计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(9):70−81 [doi: 10.15888/j.cnki.csa.008708]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度学习的医疗命名实体识别①

贾杨春, 朱定局

(华南师范大学 计算机学院, 广州 510631)
通信作者: 朱定局, E-mail: zhudingju@m.scnu.edu.cn

摘　要: 医疗命名实体识别指从海量的非结构化的医疗数据中提取关键信息, 为医学研究的发展和智慧医疗系统的普及提供了基础. 深度学习运用深层非线性的神经网络结构能够学习到复杂、抽象的特征, 可实现对数据更本质的表征. 医疗命名实体识别采用深度学习模型可明显提升效果. 首先, 本文综述了医疗命名实体识别特有的难点以及传统的识别方法; 其次, 总结了基于深度学习方法的模型并介绍了较为流行的模型改进方法, 包括针对特征向量的改进, 针对数据匮乏、 复杂命名实体识别等问题的改进; 最后, 通过综合论述对未来的研究方向进行展望.

关键词: 命名实体识别; 深度学习; 循环神经网络; 自然语言处理; 注意力机制; 人工智能

引用格式: 贾杨春,朱定局.基于深度学习的医疗命名实体识别.计算机系统应用,2022,31(9):70–81. http://www.c-s-a.org.cn/1003-3254/8708.html

Medical Named Entity Recognition Based on Deep Learning

JIA Yang-Chun, ZHU Ding-Ju

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: Medical named entity recognition refers to the extraction of key information from massive unstructured medical data, which provides a foundation for the development of medical research and the popularization of smart medical systems. Deep learning uses deep nonlinear neural network structures to learn complex and abstract characteristics, which can represent data more essentially. Deep learning models can significantly improve the effect of medical named entity recognition. First, this study introduces the unique difficulties and traditional methods of medical named entity recognition. Then, it summarizes models based on deep learning and popular model improvement methods, including the improvement of feature vectors and the ways to deal with difficulties such as a lack of data and the recognition of complex named entities. Finally, the study provides an outlook on future research direction through a comprehensive discussion.

Key words: named entity recognition (NER); deep learning; recurrent neural network (RNN); natural language processing (NLP); attention mechanism; artificial intelligence

随着现代化医疗系统的普及, 如今已产生了海量的医疗数据, 如诊断报告、临床研究数据、药品说明以及电子病历等. 医疗数据的有效利用对医学研究、医疗诊断和公共防疫等方面起着至关重要的作用, 比如, 统计研究医学临床数据可为医务人员诊疗决策提供信息支撑; 又比如, 精准医疗通过对患者的基因大数据进行挖掘与分析, 可为其提供有针对性的诊疗方案; 还有, 对于疫情防控, 跨省市的医疗大数据的合理利用可协助专家及时、全面地提出防控建议, 实现对医疗资源的合理配置. 然而, 医疗数据一般是半结构化或非结构化的文本, 对后续的研究带来了一定难度. 因此, 对海量的非结构化医疗文本通过数据分析与挖掘的方式来获取有价值的医学数据已成为一个研究热点.

命名实体识别 (named entity recognition, NER)[1]

① 基金项目: 广东省普通高校“人工智能”重点领域专项 (2019KZDZX1027); 中国高等教育学会专项课题 (2020JXD01); 广东高校省级重点平台和重大科研项目 (2017KTSCX048); 广东省中医药局科研项目 (20191411)
收稿时间: 2021-12-06; 修改时间: 2022-01-11; 采用时间: 2022-02-17; csa 在线出版时间: 2022-06-24

指从自然语言文本中发现特定的目标实体, 对文本信息结构化起着十分重要的作用. 医疗命名实体识别指从医疗文本中识别医疗实体的边界并判断医疗实体的类别, 常见的医疗实体类别包括疾病名称、身体部位、药品信息、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着事件抽取、关系抽取等任务的效果, 是医疗文本数据挖掘的关键任务, 为构建健康医疗系统、智能医疗问答系统、医疗知识图谱提供了关键基础.

如今, 命名实体识别技术在学术界已经较为成熟, 但还无法较好的应用到工业界中. 主要是因为不同领域, 有不同的语言风格以及规则, 命名实体模型的泛化能力差, 无法找到一个统一高效的模型. 如何将命名实体识别任务部署在医疗领域, 首先需要分析该领域的特点, 进而总结出需要解决的难点. 现有的 BiLSTM-CRF 模型是该任务的主要模型, 为了提升模型效果, 需要在传统模型的基准上加以优化改进. 本文按照本思路, 在第 1 节总结了命名实体识别在医疗领域面临的难点, 第 2 节和第 3 节总结了传统的方法以及基于深度学习的方法, 在第 4 节介绍了集中主流的改进方法, 包括针对特征向量、数据匮乏、复杂命名实体识别等问题的改进.

1 医疗命名实体识别的难点

分析研究医疗命名实体识别的特有难点, 对提高实体识别效果具有一定的指导意义. 对于医疗命名实体识别的难点, 一方面是医疗文本特有的语言特点给实体识别任务带来的困难, 另一方面是复杂的命名实体难以准确地被识别, 还有就是命名实体识别应用到医疗领域所面临的数据匮乏、可迁移性差、可解释性弱等问题.

1.1 医疗文本的语言特点

中文医疗文本的特点主要集中在中文的语言特点以及医疗文本的语言特点两个方面.与英文不同, 中文命名实体识别的难度更大, 主要因为中文文本没有显著的大小写特征以及单词特征, 相关实体边界难以确定. 容易造成以下问题: (1) 中文分词错误. 若采用基于词的命名实体识别, 会因不正确的中文分词导致命名实体识别错误, 如“痛风性关节炎”会被错误识别为“痛风”和“关节炎”两个实体. (2) 语义信息无法完整提取. 若采用基于字的方法, 可避免分词错误, 但没有考虑文本中词和词边界信息, 而这些词义信息可能对识别效果有潜在的提升效果. (3) 字、词多义问题. 无论是基于字还是基于词的命名实体识别都无法避免由于同一个字、词在不同上下文中的含义不同所造成的歧义问题.

医疗文本中存在着大量的专业术语、英文缩写等特殊表达, 以及很多如嵌套、略写等不规范的表达, 这给医疗命名实体识别带来了一定的难度, 主要表现在以下几个方面: (1) 专业性强是医疗领域的一大特点. 医疗文本中存在着大量的医学名词和专业术语, 没有医学背景的非专业人士对其很难理解, 无法对其准确标注. (2) 医疗文本中存在很多英文缩写. 医务人员为了简便以及提高通用性与可读性, 通常使用英文缩写代替复杂的中文名称, 但命名实体识别模型很难对其辨别. 如“HR”是“心率”的英文缩写, “BP”是“血压”的英文缩写. (3) 大量的医学名词都是由外文音译而来, 同一个英文医学名词可能对应着不同的音译而来中文名词, 给命名实体识别引入了噪声. 如“克雷伯杆菌”和“克雷白杆菌”“艾乐替尼”和“阿来替尼”. (4) 医疗文本中存在句子成分缺失、表达不完整现象.如缩略表达“畏寒 (经常怕冷感)”“律齐 (心跳正常)”. (5) 医疗实体名词种类繁多, 数量庞大, 并且, 随着现代化医疗技术的发展, 未登陆词不断涌现, 难以构建一个系统全面的医学词典.

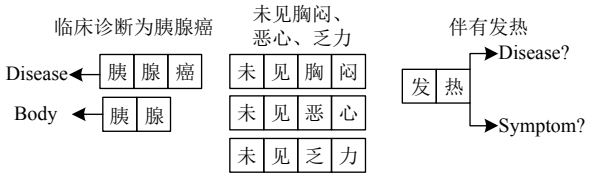
1.2 复杂命名实体识别

对于医疗命名实体识别而言, 复杂命名实体通常难以被准确地识别. 主要包括嵌套医疗命名实体、类别易混淆命名实体, 非连续表达医疗命名实体. 实例如图 1 所示, 其中嵌套命名实体是指一个命名实体中存在多个其他的命名实体, 如“胰腺癌”指一个疾病实体, 而“胰腺癌”中的“胰腺”指的是身体部位实体; 非连续命名实体指由于表达不完整造成歧义现象, 如“未见胸闷憋气、恶心呕吐、乏力”, 应指“未见胸闷憋气, 未见恶心呕吐, 未见乏力”; 类型易混淆命名实体指的某些医疗命名实体所属多个类别, 无法准确判断当前实体属于哪一类别, 如“发热”一般是指一种疾病, 但有时也可作为症状.

1.3 领域命名实体识别所面临的问题

近些年来, 随着深度学习技术的发展, 命名实体识别在学术界已获得了很多不错的成果, 但在工业界还面临着很多挑战, 针对本文所研究的医疗领域, 主要存在以下问题: (1) 数据匮乏. 医疗实体标注是一个较为复杂的任务, 对专业知识以及标注规范的要求较高, 标注成本高昂, 现存的医疗数据集很稀缺, 这对实体识别任务带来了很大的挑战. 除此之外, 在数据层面医疗命名实体识别任务还面临着冷启动、噪声、数据不平衡等问题. (2) 可迁性差. 不同的医疗系统, 不同的医院,

不同的医药厂商对医疗文本的描述不同, 并且具有不同的规范, 这会导致医疗命名实体识别技术的可迁移性差. (3) 可解释性弱. 深度学习端到端的过程会导致模型的可解释性弱, 而医疗领域是一个极度严谨的领域, 使用完全的黑箱算法是不可取的.



(a) 嵌套命名实体 (b) 非连续命名实体 (c) 类型易混淆命名实体

图 1 复杂命名实体实例

2 传统医疗命名实体识别方法

2.1 基于规则的方法

基于规则的医疗命名实体识别方法依赖于手工制定的规则, 即利用领域专家制定规则模板结合医学词典通过模式匹配的方法识别出医疗实体. 规则一般包括句法、语法以及医疗领域知识. 例如, “胰腺癌多发于成年女性, 症状通常为腹部疼痛”, 现有规则: (1) [身体部位+癌/炎/瘤/…] 代表一种疾病. (2) [身体部位+疼/痛/疼痛]代表症状, 通过模式匹配能够识别出疾病实体“胰腺癌”和症状实体“腹部疼痛”. 早期, 有学者研究如何设计处理医疗文本信息的系统以实现信息提取、构建知识库、信息编码等功能, 这些系统一般是基于模式匹配结合领域知识来设计的. Canfield 等人[2] 通过分析临床报告中的语义和句法结构, 结合医学词典构建医疗信息处理模型. Sager 等人[3] 设计了一种文本信息提取系统并将其运用在医疗领用, 该系统通过分析医学文本特有的结构以及语法规则, 可提取医学报告、临床记录中的关键信息. Friedman 等人[4] 提出了一个通用的医疗文本信息提取系统 MEDLEE, 利用领域知识库和专家总结的规则实现其功能, 具有很好兼容性. Zingmond 等人[5] 通过分析统计医学语料中的规则结合语言处理工具构建了一个文本处理器, 用于处理医学文本报告. 这些早期的医疗文本信息处理系统为基于规则的医疗命名实体识别提供了基础. 特定的领域具有特有的语言规则和词典, 当专家能够较为完备地总结出目标领域的规则时, 基于规则方法在领域命名实体识别将会有很好的表现. 李楠等人[6] 通过分析化学文献中化学物质命名的构词规律, 总结化学领域的启发式规则, 有效提高了该领域实体识别的准确率.

2.2 基于机器学习的方法

对于基于机器学习的方法, 命名实体识别可被形式化为实体标签分类任务或文本序列标注任务. 专家应用医学领域知识与特征工程对样本数据进行表征, 利用大量标注好的医疗数据作为训练样本, 然后应用机器学习算法训练模型使其对数据的模式进行学习, 即可使用训练好的模型实现标签分类或序列标注任务: 对于实体标签分类任务, 常用的模型有支持向量机 (support vector machine, SVM)[7]、决策树 (decision tree)[8] 等, 将医疗语料中每个字符的标签当做一个类别进行文本分类; 条件随机场 (condition random gield, CRF)[9]、隐马尔可夫模型 (hidden Markov models, HMM)[10] 等模型是文本序列标注任务常用的模型, 通常将命名实体识别任务理解为一个最大概率序列问题, 即根据观测序列 (一般指字符) 预测隐藏序列 (一般指字符的标签). Li 等人[11] 使用基于隐马尔可夫模型的方法训练医疗临床笔记, 识别临床笔记中的各个模块, 该方法的准确率到达了 93%, 明显优于基线. Zhou 等人[12] 为了高效挖掘医学文本中的信息, 提出了一个识别生物医学实体的系统, 该系统通过基于 HMM 的命名实体识别器集成如构词模式、词性、语义触发等针对生物医学领域的特征, 评价显示, 该系统能够有效处理实体嵌套问题, 在 GENIA 语料库中的 $F1$ 值可到达 90%. Lee 等人[13] 使用基于 SVM 的将方法命名实体识别分为识别和语义分类两个子任务, 可以解决实体类别过多并且分布不均匀对识别效果造成的影响. Settles 等人[14] 使用条件随机场 (CRF) 结合丰富的特征集实现在生物医学领域的命名实体识别. 在中文领域, 叶枫等人[15] 提出使用基于 CRF 的方法识别电子病历的实体, 通过构建医学数据的特征模板用小规模的语料库训练模型, 获得了较为理想的 $F1$ 值. 燕杨等人[16] 提出了一种基于级联条件随机场模型, 识别电子病历中的复杂疾病和临床症状的嵌套实体, 与传统的 CRF 相比, $F1$ 值提高了 7%.

在领域词典足够完善的情况下, 当制定的规则能够对目标领域文本的特征精准描述时, 基于规则的方法将会有比其他方法更好的表现. 但是, 制定领域规则模板和维护领域词典耗时耗力, 并且对专业知识的要求很高, 不同的领域具有不同的规则与词典, 导致该方法的可迁移性较差. 基于机器学习的方法取得了很大的进展, 在一定程度上改善了上述问题, 降低了对医学

领域知识的要求, 但基于机器学习的方法需要大量人工标记的数据集对模型参数进行训练, 而现有可用的大规模医疗数据集比较稀缺. 并且, 基于机器学习的方法需要专家手动选择对命名实体识别任务有影响的各种特征, 但特征提取通常是困难且昂贵的.

3 基于深度学习的方法

深度学习[17] 运用深层非线性的神经网络结构能够学习得到更复杂、抽象的特征, 以实现对数据更本质地表征. 与传统的机器学习不同, 深度学习不依靠人工识别特征, 可以自动提取特征. 基于深度学习的方法在命名实体识别任务上取得了不错的效果, 受到了研究人员的广泛关注.

基于深度学习的命名实体识别任务的网络构架一般分为 3 类: 卷积神经网络 (convolutional neural networks, CNN)[18], 循环神经网络 (recurrent neural networks, RNN)[19] 以及基于自注意机制的 Transformer[20]. 曹依依等人[21] 提出用基于 CNNs 的模型处理医疗领域的 NER, 作者采用迭代扩张卷积作为编码器提取特征, 降低了模型训练难度并实现了并行运算. Liu 等人[22] 采用循环神经网络的变体 LSTM 建模, 用基于 BiLSTM-CRF 的模型识别医学文本中的健康信息和临床实体, 该方法避免了繁琐的特征工程, 在 i2b2 数据集上提取医学实体, *F*1 值达到了 94.37%, 证明了该模型的有效性. 李博等人[23] 考虑使用基于 Transformer 的模型在自建数据集上识别医疗实体, 有效提高了识别准确率和改善了对较长语句的处理性能. 表 1 总结了部分比较有代表性的基于深度学习方法的论文.

表 1 基于深度学习方法的典型论文

特征提取器	模型	领域	数据集	语言	创新点	*F*1值 (%)
CNNs	CNN-CRF[24]	通用领域	CoNLL	英文	开创性提出使用CNN-CRF处理NER	89.59
	门控CNN-CRF[25]	通用领域	SIGHAN	中文	采用门控单元过滤CNN的输出, 改善梯度弥散; 可实现并行计算	90.49
	LR-CNN[26]	通用领域	Resume数据集	中文	融合词典信息; 增加反馈层, 利用全局信息	95.11
	IDCNN-CRF[21]	医疗	CCKS	中文	采用迭代扩张卷积改善过拟合现象, 训练参数少	90.31
RNNs	BiLSTM-CRF[27]	通用领域	CoNLL	英文	用BiLSTM-CRF处理NER的开山之作	90.10
	BiLSTM-CRF[28]	生物医学	JNLPBA等	英文	提高了对特殊字符的识别准确率, 采用CNN捕获更细粒度的字符级特征	89.09
	BiLSTM-CRF[29]	生物医学	GENIA	英文	解决医学实体嵌套问题, 在基础模型上引入一个双向LSTM子分类模型	70.08
	BiLSTM-CRF[22]	医疗	i2b2	英文	—	94.37
	BiLSTM-CRF[30]	通用领域	MSRA	中文	将偏旁级特征结合到字符向量中	90.95
	BERT-BiLSTM-CRF[31]	医疗	CCKS	中文	引入BERT预训练语言模型	88.45
Transformers	Transformer-CRF[23]	医疗	自建数据集	中文	—	最好达到95.02
	Transformer-CRF[32]	通用领域	Resume数据集	中文	引入相对位置编码	94.7

3.1 命名实体识别模型的一般构架

构建命名实体识别模型时, 需要把不可计算的自然语言文本进行向量化表示. 传统的独热码编码根据每个字或词在词汇表中的位置信息, 为其分配一个长度为词汇表大小的向量, 实现文本编码. 独热码编码方法简单、便于实现, 但存在一些本质的缺点: (1) 将各个字或词都视为是独立的, 无法体现某些字词之间存在的潜在联系, 比如“肺”和“肝”都为身体器官, 但用独热码编码方法会忽略它们之间的相似性. (2) 向量维度为字典大小, 且只有一位是有意义的, 具有离散稀疏性. (3) 无法考虑词序信息. (4) 有些词不在词汇表中, 无法对其编码. 词嵌入方法在一定程度上改善了以上问题, 它将词用一个定长的低维实数向量表示以实现文本向量化, 预训练词向量模型是实现词嵌入较为主流的方法. Word2Vec[33] 是一种预训练词向量的工具, 对词之间的相似性能够较好地“刻画”. 一个词的上下文信息能够很大程度地决定该词的语义信息, Word2Vec 能够捕获目标词的上下文信息并融入该词的词向量中, 包括通过上下文中的词预测中心词的 (continuous bag-of-words model, CBOW)[34] 和通过中心词预测上下文中的词的跳字模型 (continuous skip-gram model, Skip-gram) 两种语言模型. GloVe[35] 是对 Word2Vec 词嵌入方法的扩展, 采用指定大小的窗口构建自然语言文本中词的共现矩阵, 把全局词频统计与 Word2Vec 的基

于上下文的学习结合起来, 以实现将全局语义信息考虑到词向量中. BERT[36] 自提出后就受到了广泛的关注, 该模型基于深层双向 Transformer 结构能动态地生成词向量, 并且使词向量中蕴含了更多的语义、语法知识, 有效解决了一词多义问题, 在众多自然语言处理任务上都取得了令人惊叹的效果.

基于深度学习的医疗命名实体识别模型如图 2 所示, 包括输入层、嵌入层、编码层、解码层和输出层. 嵌入层的语言模型可从输入的医疗文本中学习到语义、语法知识并将其向量化表示, 主要有基于词的表示、基于字的表示以及基于字信息和词信息的混合表示; 编码层使用 CNN、RNN、Transformer 等特征提取器对嵌入层传入的信息进行特征提取并编码; 解码层利用解码器对编码层的输出结果进行标签预测, 最终输出最佳标签序列. 有不少研究基于 3 种特征提取器, 加以改进, 实现更高的识别性能.

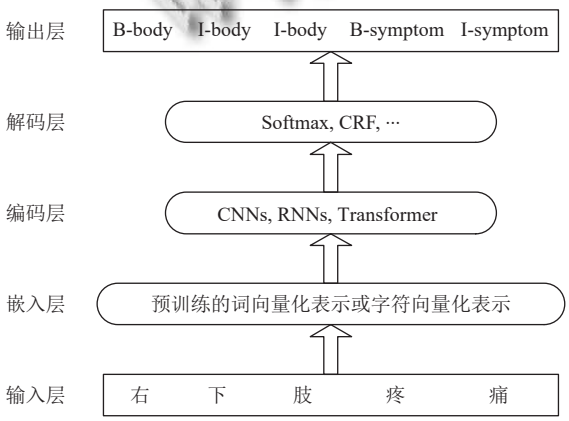


图 2 命名实体识别模型框架

标注命名实体识别的数据集需要标注出实体边界和类别, 常用方案有: (1) BIO. 该方法用符号 B 和 I 标记实体边界, B 表示一个命名实体的开始位置, I 代表处于命名实体内部的位置, 用符号 O 标记非实体或非目标类别的实体. (2) BIOES. 相较于前一种方法, 该方法对实体边界的描述更详加细, 用符号 E 指一个命名实体的结束边界, 符号 S 用来表示仅包含一个字的实体名称. 对于 NER 任务, 有基于字级别和基于词级别两种标注粒度, 图 3 是一个基于字符的 BIO 标注示例, 通常用 [B/I/O-实体类别] 格式对一个字标注.

右 B-body 下 I-body 肢 I-body 伴 O
有 O 疼 B-symptom 痛 I-symptom

图 3 BIO 标注示例

3.2 基于 CNNs 的模型

基于 CNNs 的医疗命名实体识别模型的一般结构如图 4 所示, 嵌入层将输入的一维文本序列转化为二维向量并将其传递给卷积层, 通过指定数量的卷积核在输入的向量矩阵上进行窗口移动来提取文本中的特征, 得到的特征向量通过池化层的平均池化或极大池化操作后被传递给 CRF 层, 最后输出对应的标签序列. Collobert 等人[24] 开创性地使用基于 CNNs 的模型解决 NER 问题. 在此基础上, Gui 等人[26] 将词典信息融合到基于 CNNs 的模型, 处理中文 NER 任务, 该模型实现了并行处理, 并且应用新颖的反馈机制来解决词之间的歧义问题, 表现出较好的性能. 陶源等人[25] 则运用门控卷积过滤卷积层的输出, 有效解决了长距离依赖问题.

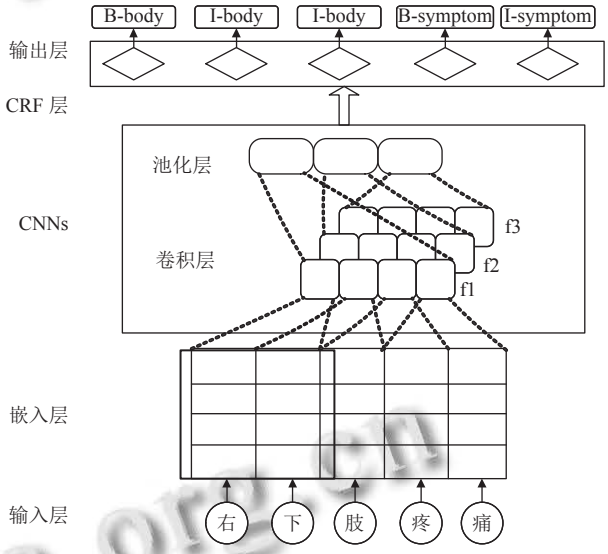


图 4 基于 CNNs 的命名实体识别模型结构

然而, 基于 CNN 的模型会存在以下问题: (1) 对于传统的 CNN 而言, 由于受到感受野的限制, 卷积后得到的只是局部特征, 无法捕获远距离的特征, 对于较长的医疗文本序列来说效果不佳. 越深的卷积层可捕获越远距离的特征, 通过增加 CNN 的神经网络层数可在一定程度上解决上述问题, 但这会给训练模型带来很大的难度. Strubell 等人[37] 提出用膨胀卷积增加序列覆盖距离以得到更多的上下文信息, 该方法在命名实体识别任务中能兼顾运算速度和长序列的特征提取. 主要原理为: 卷积核窗口在输入序列的矩阵上以指定大小的间隔跳跃滑动, 以相同大小的卷积核获得比传统卷积核更大的感受野. 作者还提出重复应用相同的卷积块实现参数共享, 这一迭代过程可以防止过拟合现象. (2) 经过池化操作后得到的特征向量会丢失位置信

息, 但位置信息对文本序列的命名实体识别来说却很关键.

3.3 基于 RNNs 的模型

RNN 能够学习到序列数据中的上下文语义与时序信息, 通常用来处理线性序列数据. 图 5 展示了循环神经网络的结构, 循环节点共享权重矩阵 W 并按时序链式连接, 某个时刻的节点输出依赖于当前时刻的输入 (即当前节点输入的词向量x_t) 和上一时刻的输出 (即上一节点的隐藏状态向量h_{t-1}). RNN 的结构使其能够处理不同长度的序列文本, 但序列过长会因梯度弥散使模型的输出被近距离的信息主导, 难以学习到远距离的依赖关系. 而长短期记忆网络 (long short-term memory, LSTM)[38] 可以选择性地“记忆”过去的关键信息以实现长期记忆, 有效改善梯度弥散现象, 使命名实体识别任务的效果显著提升. Huang 等人[27] 开创性地使用 BiLSTM-CRF 模型. 模型解决 NER 问题, 在此基础上, Dong 等人[28] 采用 CNN 捕获更细粒度的字符级特征, 提高了对特殊字符的识别准确率; Li 等人[29] 在基础模型上引入一个双向 LSTM 子分类模型, 在一定程度上解决了医学实体嵌套的问题; 下文将详细介绍 BiLSTM-CRF 模型.

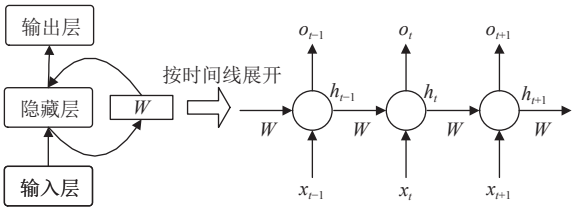


图 5 循环神经网络结构图

LSTM 的一个循环单元的模型结构如图 6 所示, 包含 4 个相互作用神经网络层, 主要通过门控机制对单元状态C中的信息更新以达到控制信息传递、避免长距离依赖问题的目的. 其中, 遗忘门利用门控函数判断上一节点的单元状态c_{t-1}对当前单元状态c_t的重要性, 对c_{t-1}中的信息部分“忘记”并将其融入c_t状态向量中; 输入门对当前输入信息x_t进行选择性“记忆”并将其选择性地加c_t中; 输出门决定当前节点的单元状态信息c_t要输出多少给当前隐藏状态h_t. LSTM 用当前输入x_t和上一时刻的隐藏状态h_{t-1}作为输入训练各个“门”的控制信号, 对隐藏状态h_t更新并得到当前输出y_t, 其计算公式如式 (1)–式 (3), 其中, W 和 b 为需要学习的参数:

$$\begin{bmatrix} \tilde{c}_t \\ f_t \\ i_t \\ o_t \end{bmatrix} = \begin{bmatrix} \tanh \\ \sigma \\ \sigma \\ \sigma \end{bmatrix} \left(W \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix} + b \right) \tag{1}$$

$$c_t = f_t \odot c_{t-1} + x_t \odot \tilde{c}_t \tag{2}$$

$$y_t = o_t \odot \tanh(c_t) \tag{3}$$

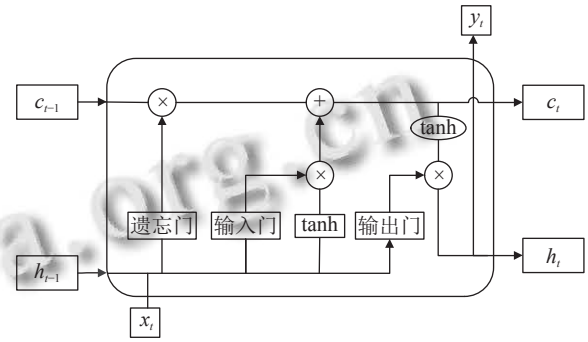


图 6 LSTM 循环单元

条件随机场是在给定一组输入随机变量的条件下得到另一组输出随机变量的条件概率分布模型, 常用于处理序列标注任务. 对于命名实体识别, CRF 可向编码层输出的标签序列中添加一些约束以条件保证输出结果的合理性, 例如: (1) 用 B 表示一个实体名称的开始边界, “左下肢”中“左”的标签应该是“B-body”而不能是“I-body”. (2) 同一个实体对应标签的类别是相同的, “左下肢”的标签为“B-body I-body I-body”, 类别都是“body”, 而不能是“B-body I-symptom I-body”或其他.

CRF 通过学习到的特征来表征各标签之间的约束关系, 使用状态特征函数和转移特征函数对标签序列进行评估. 假设给定输入序列:

$$X = (x_1, x_2, \cdots, x_n)$$

对应的输出序列为:

$$Y = (y_1, y_2, \cdots, y_n)$$

定义得分函数为:

$$F(X, Y) = \sum_{i=1}^{n} M_{i,y_i} + \sum_{i=1}^{n} N_{y_i, y_{i+1}} \tag{4}$$

其中, $F(X,Y)$表示输入文本序列X对应的实体标签序列为Y的概率分数, M_{i,y_i}代表第 i 个字符被标记为标签y_i的概率, 指由当前文本自身特征所对应的标签的得分; N 是转移矩阵, $N_{y_i,y_{i+1}}$表示标签y_i的下一个标签是y_{i+1}的概率, 即上下文、外部字典及规则等特征对当前实体标签的影响. 求出最大的概率分数, 即可得到当前最佳

的输出标签序列.

医疗文本含有大量的专业术语和专有名词等特殊实体, 需要依赖上下文信息才能准确提取这些实体, LSTM 网络能够充分考虑上下文语义信息以及长距离依赖问题. 为了增强循环神经网络对下文信息的提取, 可以考虑添加一层从序列末尾开始处理的逆序 LSTM. 条件随机场作为序列标注算法, 能有保证输出序列的具有一定的有效性. BiLSTM-CRF 模型其结构如图 7 所示, 该模型通过两层 LSTM 对嵌入层中的序列信息进行编码, 并将这两层的编码结果连接在一起输入 CRF 层, 解码输出对应的最佳标签序列.

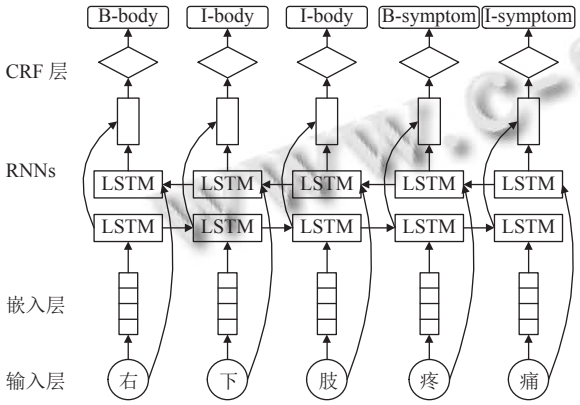


图 7 BiLSTM-CRF 命名实体识别模型

3.4 基于 Transformer 的模型

Transformer[20] 是一种基于注意力机制, 旨在解决序列到序列的网络结构. 由前文可知, 传统的基于 CNNs 的模型受感受野的限制无法捕获远距离的特征; 基于 RNNs 的模型虽然能较好地解决长期依赖的问题, 但模型结构使其无法实现并行运算, 运行速率较慢. Transformer 可以避免以上模型的缺点, 该网络结构没有循环结构, 能够对序列中的单词或字符并行处理, 借助注意力机制对序列中所有字或词之间的关系进行建模, 可以解决长期依赖的问题. 基于 Transformer 的模型在医疗命名实体任务上获得了不错的效果, 其模型如图 8 所示, 将预训练的字向量结合位置编码作为 Transformer 的输入, 提取文本序列的特征信息, 然后通过 CRF 预测出最佳标签序列并输出.

虽然 Transformer 在命名实体识别任务上取得了非常不错的效果, 但是该模型也有一定的局限性: (1) 简单地抛弃 RNN 和 CNN 使 Transformer 无法捕获文本序列的局部特征. (2) 位置编码并不能改变 Transformer 无法捕获位置信息这一固有的结构缺陷.

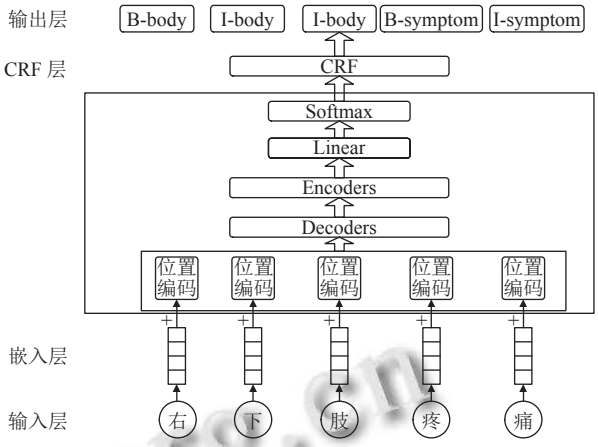


图 8 基于 Transformer 的命名实体识别模型结构

4 针对医疗命名实体识别的模型改进

近年来, 提升命名实体识别模型的性能引起了广泛关注, 具有一定的现实意义. 本小节总结了医疗领域命名实体识别模型较为流行的改进方法, 主要有: 针对嵌入层特征向量的改进, 融合词典信息, 拼音、偏旁特征以丰富嵌入层的特征向量; 为解决数据匮乏问题, 引入迁移学习的方法; 通过引入注意力机制可以提高模型的计算能力并且能有效解决长距离依赖问题; 针对复杂命名实体难以被识别的问题, 从不同角度对模型进行改进.

4.1 针对特征向量的改进

通常, 精心构造的底层特征向量可以显著提升模型的识别效果. 如引入词典信息, 可以有效避免分词错误. 再如, 通过融合拼音特征、偏旁特征来丰富嵌入层的词向量, 能够提高对“多义词”“象形字”的识别准确率. 对嵌入层的特征向量进行改进, 引入丰富的特征, 是一个提升模型效果的可行方法.

在特征向量中融合词信息可以避免中文分词错误, 命名实体识别通常采用基于字符的方法, 但这种方法忽略了文本序列中很多与词相关的语义信息, 容易带来歧义. 因此, 很多 NER 模型将词典特征融合到输入的字符序列中, 既可避免分词错带来的影响, 还可把潜在的词信息融入到特征向量中, 从而提高对实体边界识别的准确性. Zhang 等人[39] 首次提出使用晶格结构 (lattice) 获取文本序列中潜在的词信息, 有效利用序列中与词相关的语义信息还避免了分词错误. 在医疗领域, 张笑天[40] 提出了一种基于 Lattice-LSTM 的医疗文本命名实体识别模型, 并在嵌入层使用大量医学字典

训练词向量模型, 整体提升了命名实体识别效果.

医疗文本中有许多外文音译而来的医学名词, 一词多“译”给医疗命名实体识别带来了一定的难度, 同音异字是音译名词的常见现象; 大部分汉字的偏旁含有一定的语义信息, 如“脚”“腿”中的“月”字旁代表身体部位, “痛”“瘤”中的“疒”字旁代表与疾病相关的症状. 根据这些中文医疗文本的特点, 可对医疗命名实体识别的词嵌入模型加以改进: 第一, 增加拼音特征, 便于对音译词的识别; 第二, 增加偏旁特征, 从而增强汉字本身的语义. Dong 等人[28] 将来自字典的偏旁级特征结合到字符级向量中, 采用基于 LSTM 的模型处理中文领域的 NER 任务, 在 MSRA 数据集上, $F1$ 值取得了 90.95% 优秀表现. Sun 等人[41] 提出一种的名为 Chinese-BERT 模型, 该模型能够利用上下文特征和汉字本身的语义特征, 不仅将基于汉字字形的特征融入到特征向量中, 还考虑了基于汉字拼音的特征, 通过对比实验, 该模型表现出其明显的性能优势. 在医疗领域, Yin 等人[42] 使用 CNN 来提取汉字的偏旁特征, 结合字符级嵌入的模型识别医疗文本中的实体名称, 有效利用了医疗文本中汉字本身隐含的语义信息. 图 9 展示一种构建在字符向量中融合拼音、部首特征的方法, 新的特征向量由字符向量 C 与特征向量 W 的加和构成.

4.2 针对长距离依赖问题的改进

医疗文本数据中较长语句较多, 当处理的文本序列过长时, 传统的基于 CNNs 的模型受感受野的限制无法捕获远距离的特征; 而基于 RNNs 的模型虽然能较好地解决长期依赖的问题, 但模型结构使其无法实现并行运算, 运行速率较慢. 通过引入注意力机制可以挺高模型的计算能力并且能有效解决长距离依赖问题.

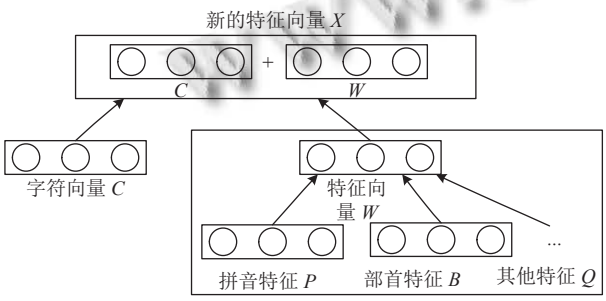


图 9 融合多特征的特征向量

注意力机制 (attention mechanism)[43] 是一种模拟人类视觉的信号处理机制, 能够将有限的注意力选择性地分配给需要关注的部分. 在命名实体识别任务中, 融入注意力机制能够使 NER 模型捕获到更为关键的句法和高层语义特征. 医疗文本中有很多超长语句, 合理利用上下文信息对正确识别出目标实体具有很大的意义, 虽然基于 RNNs 模型可以有效利用上下文信息, 但它无法体现上下文信息与当前信息的相关程度. 引入注意力机制根据各个字词对正在识别的目标实体的重要程度, 分配相应大小的关注程度, 从而优化资源配置, 提高识别效果. 例如“门诊以口唇疱疹、低烧收入我科, 考虑手足口”“门诊、收入我科”等对识别出疾病实体“手足口”作用不大, 而“口唇疱疹”“低烧”对识别出实体“手足口”却起着很大的作用, 因此注意力机制会把更多的注意力资源分配给“口唇疱疹”和“发烧”.

Luo 等人[44] 在命名实体识别模型中引入注意力机制, 用于生物医学领域的实体识别, 该模型以很少的特征工程获得了比其他最先进的方法更好的表现, 在 CHEMDNER 和 CDR 语料库上的 $F1$ 值达到 91.14% 和 92.57%. 单义栋等人[45] 从军事领域的文本中识别实体, 通过引入注意力机制和融合词向量的方法提供对实体识别任务更为关键的特征, 整体提升了模型的性能. 融合注意力机制在命名实体识别任务中获得了不错的效果, 成为了开放领域命名实体识别最好的模型之一[46–49].

4.3 针对数据匮乏问题的改进

医疗命名实体识别离不开医疗数据集的支撑, 标注好的大规模数据集十分稀缺, 并且由于医疗领域的特殊性, 很多医疗领域的数据都涉及隐私问题. 数据匮乏使得命名实体模型无法对特征进行准确的表达, 识别效果不好, 针对医疗数据匮乏的问题, 融合迁移学习的方法被广泛使用.

迁移学习[50] 指将从“源”数据集中学到的知识应用在“目标”任务中, 即利用源域中的标注数据或知识结构, 通过微调模型等方法, 完成或改进目标任务的学习效果. 训练医疗领域的 NER 模型需要大规模标注好的医疗数据集, 但医疗数据具有一定的特殊性, 并且需要有一定专业背景的人来标注, 导致现有的可训练数据集很少, 融合迁移学习可有效解决医疗命名实体识别中数据匮乏的问题[51]. Giorgi 等人[52] 使用 SSC 语料库 (大规模数据集, 含有噪声) 代替 GSC 语料库 (手工标记的数据集, 高度可靠) 训练生物医学领域的 NER 模型, 通过 SSC 到 GSC 的迁移, 既可以利用大规模 SSC 数据集训练模型, 又可以利用 GSC 数据集对模型

进行优化以减少噪声. Wang 等人[53] 提出一种融合迁移学习的 NER 模型, 该模型引入标签感知机制, 实现了医疗 NER 模型的特征和参数在不同专业间迁移.

4.4 针对复杂实体识别问题的改进

由于医疗文本中医学名词构词复杂, 复杂命名实体在医疗文本中占比很大, 如, 据统计生物医学数据集 GENIA 中含有嵌套实体的语句占到了 30%, 因此对于医疗命名实体识别而言, 复杂命名实体识别的问题不可忽视.

嵌套实体. 传统模型在识别命名实体时, 每个字符对应一个标签, 无法解决嵌套实体一个字符对应可能对于多个标签的问题. 对于嵌套实体而言, 它的构成复杂多变, 无法找到一个统一的规则对它进行“刻画”, 一般需对模型加以改进提高识别准确率. 主流的处理方法包括: 多层序列标注法, 增加模型编码器和解码器的层数将多个标签分配给一个字符, Ju 等人[54] 提出一种动态层叠式模型, 通过堆叠多个 Flat NER 层, 从内层到外层识别命名实体, 每当该模型识别出命名实体, 就会在当前基础上堆叠一个新的 Flat NER 层以识别更外层的命名实体, 直到没有更外层的命名实体被识别出来为止. 但是, 由于该模型识别内层实体时无法考虑外层实体信息, 会在一定程度上造成级联错误; 基于区域的识别方法, 抽取序列中所有可能存在实体的子序列区域来识别出所有实体, 该方法可以有效避免上述方法级联错误. Sohrab 等人[55] 提出一种基于区域的方法, 将序列中所有可能存在实体的区域进行编码, 然后通过一个分类器判断该区域是否是一个实体. 然而, 这种方法会判断大量的非实体区域, 带来较高的计算成本. 基于边界感知的方法, 这种方法综合考虑了上述两种方法的优劣之处, 用序列标注的方法提取到命名实体的位置, 用基于区域的方法确定实体所属的类型, 图 10 以“甲状腺癌”为例子概述了这种方法的模型, 首先提取实体边界, 将“B”和“E”配对并进行实体区域标记, 然后对标记区域进行实体类别判断.

非连续实体. 医疗文本序列中有很多非连续实体的表达, 多个间隔的部分构成非连续实体, 主要包括以下处理方法: 数据标注层面[56]. 扩展传统的 BIO 标记, 添加必要的标签以满足非连续实体的标注需求. 如: BH 代表非连续实体中首个部分的开始, BI 代表非连续实体中首个部分的内部, BD 代表中间部分的开始, BI 代表中间部分的内部. 针对语料句子层面[57], 判断句子由哪些非连续命名实体的部分构成. 使用超图的方式表达语句中所有非连续部分的组合, 通过解码得到所识别到的实体; 基于转移的方法[58]. 预先设定好动作, 采用堆栈的技术对非连续命名实体的组成部分进行处理.

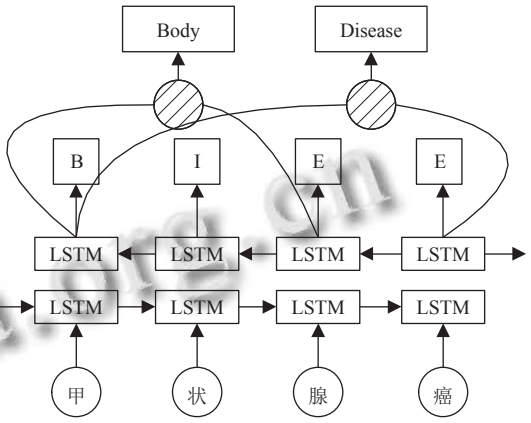


图 10 基于边界感知的方法

5 评价指标

正确识别出一个医疗命名实体, 既要正确识别出该实体的边界, 也要正确识别出其对应的类别. 医疗命名实体识别通常采用精确率 (Precision)、召回率 (Recall) 和 $F1$ 值 ($F1$-Measure) 对模型进行评估. 可通过如表 2 所示的混淆矩阵来理解, T_p代表模型预测为命名实体且预测正确的个数, F_p代表模型将非实体识别为命名实体的个数, F_n代表模型将命名实体识别为非实体的个数, T_n代表模型预测为非实体且预测正确的个数.

表 2 混淆矩阵

预测结果	实际结果	
	命名实体	非实体
命名实体	T_p	F_p
非实体	F_n	T_n

精确率 P, 指所有被模型识别为命名实体的样本中实际为命名实体的概率, 表达式为:

$$P = \frac{T_p}{T_p + F_p} \times 100\% \tag{5}$$

召回率 R, 指模型所有预测正确的结果中命名实体所占的比例, 表达式为:

$$R = \frac{T_p}{T_p + F_n} \times 100\% \tag{6}$$

当样本分布不均衡时, 仅考虑精确率或者召回率是不全面的, 由式 (5) 和式 (6) 可知它们是相互矛盾的. *F*1 值是命名实体识别的主要指标, 综合了上述两个评价指标, 表达式为:

$$F = \frac{2PR}{P+R} \times 100\% \tag{7}$$

6 总结与展望

本文对医疗命名实体识别任务进行研究, 主要做了以下工作: (1) 分析了命名实体识别对医学研究的重要意义以及其特有的难点; (2) 综述了传统的方法并详细归纳了基于深度学习的模型; (3) 介绍了当下较为流行的医疗命名实体识别模型改进方法; (4) 总结了常用数据集和评价指标.

结合医疗名实体识别任务的研究现状和趋势, 对今后的研究工作提出以下几点建议: 一方面, 就医疗领域的数据匮乏现状, 可对如何采用小规模的数据训练模型这一问题更深入地研究; 另一方面, 探索更有效的命名实体识别模型, 比如将图神经网络[59]、迁移学习等技术与现有的命名实体识别模型融合. 最后, 应注重命名实体识别在实际应用中的泛化能力, “AI+医疗”是大数据时代的一个探索性研究热点, 医疗知识图谱揭示了医学实体之间的逻辑关联, 智能问答系统为大众提供了科普性知识, 临床决策系统的普及缓解了医务人员的工作压力, 这些下游应用都离不开结构化数据的支持, 命名实体识别模型应适应于不同的应用场景, 并且能够与其他技术模块高效结合.

参考文献

1 Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investigationes, 2007, 30(1): 3–26. [doi: 10.1075/li.30.1.03nad]

2 Canfield K, Bray B, Huff S. Representation and database design for clinical information. Proceedings of the Annual Symposium on Computer Application in Medical Care. SCAMC Inc., 1990. 350–353.

3 Sager N, Lyman M, Bucknall C, *et al*. Natural language processing and the representation of clinical data. Journal of the American Medical Informatics Association, 1994, 1(2): 142–160. [doi: 10.1136/jamia.1994.95236145]

4 Friedman C, Hripcsak G, Dumouchel W, *et al*. Natural language processing in an operational clinical information system. Natural Language Engineering, 1995, 1(1): 83–108. [doi: 10.1017/S1351324900000061]

5 Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Computers and Biomedical Research, 1993, 26(5): 467–481. [doi: 10.1006/cbmr.1993.1033]

6 李楠, 郑荣廷, 吉久明, 等. 基于启发式规则的中文化学物质命名识别研究. 现代图书情报技术, 2010, 5: 13–17.

7 Hearst MA, Dumais ST, Osuna E, *et al*. Support vector machines. IEEE Intelligent Systems and Their Applications, 1998, 13(4): 18–28. [doi: 10.1109/5254.708428]

8 Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics, 1991, 21(3): 660–674. [doi: 10.1109/21.97458]

9 Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 8th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 2002. 282–289.

10 Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. Annals of Mathematical Statistics, 1966, 37(6): 1554–1563. [doi: 10.1214/aoms/1177699147]

11 Li Y, Gorman SL, Elhadad N. Section classification in clinical notes using supervised hidden Markov model. Proceedings of the 1st ACM International Health Informatics Symposium. Arlington: ACM, 2010. 744–750. [doi: 10.1145/1882992.1883105]

12 Zhou GD, Zhang J, Su J, *et al*. Recognizing names in biomedical texts: A machine learning approach. Bioinformatics, 2004, 20(7): 1178–1190. [doi: 10.1093/bioinformatics/bth060]

13 Lee KJ, Hwang YS, Rim HC. Two-phase biomedical NE recognition based on SVMs. Association for Computational Linguistics, 2003, 13(6): 33–34. [doi: 10.3115/1118958.1118963]

14 Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications. Geneva: Association for Computational Linguistics, 2004. 104–107.

15 叶枫, 陈莺莺, 周根贵, 等. 电子病历中命名实体的智能识别. 中国生物医学工程学报, 2011, 30(2): 256–262. [doi: 10.3969/j.issn.0258-8021.2011.02.014]

16 燕杨, 文敦伟, 王云吉, 等. 基于层叠条件随机场的中文病

历命名实体识别. 吉林大学学报 (工学版), 2014, 44(6): 1843–1848. [doi: 10.13229/j.cnki.jdxbgxb201406047]
17 Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527–1554. [doi: 10.1162/neco.2006.18.7.1527]
18 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90. [doi: 10.1145/3065386]
19 Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. arXiv: 1506.00019, 2015.
20 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Processing of Advances in Neural Information Processing Systems. Long Beach: NIPS, 2017. 5998–6008.
21 曹依依, 周应华, 申发海, 等. 基于 CNN-CRF 的中文电子病历命名实体识别研究. 重庆邮电大学学报 (自然科学版), 2019, 31(6): 869–875. [doi: 10.3979/j.issn.1673-825X.2019.06.017]
22 Liu ZJ, Yang M, Wang XL, *et al*. Entity recognition from clinical texts via recurrent neural network. BMC Medical Informatics and Decision Making, 2017, 17(S2): 67. [doi: 10.1186/s12911-017-0468-7]
23 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. 计算机工程与应用, 2020, 56(5): 153–159. [doi: 10.3778/j.issn.1002-8331.1909-0211]
24 Collobert R, Weston J, Bottou L, *et al*. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011, 12: 2493–2537.
25 陶源, 彭艳兵. 基于门控 CNN-CRF 的中文命名实体识别. 电子设计工程, 2020, 28(4): 42–46, 51. [doi: 10.14022/j.issn1674-6236.2020.04.009]
26 Gui T, Ma RT, Zhang Q, *et al*. CNN-based Chinese NER with lexicon rethinking. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: IJCAI, 2019. 4982–4988. [doi: 10.24963/ijcai.2019/692]
27 Huang ZH, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv: 1508.01991, 2015.
28 Dong CH, Zhang JJ, Zong CQ, *et al*. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. Natural Language Understanding and Intelligent Applications. Kunming: Springer, 2016. 239–250. [doi: 10.1007/978-3-319-50496-4_20]
29 Li F, Zhang MS, Tian B, *et al*. Recognizing irregular entities in biomedical text via deep neural networks. Pattern Recognition Letters, 2018, 105: 105–113. [doi: 10.1016/j.patrec.2017.06.009]
30 李丽双, 郭元凯. 基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别. 中文信息学报, 2018, 32(1): 116–122. [doi: 10.3969/j.issn.1003-0077.2018.01.015]
31 Zhang WT, Jiang SH, Zhao S, *et al*. A BERT-BiLSTM-CRF model for Chinese electronic medical records named entity recognition. 12th International Conference on Intelligent Computation Technology and Automation. Xiangtan: IEEE, 2019. 166–169. [doi: 10.1109/ICICTA49267.2019.00043]
32 郭晓然, 罗平, 王维兰. 基于 Transformer 编码器的中文命名实体识别. 吉林大学学报 (工学版), 2021, 51(3): 989–995. [doi: 10.13229/j.cnki.jdxbgxb20200640]
33 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 3111–3119.
34 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv: 1301.3781, 2013.
35 Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014. 1532–1543. [doi: 10.3115/v1/D14-1162]
36 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional Transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]
37 Strubell E, Verga P, Belanger D, *et al*. Fast and accurate entity recognition with iterated dilated convolutions. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 2670–2680.
38 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]
39 Zhang Y, Yang J. Chinese NER using lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 1554–1564. [doi: 10.18653/v1/P18-1144]
40 张笑天. 基于 Lattice LSTM 的医学文本中文命名实体识别

研究与实现 [硕士学位论文]. 成都: 电子科技大学, 2019.
41 Sun ZJ, Li XY, Sun XF, *et al*. ChineseBERT: Chinese pretraining enhanced by glyph and pinyin information. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021. 2065–2075.
42 Yin MW, Mou CJ, Xiong KN, *et al*. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism. Journal of Biomedical Informatics, 2019, 98: 103289. [doi: 10.1016/j.jbi.2019.103289]
43 Mnih V, Heess N, Graves A, *et al*. Recurrent models of visual attention. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2204–2212.
44 Luo L, Yang ZH, Yang P, *et al*. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 2018, 34(8): 1381–1388. [doi: 10.1093/bioinformatics/btx761]
45 单义栋, 王衡军, 黄河, 等. 基于注意力机制的命名实体识别模型研究——以军事文本为例. 计算机科学, 2019, 46(S1): 111–114, 119.
46 刘晓俊, 辜丽川, 史先章. 基于 Bi-LSTM 和注意力机制的命名实体识别. 洛阳理工学院学报 (自然科学版), 2019, 29(1): 65–70, 77. [doi: 10.3969/j.issn.1674-5043.2019.01.014]
47 李明扬, 孔芳. 融入自注意力机制的社交媒体命名实体识别. 清华大学学报 (自然科学版), 2019, 59(6): 461–467. [doi: 10.16511/j.cnki.qhdxxb.2019.25.005]
48 严红, 陈兴蜀, 王文贤, 等. 基于深度神经网络的法语命名实体识别模型. 计算机应用, 2019, 39(5): 1288–1292. [doi: 10.11772/j.issn.1001-9081.2018102155]
49 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. 计算机应用, 2020, 40(S1): 98–102. [doi: 10.11772/j.issn.1001-9081.2019081371]
50 Yang ZL, Salakhutdinov R, Cohen WW. Transfer learning for sequence tagging with hierarchical recurrent networks. Proceedings of the 5th International Conference on Learning Representations. Toulon: ICLR, 2017. 1–10.
51 Wang X, Zhang Y, Ren X, *et al*. Cross-type biomedical named entity recognition with deep multi-task learning. Bioinformatics, 2019, 35(10): 1745–1752. [doi: 10.1093/bioinformatics/bty869]
52 Giorgi JM, Bader GD. Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics, 2018, 34(23): 4087–4094. [doi: 10.1093/bioinformatics/bty449]
53 Wang ZH, Qu YR, Chen LH, *et al*. Label-aware double transfer learning for cross-specialty medical named entity recognition. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018. 1–15. [doi: 10.18653/v1/N18-1001]
54 Ju MZ, Miwa M, Ananiadou S. A neural layered model for nested named entity recognition. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans: Association for Computational Linguistics, 2018. 1446–1459. [doi: 10.18653/v1/N18-1131]
55 Sohrab MG, Miwa M. Deep exhaustive model for nested named entity recognition. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 2843–2849. [doi: 10.18653/v1/D18-1309]
56 Tang BZ, Hu JL, Wang XL, *et al*. Recognizing continuous and discontinuous adverse drug reaction mentions from social media using LSTM-CRF. Wireless Communications and Mobile Computing, 2018, 2018: 2379208. [doi: 10.1155/2018/2379208]
57 Wang BL, Lu W. Neural segmental hypergraphs for overlapping mention recognition. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 204–214. [doi: 10.18653/v1/D18-1019]
58 Dai X, Karimi S, Hachey B, *et al*. An effective transition-based model for discontinuous NER. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020. 5860–5870. [doi: 10.18653/v1/2020.acl-main.520]
59 Gui T, Zou YC, Zhang Q, *et al*. A lexicon-based graph neural network for Chinese NER. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 1040–1050. [doi: 10.18653/v1/D19-1096]

(校对责编: 孙君艳)