] KE X, LI S Z. Chinese organization name recognition based on co-training algorithm[C]//2008 3rd International Conference on Intelligent System and Knowledge Engineering. Xiamen: IEEE, 2008: 771-777.

[14] ANDO R, ZHANG T. A framework for learning predictive structures from multiple tasks and unlabeled data[J]. Journal of Machine Learning Research, 2005, 6: 1817-1853.

[15] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[16] WANG Z H, YANG B. Attention-based bidirectional long short-term memory networks for relation classification using knowledge distillation from BERT[C]//2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress. Calgary: IEEE, 2020: 562-568.

[17] 曹春萍, 关鹏举. 基于E-CNN和BLSTM-CRF的临床文本命名实体识别[J]. 计算机应用研究, 2019, 36(12): 3748-3751.
CAO C P, GUAN P J. Clinical text named entity recognition based on E-CNN and BLSTM-CRF[J]. Application Research of Computers, 2019, 36(12): 3748-3751. (in Chinese)

[18] STRUBELL E, VERGA P, BELANGER D, et al. Fast and accurate entity recognition with iterated dilated convolutions[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 2670-2680.

[19] 许力, 李建华. 基于BERT和BiLSTM-CRF的生物医学命名实体识别[J]. 计算机工程与科学, 2021, 43(10): 1873-1879.
XU L, LI J H. Biomedical named entity recognition based on BERT and BiLSTM-CRF[J]. Computer Engineering & Science, 2021, 43(10): 1873-1879. (in Chinese)

[20] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]//International Conference on Learning Representations. Scottsdale: ICLR, 2013: 1-12.

[21] PENNINGTON J, SOCHER R, MANNING C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014: 1532-1543.

[22] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11)[2022-01-05]. https://arxiv.org/abs/1810.04805.

[23] WU Y, HUANG J, XU C E, et al. Research on named en-

tity recognition of electronic medical records based on RoBERTa and radical-level feature[J]. Wireless Communications and Mobile Computing, 2021, 2021: 2489754.

[24] 杨锦锋, 关毅, 何彬, 等. 中文电子病历命名实体和实体关系语料库构建[J]. 软件学报, 2016, 27(11): 2725-2746.
YANG J F, GUAN Y, HE B, et al. Corpus construction for named entities and entity relations on Chinese electronic medical records[J]. Journal of Software, 2016, 27 (11): 2725-2746. (in Chinese)

[25] 全国知识图谱与语义计算大会. 任务二: 电子病历命名实体识别[EB/OL]. (2017)[2022-01-05]. http://www.sigkg.cn/ccks2017/?page_id=51.
National Knowledge Graph and Semantic Computing Conference. Task 2: Electronic medical record named entity recognition[EB/OL]. (2017)[2022-01-05]. http://www.sigkg.cn/ccks2017/?page_id=51. (in Chinese)

[26] 全国知识图谱与语义计算大会. 任务一: 面向中文电子病历的命名实体识别[EB/OL]. (2018)[2022-01-05]. http://www.sigkg.cn/ccks2018/?page_id=16.
National Knowledge Graph and Semantic Computing Conference. Task 1: Named entity recognition for Chinese electronic medical Record[EB/OL]. (2018) [2022-01-05]. http://www.sigkg.cn/ccks2018/?page_id=16. (in Chinese)

[27] 全国知识图谱与语义计算大会. 任务一: 面向中文电子病历的命名实体识别[EB/OL]. (2019)[2022-01-05]. http://www.sigkg.cn/ccks2019/?page_id=62.
National Knowledge Graph and Semantic Computing Conference. Task 1: Named entity recognition for Chinese electronic medical record[EB/OL]. (2019) [2022-01-05]. http://www.sigkg.cn/ccks2019/?page_id=62. (in Chinese)

[28] 全国知识图谱与语义计算大会. 任务三: 面向中文电子病历的医疗实体及事件抽取[EB/OL]. (2020)[2022-01-05]. http://sigkg.cn/ccks2020/?page_id=69.
National Knowledge Graph and Semantic Computing Conference. Task three: For electronic medical records in Chinese medical entities and event extraction[EB/OL]. (2020) [2022-01-05]. http://sigkg. cn/ccks2020/? page_id=69. (in Chinese)

[29] 全国知识图谱与语义计算大会. 任务四: 面向中文电子病历的医疗实体及事件抽取[EB/OL]. (2021)[2022-01-05]. http://sigkg.cn/ccks2021/?page_id=27.
National Knowledge Graph and Semantic Computing Conference. Task 4: For electronic medical records in Chinese medical entity and event extraction[EB/OL]. (2021) [2022-01-05]. http://sigkg. cn/ccks2021/? page_id=27. (in Chinese)

[30] 王正宏. 区域健康医疗数据集成模式研究与实现[D]. 合肥: 合肥工业大学, 2020.
WANG Z H. Research and Implementation of Regional Health Medical Data Integration Model[D]. Hefei: Hefei University of Technology, 2020. (in Chinese)

[31] 韩丽珍. PDCA循环法应用前后肿瘤科病案缺陷状况对比分析[J]. 中国卫生统计, 2019, 36(5): 745-747.
HAN L Z. Comparative analysis of medical record defects in oncology department before and after application of PDCA cycle[J]. Chinese Journal of Health Statistics, 2019, 36(5): 745-747. (in Chinese)

[32] 邱炎龙. 基于电子病历的心血管疾病预测技术研究[D]. 兰州: 西北师范大学, 2021.
QIU Y L. Research on Cardiovascular Disease Prediction Technology Based on Electronic Medical Records[D]. Lanzhou: Northwest Normal University, 2021. (in Chinese)

[33] 余健, 胡孔法, 丁有伟. 一种面向中医药数据的高效脱敏算法[J]. 世界科学技术-中医药现代化, 2020, 22(12): 4169-4174.
YU J, HU K F, DING Y W. An efficient desensitization algorithm for Chinese medicine data[J]. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2020, 22(12): 4169-4174. (in Chinese)

[34] 唐观根. 中文电子病历命名实体识别研究[D]. 杭州: 杭州电子科技大学, 2020.
TANG G G. Research on Named Entity Recognition of Chinese Electronic Medical Records[D]. Hangzhou: Hangzhou Dianzi University, 2020. (in Chinese)

[35] GRISHMAN R, SUNDHEIM B. Message Understanding Conference-6: A brief history[C]//Proceedings of the 16th Conference on Computational linguistics-Volume 1. Copenhagen: Association for Computational Linguistics, 1996: 466-471.

[36] DODDINGTON G, MITCHELL A, PRZYBOCKI M A, et al. The automatic content extraction(ACE) program - tasks, data, and evaluation[C]//Language Resources and Evaluation Conference. Lisbon: LREC, 2004: 1-4.

[37] XU L, TONG Y, DONG Q Q, et al. CLUENER2020: Fine-grained named entity recognition dataset and benchmark for Chinese[EB/OL]. (2020-01-13) [2022-01-05]. https://arxiv.org/abs/2001.04351.

[38] ZHANG N Y, CHEN M S, BI Z, et al. CBLUE: A Chinese biomedical language understanding evaluation benchmark[EB/OL]. (2021-01-15) [2022-01-05]. https://

arxiv.org/abs/2106.08087.

[39] 龚乐君, 张知菲. 基于领域词典与CRF双层标注的中文电子病历实体识别[J]. 工程科学学报, 2020, 42(4): 469-475.
GONG L J, ZHANG Z F. Clinical named entity recognition from Chinese electronic medical records using a double-layer annotation model combining a domain dictionary with CRF[J]. Chinese Journal of Engineering, 2020, 42(4): 469-475. (in Chinese)

[40] GORINSKI P J, WU H H, GROVER C, et al. Named entity recognition for electronic health records: A comparison of rule-based and machine learning approaches [EB/OL]. (2019-03-10)[2022-01-05]. https://arxiv.org/abs/1903.03985.

[41] 高冰涛, 张阳, 刘斌. BioTrHMM: 基于迁移学习的生物医学命名实体识别算法[J]. 计算机应用研究, 2019, 36(1): 45-48.
GAO B T, ZHANG Y, LIU B. BioTrHMM: Named entity recognition algorithm based on transfer learning in biomedical texts[J]. Application Research of Computers, 2019, 36(1): 45-48. (in Chinese)

[42] 杨丽静, 唐俊, 沈伟富, 等. 基于命名实体识别的恶性肿瘤诊断文本信息提取研究[J]. 医院管理论坛, 2020, 37(8): 74-77.
YANG L J, TANG J, SHEN W F, et al. Research on text information extraction of malignant tumor diagnosis based on named entity recognition[J]. Hospital Management Forum, 2020, 37(8): 74-77. (in Chinese)

[43] 张华丽, 康晓东, 李博, 等. 结合注意力机制的Bi-LSTM-CRF中文电子病历命名实体识别[J]. 计算机应用, 2020, 40(S1): 98-102.
ZHANG H L, KANG X D, LI B, et al. Medical name entity recognition based on Bi-LSTM-CRF and attention mechanism[J]. Journal of Computer Applications, 2020, 40(S1): 98-102. (in Chinese)

[44] DOS SANTOS C N, ZADROZNY B. Learning character-level representations for part-of-speech tagging[C]//Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32. Beijing: JMLR.org, 2014: II(1818-1826).

[45] 乔锐, 杨笑然, 黄文亢. 基于BERT与模型融合的医疗命名实体识别[C]//2019年全国知识图谱与语义计算大会. 杭州: 中国中文信息学会, 2019: 1-6.

[46] 曲春燕, 关毅, 杨锦锋, 等. 中文电子病历命名实体标注语料库构建[J]. 高技术通讯, 2015, 25(2): 143-150.
QU C Y, GUAN Y, YANG J F, et al. The construction of annotated corpora of named entities for Chinese electronic medical records[J]. Chinese High Technology Letters, 2015, 25(2): 143-150. (in Chinese)

[47] 刘一斌. 中医中文电子病历命名实体语料库构建及研究[D]. 广州: 广州中医药大学, 2020.
LIU Y B. Construction and Research of Chinese Electronic Medical Record Named Entity Recognition Corpus[D]. Guangzhou: Guangzhou University of Chinese Medicine, 2020. (in Chinese)

[48] 陈曙东, 罗超, 欧阳小叶, 等. 基于动态词典匹配的语义增强中文命名实体识别算法[J]. 无线电工程, 2021, 51(7): 519-525.
CHEN S D, LUO C, OUYANG X Y, et al. A semantic-enhanced Chinese named entity recognition algorithm based on dynamic dictionary matching[J]. Radio Engineering, 2021, 51(7): 519-525. (in Chinese)

[49] WANG Q, ZHOU Y M, RUAN T, et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2019, 92: 103133.

[50] CHEN X L, OUYANG C P, LIU Y B, et al. Improving the named entity recognition of Chinese electronic medical records by combining domain dictionary and rules[J]. International Journal of Environmental Research and Public Health, 2020, 17(8): 2687.

[51] JUSTYNA S W, ALEKSANDER W, ALEKSANDER P, et al. Detecting formal thought disorder by deep contextualized word representations[J]. Psychiatry Research, 2021, 304: 114135.

[52] ZHANG Y, YANG J. Chinese NER using lattice LSTM [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 1554-1564.

[53] ANDREW B F. A Maximum Entropy Approach to Named Entity Recognition[D]. New York: New York University, 1999.

[54] 陈琛. 基于BiGRU_CRF模型的医疗领域命名实体识别[J]. 电子技术与软件工程, 2020(14): 180-182.
CHEN C. Named entity recognition in medical field based on BiGRU_CRF mode[J]. Electronic Technology & Software Engineering, 2020(14): 180-182. (in Chinese)

[55] FINE S, SINGER Y, TISHBY N. The hierarchical hidden Markov model: Analysis and applications[J]. Machine Learning, 1998, 32(1): 41-62.

[56] MCCALLUM A, FREITAG D, PEREIRA F C. Maximum entropy Markov models for information extraction

and segmentation[C]//Proceedings of the Seventeenth International Conference on Machine Learning. Stanford: Morgan Kaufmann Publishers Inc., 2000: 591-598.

[57] 李博, 康晓东, 张华丽, 等. 采用Transformer-CRF的中文电子病历命名实体识别[J]. 计算机工程与应用, 2020, 56(5): 153-159.
LI B, KANG X D, ZHANG H L, et al. Named entity recognition in Chinese electronic medical records using transformer-CRF[J]. Computer Engineering and Applications, 2020, 56(5): 153-159. (in Chinese)

[58] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. (2014-08-25)[2022-01-05]. https://arxiv.org/abs/1408.5882.

[59] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL]. (2014-06-03)[2022-01-05]. https://arxiv.org/abs/1406.1078.

[60] CHIU J P, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[61] WANG Y Q, HUANG M L, ZHU X Y, et al. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016: 606-615.

[62] JI B, LIU R, LI S S, et al. A BiLSTM-CRF method to Chinese electronic medical record named entity recognition[C]//Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence. Sanya: ACM, 2018: 1-6.

[63] LI X N, YAN H, QIU X P, et al. FLAT: Chinese NER using flat-lattice transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual Conference: Association for Computational Linguistics, 2020: 6836-6842.

[64] ADHIKARI A, RAM A, TANG R, et al. DocBERT: BERT for document classification[EB/OL]. (2019-04-17)[2022-01-05]. https://arxiv.org/abs/1904.08398.

[65] ALBERTI C, LEE K, COLLINS M. A BERT baseline for the natural questions[EB/OL]. (2019-01-24)[2022-01-05]. https://arxiv.org/abs/1901.08634.

[66] YANG W, ZHANG H T, LIN J. Simple applications of BERT for ad hoc document retrieval[EB/OL]. (2019-03-26)[2022-01-05]. https://arxiv.org/abs/1903.10972.

[67] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training[J]. Computer Science, 2018: 1-12.

[68] 朱岩, 张利, 王煜. 基于RoBERTa-WWM的中文电子病历命名实体识别[J]. 计算机与现代化, 2021(2): 51-55.
ZHU Y, ZHANG L, WANG Y. Named entity recognition on Chinese electronic medical records based on RoBERTa-WWM[J]. Computer and Modernization, 2021(2): 51-55. (in Chinese)

[69] 刘司宇. 基于深度学习的中文命名实体识别方法改进研究[D]. 成都: 成都理工大学, 2020: 70-71.
LIU S Y. The Research on Improvement of Chinese Named Entity Recognition Method Based on Deep Learning[D]. Chengdu: Chengdu University of Technology, 2020: 70-71. (in Chinese)

[70] MA X Z, HOVY E. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF[EB/OL]. (2016-03-04)[2022-01-05]. https://arxiv.org/abs/1603.01354.

[71] 晏阳天, 赵新宇, 吴贤. 基于BERT与字形字音特征的医疗命名实体识别[C]//2020年全国知识图谱与语义计算大会. 南昌: 中国中文信息学会, 2020: 1-7.

[72] 殷章志, 李欣子, 黄德根, 等. 融合字词模型的中文命名实体识别研究[J]. 中文信息学报, 2019, 33(11): 95-100, 106.
YIN Z Z, LI X Z, HUANG D G, et al. Chinese named entity recognition ensembled with character[J]. Journal of Chinese Information Processing, 2019, 33(11): 95-100, 106. (in Chinese)

[73] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation [C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich: Springer, 2015: 234-241.

[74] WANG S H, JIANG J. Machine comprehension using match-LSTM and answer pointer[EB/OL]. (2016-08-29)[2022-01-05]. https://arxiv.org/abs/1608.07905.

[75] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for Chinese BERT[EB/OL]. (2019-06-19)[2022-01-05]. https://arxiv.org/abs/1906.08101.

[76] HU J L, SHI X, LIU Z J, et al. HITSZ_CNER: A hybrid system for entity recognition from Chinese clinical text [C]//Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing. Chendu: Springer, 2017: 1-6.

[77] LU N J, ZHENG J, WU W, et al. Chinese clinical named entity recognition with word-level information incorporating dictionaries[C]//2019 International Joint Conference on Neural Networks. Budapest: IEEE, 2019: 1-8.

[78] 罗凌, 杨志豪, 宋雅文, 等. 基于笔画ELMo和多任务学习的中文电子病历命名实体识别研究[J]. 计算机学报, 2020, 43(10): 1943-1957.
LUO L, YANG Z H, SONG Y W, et al. Chinese clinical named entity recognition based on stroke ELMo and multi-task learning[J]. Chinese Journal of Computers, 2020, 43(10): 1943-1957. (in Chinese)

[79] LI X Y, ZHANG H, ZHOU X H. Chinese clinical named entity recognition with variant neural structures based on BERT methods[J]. Journal of Biomedical Informatics, 2020, 107: 103422.

[80] 唐国强, 高大启, 阮彤, 等. 融入语言模型和注意力机制的临床电子病历命名实体识别[J]. 计算机科学, 2020, 47(3): 211-216.
TANG G Q, GAO D Q, RUAN T, et al. Clinical electronic medical record named entity recognition incorporating language model[J]. Computer Science, 2020, 47(3): 211-216. (in Chinese)

[81] QIU J H, ZHOU Y M, WANG Q, et al. Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field[J]. IEEE Transactions on NanoBioscience, 2019, 18(3): 306-315.

[82] JI B, LIU R, LI S S, et al. A hybrid approach for named entity recognition in Chinese electronic medical record [J]. BMC Medical Informatics and Decision Making, 2019, 19(Suppl 2): 64.

[83] WANG C Y, WANG H, ZHUANG H, et al. Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree[J]. Journal of Biomedical Informatics, 2020, 111: 103583.

[84] 潘璀然, 王青华, 汤步洲, 等. 基于句子级Lattice-长短记忆神经网络的中文电子病历命名实体识别[J]. 第二军医大学学报, 2019, 40(5): 497-506.
PAN C R, WANG Q H, TANG B Z, et al. Chinese electronic medical record named entity recognition based on sentence-level Lattice-long short-term memory neural network[J]. Academic Journal of Second Military Medical University, 2019, 40(5): 497-506. (in Chinese)

[85] YANG X R, HUANG W K. A conditional random fields approach to clinical name entity recognition[C]//China Conference on Knowledge Graph and Semantic Computing. Tianjin: Springer, 2018: 1-6.

[86] LUO L, LI N, LI S, et al. DUTIR at the CCKS-2018 Task1: A neural network ensemble approach for Chinese clinical named entity recognition[C]//China Conference on Knowledge Graph and Semantic Computing. Tianjin: Springer, 2018: 7-12.

[87] 何云琪, 刘苏文, 钱龙华, 周国栋. 基于句法和语义特征的疾病名称识别[J]. 中国科学: 信息科学, 2018, 48(11): 1546-1557.

[88] 盛剑, 向政鹏, 秦兵, 等. 多场景文本的细粒度命名实体识别[J]. 中文信息学报, 2019, 33(6): 80-87.
SHENG J, XIANG Z P, QIN B, et al. Fine-grained named entity recognition for multi-scenario[J]. Journal of Chinese Information Processing, 2019, 33(6): 80-87. (in Chinese)

[89] LIU M L , ZHOU X S, CAO Z, et al. Team MSIIP at CCKS 2019 Task 1[C]//2019 China Conference on Knowledge Graph and Semantic Computing. Hangzhou: Chinese Information Processing Society of China, 2019: 1-11.

[90] LI N, LUO L, DING Z, et al. DUTIR at the CCKS-2019 Task1: Improving Chinese clinical named entity recognition using stroke ELMo and transfer learning[C]//Proceedings of the 4th China Conference on Knowledge Graph and Semantic Computing. Hangzhou: Chinese Information Processing Society of China, 2019: 24-27.

[91] 赵刚, 张腾, 王晨骁, 等. Team MSIIP at CCKS 2019 Task 2[EB/OL]. (2019) [2022-01-05]. https://conference.bj.bcebos.com/ccks2019/eval/webpage/pdfs/eval_paper_1_2_2.pdf.

[92] JI B, LI S S, YU J, et al. Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models[J]. Journal of Biomedical Informatics, 2020, 104: 103395.

[93] LI Z C, GAN Z, ZHANG B L, et al. Semi-supervised noisy label learning for Chinese clinical named entity recognition[J]. Data Intelligence 2021, 3(3): 389-401

[94] 杨文明, 毕金良, 邹佳丽, 等. 基于ChiEHRBert与多模型融合的医疗命名实体识别[C]//2020年全国知识图谱与语义计算大会. 南昌: 中国中文信息学会, 2020: 1-9.

[95] ZHENG H Y, QIN B, XU M. Chinese medical named entity recognition using CRF-MT-Adapt and NER-MRC [C]//2021 2nd International Conference on Computing and Data Science. Stanford: IEEE, 2021: 362-365.

[96] 温超杰, 陈涛, 朱江. 基于预训练模型和领域词典的医疗命名实体识别方法研究[C]//2020年全国知识图谱与语义计算大会. 南昌: 中国中文信息学会, 2020: 1-11.

[97] MA C, HUANG W K. Named entity recognition and event extraction in Chinese electronic medical records [C]//China Conference on Knowledge Graph and Semantic Computing. Qinhuangdao: Springer, 2022: 133-138.

[98] GAN Z, LI Z C, ZHANG B L, et al. Enhance both text

and label: Combination strategies for improving the generalization ability of medical entity extraction[C]//China Conference on Knowledge Graph and Semantic Computing. Qinhuangdao: Springer, 2022: 92-101.

[99] 晏阳天, 张昕楠, 吴喆, 等. 基于多特征融合的预训练医疗实体和事件抽取模型[C]//2021年全国知识图谱与语义计算大会. 广州: 中国中文信息学会, 2021: 1-8.

[100] WANG Y, LI Y, TONG H H, et al. HIT: Nested named entity recognition via head-tail pair and token interaction [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Virtual Conference: Association for Computational Linguistics, 2020: 6027-6036.

[101] BEKOULIS G, DELEU J, DEMEESTER T, et al. Joint entity recognition and relation extraction as a multi-head selection problem[EB/OL]. (2018-04-20) [2022-01-05]. https://arxiv.org/abs/1804.07847.

[102] EBERTS M, ULGES A. Span-based joint entity and relation extraction with transformer pre-training[EB/OL]. (2019-09-17)[2022-01-05]. https://arxiv.org/abs/1909.07755.

[103] WADDEN D, WENNBERG U, LUAN Y, et al. Entity, relation, and event extraction with contextualized span representations[EB/OL]. (2019-09-08) [2022-01-05]. https://arxiv.org/abs/1909.03546.

[104] KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. (2014-10-22) [2022-01-05]. https://arxiv.org/abs/1412.6980.

[105] LOSHCHILOV I, HUTTER F. Fixing weight decay regularization in adam[C]//2018 International Conference on Learning Representations. Vancouver: ICLR, 2018: 1-14.

[106] 卢宁杰. 结合主动学习的中文医疗命名实体识别研究[D]. 上海: 华东师范大学, 2020.
LU N J. Research on Chinese Medical Named Entity Recognition Combined with Active Learning[D]. Shanghai: East China Normal University, 2020. (in Chinese)

[107] 钟志农, 刘方驰, 吴烨, 等. 主动学习与自学习的中文命名实体识别[J]. 国防科技大学学报, 2014, 36(4): 82-88.
ZHONG Z N, LIU F C, WU Y, et al. Chinese named entity recognition combined active learning with self-training[J]. Journal of National University of Defense Technology, 2014, 36(4): 82-88. (in Chinese)

[108] 李猛, 李艳玲, 林民. 命名实体识别的迁移学习研究综述[J]. 计算机科学与探索, 2021, 15(2): 206-218.
LI M, LI Y L, LIN M. Review of transfer learning for named entity recognition[J]. Journal of Frontiers of Computer Science and Technology, 2021, 15(2): 206-218. (in Chinese)

[109] 王浩畅, 李钰, 赵铁军. 面向生物医学命名实体识别的多Agent元学习框架[J]. 计算机学报, 2010, 33(7): 1256-1262.
WANG H C, LI Y, ZHAO T J. Biomedical named entity recognition through a multi-agent meta-learning framework[J]. Chinese Journal of Computers, 2010, 33(7): 1256-1262. (in Chinese)

[110] SUI D B, CHEN Y B, ZHAO J, et al. Feded: Federated learning via ensemble distillation for medical relation extraction[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Seattle: Association for Computational Linguistics, 2020: 2118-2128.

作者简介

杜晋华　男,2000年4月出生于山西省大同市. 现为清华大学北京信息科学与技术国家研究中心博士研究生. 主要研究方向为机器学习、大数据与自然语言处理.
E-mail: dujh22@mails.tsinghua.edu.cn

尹　浩(通讯作者)　男,1974年10月出生于湖南省益阳市. 现为清华大学北京信息科学与技术国家研究中心研究员、博士生导师. 主要研究方向为计算机网络、大数据与区块链.
E-mail: h-yin@mail.tsinghua.edu.cn

冯　嵩　男,1970年9月出生于湖南省常德市. 现为中南大学湘雅医院高级工程师. 主要研究方向为医疗信息化.
E-mail: fs@sina.com