深度学习在医药领域命名实体识别中的研究进展

陈瑶[1]，葛卫红[2,3]，廖俊[1*]

（1. 中国药科大学理学院，江苏 南京 211198；2. 南京大学医学院附属鼓楼医院药学部，江苏 南京 210008；3. 中国药科大学基础医学与临床药学学院，江苏 南京 211198）

[摘要] 医药领域中文本作为一种主要的信息载体，其非结构化特征导致很难利用计算机直接进行批量分析。自然语言处理技术是自然语言与计算机语言之间转换的一种工具，近几年随着深度学习的发展在文本处理领域中有了广泛的应用，而命名实体识别作为自然语言处理的一个分支，在知识库构建、信息抽取等任务中发挥着重要的作用。针对命名实体识别在医药文本中的应用，介绍了当前主流的命名实体识别研究方法及主要数据来源，突出深度学习在医药领域实体识别应用中的优势，为该领域相关研究提供参考。

[关键词] 医药文本；深度学习；命名实体识别

[中图分类号] R9-39 [文献标志码] A [文章编号] 1001-5094（2020）01-0028-07

Research Progress in the Application of Deep Learning in Medical Named Entity Recognition

CHEN Yao[1], GE Weihong[2,3], LIAO Jun[1]

(1.School of Science, China Pharmaceutical University, Nanjing 211198, China; 2. Department of Pharmacy, Nanjing Drum Tower Hospital, Nanjing 210008, China; 3. School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing 211198, China)

[Abstract] As a main carrier of information in medical area, texts can hardly be analyzed directly in bulk because of their unstructured formats. Natural language processing is a tool to convert the natural language into computer language, which has been widely applied with the development of deep learning in text processing. Named entity recognition, a subtask of natural language processing, plays an important role in knowledge base construction and information extraction. In regard to the application of named entity recognition in medical text analysis, this article introduces the mainstream methods and data sources to illustrate the advantages of deep learning in this area, so as to give more reference for researchers in the field.

[Key words] medical text; deep learning; named entity recognition

命名实体识别（named entity recognition，NER）旨在根据预定义的实体类别从非结构化文本中将目标实体定位并分类，识别出的实体可进一步运用于实体关系抽取等自然语言处理任务中，是信息抽取的一部分，所抽取出的信息可作为目标对象知识库构建的基础。命名实体识别最初被提出主要用于对文本中人名、地名、机构名等实体的提取[1]，近几年在医药领域的文本中也得到了广泛应用。

本文选取在命名实体识别研究中应用最为广泛的3种模型进行介绍：传统机器学习的条件随机场模型（conditional random field，CRF）、深度学习的长短期记忆模型（long short-term memory，LSTM）及将这2种方法结合的双向LSTM-CRF模型（bidirectional LSTM-CRF，BiLSTM-CRF）。总结了医药领域命名实体识别研究中常用的数据源及相关公开标注语料集，并综述了深度学习在医药领域不同语言以及不同实体种类中的命名实体识别应用现状。本文旨在通过对这些方法、数据源以及应用的总结，为当前医药领域命名实体识别研究提供新思路。

接受日期：2019-05-05

项目资助：双一流创新团队生物医药大数据与人工智能（No. CPU2018GY19）；江苏省食品药品监督管理局2017—2018年度科研项目（No. 20170308）

*通讯作者：廖俊，副教授；

研究方向：药学信息学，医学生物信息学，医药大数据与人工智能；

Tel：025-86185122；E-mail：liaojun@cpu.edu.cn

1 命名实体识别常用方法

命名实体识别常用的方法可分为基于词典的方法[2]、基于规则的方法[3]和基于机器学习的方法[4]。基于词典的命名实体识别准确率很大程度上依赖所构建的词典库的丰富度，对于诸如医药相关的专业领域的实体，词典的匮乏导致很多专业词汇因为词典未覆盖而出现不能识别的情况。基于规则的方法涉及大量人工制定规则的过程，在目标文本较为复杂或不规则时，预先设定的规则很难满足所有情况，对数据适应性差且对于大数据集的处理存在一定的困难。基于机器学习的命名实体识别是当前热门的研究方法，常用的机器学

习算法有隐马尔可夫模型（Hidden Markov Model，HMM）、CRF、支持向量机（support vector machine，SVM）等，其中 CRF 是应用最为广泛的一种。深度学习作为机器学习的一种，由 Mikolov 等 [5] 于 2010 年提出的循环神经网络（recurrent neural network，RNN）使得深度学习在命名实体识别等序列文本处理中有了很好的应用，其中 LSTM 因为很好地解决了 RNN 处理大数据集时存在的梯度消失和梯度爆炸问题 [6]，成为近几年的研究热点。本文选取了这当中热门的 CRF、LSTM 及复合模型 BiLSTM-CRF 分别进行介绍。

1.1 条件随机场模型

CRF 是最初由 Lafferty 等 [7] 提出的用于序列数据的标识与切分的概率分布模型，在诸多实体识别任务中都取得了较好的表现 [8-9]，可用于解决特征挑选、参数训练和解码问题。CRF 是一种判别式的概率无向图模型，在自然语言处理任务中是用于标注和划分序列数据的概率化模型，其概率分布函数 P（X|Y）表示在满足马尔可夫随机场时的线性链 CRF，X 表示观测序列，Y 表示标注序列。

命名实体识别任务中常用的序列标注模式有‘BIO’和‘BIOES’2 种，其中‘B’表示实体开端字/词，‘I’表示实体内字/词，‘O’表示实体外字/词，‘E’表示实体末端字/词，‘S’表示单字 / 词实体。以‘BIO’模式为例，在使用 CRF 进行命名实体识别任务时，句子中的每个字或词共同组成观测序列 X，例如，X={ 患者同服奥美拉唑和克拉霉素 }，则与之对应的标注序列 Y={O O O O B-Drug I-Drug I-Drug I-Drug O B-Drug I-Drug I-Drug I-Drug}(其中 Drug 表示实体“药品”)，该过程中包括 2 个关键任务：一是实体边界的识别，即‘BIO’的界定；二是对实体类别的识别，即该实体是‘Drug’或其他类别。

CRF++ 是一种常被用来实现命名实体识别的开源工具，训练集和测试集由多个标记（token）组成，每个标记由字/词及该字/词对应的特征与标注组成，特征可由用户自定义，例如词性、字类型等。CRF++ 中有 4 种可根据语言特征选择的模板，包括 basenp、chunking、JapaneseNE 和 seg，因中文与日文中词与词之间均无明显界限，在进行中文命名实体识别任务时常采用根据日语环境设定的 JapaneseNE 模板。

1.2 长短期记忆模型

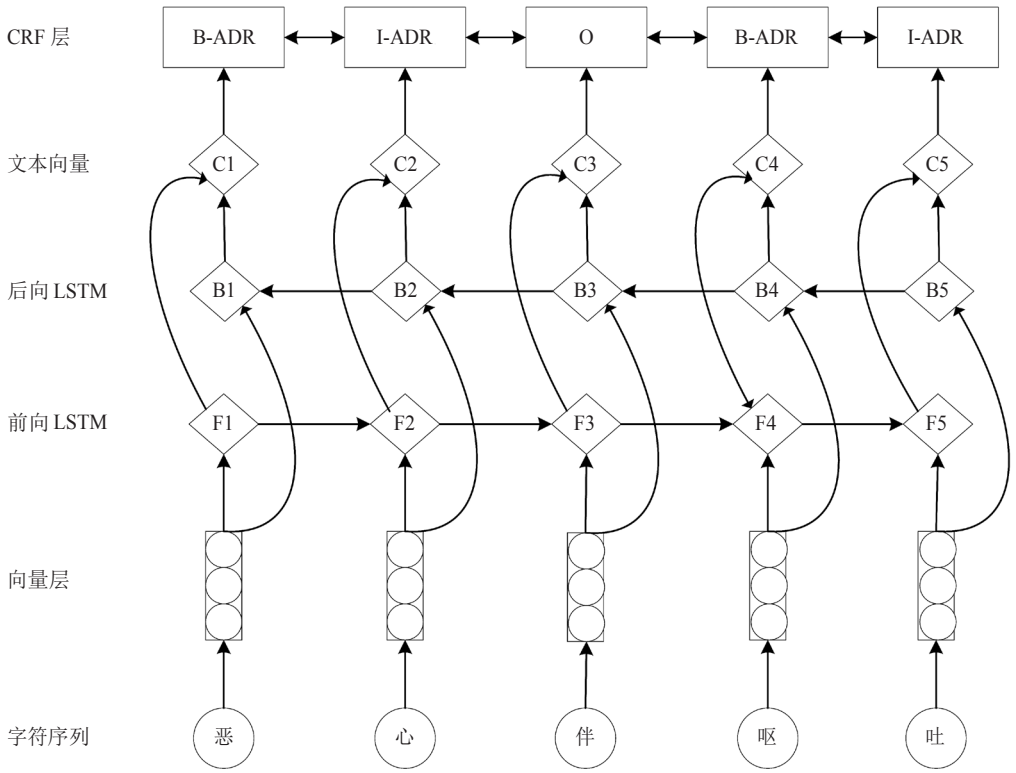
LSTM 是一种特殊的循环神经网络模型，其通过遗忘门、输入门和输出门这 3 个门来控制输入数据传入记忆单元的比例，实现对记忆单元中信息的交互，最终解决远距离依赖问题。双向长短期记忆模型（bidirectional LSTM，BiLSTM）的提出是为了充分利用序列标签问题中目标对象的前向及后向信息。

谷歌于 2013 年推出的开源工具 —word2vec [10] 将词表示为分布式词向量，在能够很好地表示词语特征的同时，也兼顾了后续网络模型维度的设置，使得深度学习在诸多自然语言处理中的应用有了进一步的突破。将其运用于 BiLSTM 模型进行的命名实体识别任务中，通过对字/词的分布式表达，省去了人工构建特征的过程，从而获得相较于 CRF 等传统机器学习算法更大的优势。例如，Cocos 等 [11] 和 Xie 等 [12] 在分别使用字典匹配、CRF 和 BiLSTM 模型从社交媒体中提取药品不良反应（adverse drug reaction，ADR）实体时，实验结果均表明 BiLSTM 的实体识别准确度更高。

1.3 双向 LSTM-CRF 模型

BiLSTM-CRF 模型可以理解为是将 CRF 引入到 BiLSTM 网络结构中的混合模型 [13]，是近期应用于命名实体识别任务中最为热门的深度学习模型之一 [14]。在进行实体识别任务时，传统 BiLSTM 模型的输出为挑选出来的得分最高的独立标签，这就导致前后标签无互相联系与制约，易造成实体之间的混淆，而 CRF 层的引入则可以加强这种制约。在 BiLSTM-CRF 模型中，BiLSTM 负责从训练语料中自动学习特征，将学习到的特征向量传入到 CRF 层并输出概率值最高的标签作为预测结果 [15]，在这当中可以添加词向量（word embedding）、字向量（character embedding）、位置向量（position embedding）等作为潜在特征，字/词向量可通过 word2vec 等工具对语料预训练获得。

以中文药品不良反应实体识别的研究为例，BiLSTM-CRF 模型框架如图 1 所示，图中的向量层（embedding layer）可以是词向量、字向量、位置向量等的组合，在经过前向、后向 LSTM 的处理后转化为上下文向量（vector），最后输入到 CRF 层进行概率值的计算与结果的输出。Unanue 等 [16] 分别将词向量、字向量和自定义特征融入 CRF、BiLSTM 和 BiLSTM-CRF 模型中，结果显示 BiLSTM-CRF 具有比单独的 CRF 及 BiLSTM 更高的 F 值，F 值为机器学习中常用评价指标，值越大表示模型效果越佳。



B-ADR：ADR 实体首字符；I-ADR：ADR 实体除首字符外的字符；O：实体外字符；B：后向；F：前向

图 1 BiLSTM-CRF 模型结构

Figure 1 Structure of BiLSTM-CRF Model

1.4 其他方法

以上介绍的 CRF、LSTM、BiLSTM-CRF 模型都属于有监督的机器学习，其特征是需要标注好的数据作为模型的训练语料，在标注语料不足的情况下则很难获得较好的模型准确率。半监督学习则可以在少量已有标注语料的基础上实现训练集的扩充，从而进一步提高模型准确率，常见的半监督学习方法有自训练（self-training）[17]、协同训练（co-training）[18] 以及三体训练（tri-training）[19] 等。此外，还有一些基于无监督的命名实体识别应用 [20-21]，但目前在医药领域的应用尚未成熟。除了使用单一模型进行训练外，还有学者采用多种方法集成的模式，例如，Wei 等 [22] 通过 SVM 将 CRF 的结果与双向 RNN 的结果相融合，最终取得的 F 值比单独使用 CRF 或 Bi-RNN 模型 F 值都要高。

2 医药领域命名实体识别任务常用数据来源介绍

医药领域命名实体识别常用的数据来源包括医学文献、电子病历、社交媒体等，其中 Gurulingappa 等 [23] 及 Mulligen 等 [24] 选取生物医学数据库 Medline 中的文献摘要作为源数据进行语料标注，所构建的标注语料集中涵盖了药品、不良反应、剂量、基因等实体；电子病历中涉及的医学相关实体更为广泛，可以从中进行患病、治疗、用药、检查等相关实体的识别研究 [25-26]；社交媒体因其用户覆盖面广、数据具有即时性等特征，成为近几年热门的命名实体识别数据来源，尤其是在药品不良反应相关的实体识别应用中 [11-12,27]。在 3 种常用数据来源中，电子病历中的医药相关实体密度最高，在利用该类数据进行实体识别研究时，需要对各类实体有明确的界定，防止因定义模糊或概念交叉而产生误差，此外对于病人一些隐私信息也要作相关处理；社交媒体中医药相关文本易获取且数目庞大，但其中的信息密度也最为稀疏，在进行医药实体识别研究时需要对其中大量的嘈杂信息进行剔除；医学文献中的医药实体信息密度介于两者之间，且当前有很多相关的公开标注语料集，是应用较为成熟的一种医药资源。图 2 展示了基于不同数据来源的医药领域命名实体识别在爱思唯尔 ScienceDirect 数据库中的研究数量分布，其中选择医学文献作为数据来源的研究数目最多，占比达 61.74%。

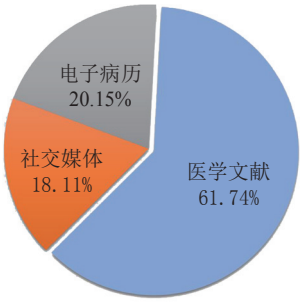


图 2 基于不同数据源的命名实体识别研究分布

Figure 2 The distribution of NER researches based on different data sources

无论是传统机器学习还是深度学习，都需要标注语料对模型进行训练，在缺乏标注语料的情况下，很多研究者倾向使用公开标注语料集展开研究，表 1 中总结了一些常用的医药领域命名实体识别开放数据集及其简介与链接。

命名实体识别技术在医药领域中的应用，主要的目标实体包括药物 / 化学物质、蛋白质 / 基因、疾病、药品不良反应等，表 2 从这些实体的角度出发，整理了近几年一些深度学习在不同类型的实体识别中的应用以及相应的数据来源。

表 1 常用医药相关命名实体识别开放数据集

Table 1 Frequently used open datasets related to medical NER

数据集	简介	链接
i2b2 challenge	临床记录标注文本，子任务包括实体识别、断言分类、关系分类等，可用于患病、治疗、医学检查等实体的识别研究	https://i2b2.org/NLP/DataSets
CHEMDNER	BioCreative challenge挑战赛的子任务，可用于化学物质和药品名称的实体识别研究	https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-2-chemdner/
CDR	BioCreative challenge挑战赛的子任务，可用于化学物质与疾病的实体识别研究以及化学物质-疾病关系提取研究	https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/
GPRO	BioCreative challenge挑战赛的子任务，可用于基因和蛋白质相关实体识别的研究	http://www.biocreative.org/
DDIExtraction2013 task	可用于药物实体识别、药物相互作用提取研究	https://www.cs.york.ac.uk/semeval-2013/task9/
JNLPBA	PubMed摘要的标注语料集，可用于基因、蛋白质、细胞等实体的识别研究	http://iasl-btm.iis.sinica.edu.tw/BNER/Home/Download
NCBI Disease	PubMed摘要的标注语料集，可用于疾病的实体识别研究	http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/
MADE 1.0	基于电子病历的标注语料集，可进行药物治疗、药品不良反应等实体识别研究	http://bio-nlp.org/index.php/projects/39-nlp-challenges
PHAEDRA corpus	基于Medline摘要文本的标注语料集，可进行药理学相关物质（基因产物、药物、代谢物等）、临床异常表现（症状、病理过程等）以及微生物、细胞、病毒等对象的实体识别研究	http://www.nactem.ac.uk/PHAEDRA/

表 2 深度学习在不同类别实体识别中的应用

Table 2 Application of deep learning in NER of different types

实体类别	数据来源	参考文献
药物/化学物质	中文电子病历	[28]
	CHEMDNER	[29]
	CHEMDNER、CDR	[30]
蛋白质/基因	GPRO	[31]
	JNLPBA	[32]
疾病	NCBI disease	[33]
	在线医疗诊断数据	[34]
药品不良反应	MADE 1.0	[35]
	社交媒体	[36]

3 深度学习在中英文医药领域命名实体识别中的应用

3.1 英文医药领域文本

命名实体识别任务最初是在英文环境下提出，无论是规则的制定还是特征的提取都有很多经验可以借鉴，加上大多数医药领域公开标注语料库及可供验证的医药类数据库都是英文，使得各类方法在英文医药领域命名实体识别中的应用都相对较成熟。

BiLSTM-CRF 因其兼具 BiLSTM 和 CRF 的优势，自提出以来就一直深受欢迎，例如，Zeng 等 [15]、Luo 等 [37]、Gridach 等 [32] 分别使用 BiLSTM-CRF 方法从医药文本中进行了药品实体、化学物质实体及基因实体的识别研究且都取得了较好的识别效果。在使用深度学习进行实体识别研究时，很多研究者选择通过 word2vec 等工具对无标注文本进行预训练，通过自动学习到的文本特征实现模型性能的提高 [38-39]；注意力机制（attention）也是经常用于深度学习神经网络结构中的一个重要因素，例如，杨培等 [40] 在对 CHEMDNER 数据集中的化学物质进行实体识别研究

时，在 BiLSTM 层和 CRF 层之间加入 attention 层，将由 BiLSTM 获得的词的上下文表示转化为该词在全文范围内的上下文表示，联合该词的邻近上下文表示一同传入 CRF 层以获得标签序列；对于医药领域的命名实体识别，领域知识或专有词典的引入也能在一定程度上提高模型识别效果，可将词典转化为词特征与当前文本训练得到的词向量一起作为神经网络的输入，从而将专有词汇信息传输到模型进行学习 [41]。

3.2 中文医药领域文本

不同于英文命名实体识别研究，中文在这一领域的技术还不够成熟且面临诸多挑战。一方面，中文词与词之间无类似英文中的空格作为明显分割界限，因此中文自然语言处理任务中首先需要解决分词的问题，而特殊领域中的专业词汇、缩写等使得难以完全依靠分词工具实现准确的词切分 [42]；另一方面，中文开放性标注语料以及可用于验证的医药相关词典或数据库的缺乏也进一步限制了中文命名实体识别的研究 [43]。

就深度学习在中文命名实体识别中的应用而言，很多研究者选择了包含人名、地名和机构名 3 种实体的开放标注中文语料作为研究对象 [44-45]，而医药领域因目前尚缺乏公开标注语料集，针对医药领域的中文命名实体识别应用也因此相对较少。2015 年 Wu 等 [46] 首次尝试将深层神经网络（deep neural network，DNN）运用于中文临床文本的实体识别研究中，并将 DNN 的识别效果与 CRF 进行对比，结果显示将未标注的数据加入神经网络训练出的模型得到了较高的 F 值；夏宇彬等 [47] 选取 200 份糖尿病患者电子病历中的入院记录进行标注，并分别使用了多层感知机、CRF、LSTM 模型进行实体识别研究，结果表明 LSTM 的识别效果最好；张艺品等 [48] 和高甦等 [49] 则分别采用 BiLSTM-CRF 模型对中医典籍进行了命名实体识别研究，前者目标实体为病症、方剂和中药材，后者目标实体则为中医认识方法、中医生理、中医病理、中医自然及治则治法等，两者均取得了较理想的研究结果。

4 结语

命名实体识别作为自然语言处理任务的分支，将其运用于自由文本中可以实现大量非结构化信息的结构化转换，对于目标对象的信息抽取及知识库构建都具有重要意义。医药领域的命名实体识别，除了基本的语言学特征外，还需要考虑各类实体的定义与概念区分、专有词汇的补充、不同数据源的融合等问题，而传统的人工构建规则或特征的方法很难满足需求。CRF 对序列问题的处理使得其在命名实体识别任务中展现出相较其他传统机器学习的优势，但诸如 BiLSTM 等深度学习模型通过文本的分布式表达自动获取字、词以及句子层面特征的能力省去了传统机器学习的人工特征构建工程，同时在处理大数据集时也更具优势，是近年来命名实体识别以及自然语言处理其他领域的研究热点，在医药领域也有了很好的利用。

中文的语言学特征以及医药领域公开标注语料集的缺乏，使得深度学习在中文医药领域命名实体识别中的应用还不够广泛，大多数是基于个人标注的语料集进行，规模小且泛化能力弱。医药领域的命名实体识别研究，一方面需要进一步结合已有在其他领域成熟应用的深度学习方法或者探索更多不同方法应用，另一方面也要加强基础语料库的建设以及特定方向相关标准的规范与统一。

[参考文献]

[1] Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. *Lingvis Invest*, 2007, 30(1): 3-26.

[2] Ekbal A, Saha S. Simultaneous feature and parameter selection using multiobjective optimization: application to named entity recognition[J]. *Intern J Mach Lear & Cyber*, 2016, 7(4): 597-611.

[3] Mai O, Khaled S. NERA 2.0: improving coverage and performance of rule-based named entity recognition for Arabic[J]. *Nat Lang Engi*, 2016, 23(3): 441-472.

[4] Kanimozhi U, Manjula D. A CRF based machine learning approach for biomedical named entity recognition[C/OL]. 2017, 335-342[2019-05-05]. https://www.computer.org/csdl/proceedings-article/icrtccm/2017/4799a335/12OmNxX3uNo. Doi: 10.1109/ICRTCCM.2017.23.

[5] Mikolov T, Karafiat M, Burget L, *et al*. Recurrent neural network based language model[C/OL]. 2010, 1045-1048[2019-05-05]. https://www.researchgate.net/publication/221489926_Recurrent_neural_network_based_language_model/link/0c960523991065d41b000000/download.

[6] Graves A. *Supervised sequence labelling with recurrent neural networks*[M/OL]. Heidelberg: Springer-Verlag Berlin Heidelberg, 2012: 385[2019-05-05]. https://www.springer.com/cn/book/9783642247965. Doi:10.1007/978-3-642-24797-2.

[7] Lafferty J, Mccallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C/OL]. 2001, 282-289[2019-05-05]. https://www.researchgate.net/publication/2529190_Conditional_Random_Fields_Probabilistic_Models_for_Segmenting_and_Labeling_Sequence_Data.

[8] Han A L, Wong D F, Chao L S. *Chinese named entity recognition with conditional random fields in the light of chinese characteristics*[M/OL]. Heidelberg: Springer Berlin Heidelberg, 2013:57-68[2019-05-05]. https://link.springer.com/chapter/10.1007%2F978-3-642-38634-3_8. Doi: https://doi.org/10.1007/978-3-642-38634-3_8.

[9] Ekbal A, Bandyopadhyay S. A conditional random field approach for named entity recognition in bengali and hindi[J]. *Linguist Issues in Lang Technol*, 2009, 2(1): 1-44.

[10] Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality[EB/OL]. (2013-10-16)[2019-05-05]. https://arxiv.org/abs/1310.4546.

[11] Cocos A, Fiks A G, Masino A J. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts[J]. *J Am Med Inform Assoc*, 2017, 24(4): 813-821.

[12] Xie J H, Liu X, Zeng D D. Mining e-cigarette adverse events in social media using Bi-LSTM recurrent neural network with word embedding representation[J]. *J Am Med Inform Assoc*, 2018, 25(1): 72-80.

[13] Huang Z H, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. (2015-8-9)[2019-05-05]. https://arxiv.org/abs/1508.01991v1.

[14] Lample G, Ballesteros M, Subramanian S, *et al*. Neural architectures for named entity recognition[EB/OL]. (2016-4-7)[2019-05-05]. http://arxiv.org/abs/1603.01360v3.

[15] Zeng D, Sun C, Lin L, *et al*. LSTM-CRF for drug-named entity recognition[J/OL]. *Entropy*, 2017, 19(6): 283[2019-05-05]. https://www.mdpi.com/1099-4300/19/6/283. Doi: 10.3390/e19060283.

[16] Unanue I J, Borzeshi E Z, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition[J]. *J Biomed Inform*, 2017, 76: 102-109[2019-05-05]. https://www.sciencedirect.com/science/article/pii/S1532046417302447. Doi:10.1016/j.jbi.2017.11.007

[17] Riloff E, Wiebe J, Wilson T. Learning subjective nouns using extraction pattern bootstrapping[C/OL]. 2003, 25-32[2019-05-05]. https://www.researchgate.net/publication/2901394_Learning_Subjective_Nouns_Using_Extraction_Pattern_Bootstrapping. Doi: 10.3115/1119176.1119180.

[18] Blum A, Mitchell T M. Combining labeled and unlabeled data with co-training[C/OL]. 1998, 92-100[2019-05-05]. https://dl.acm.org/doi/10.1145/279943.279962.

[19] Zhou Z H, Li M. Tri-training: exploiting unlabeled data using three classifiers[J]. *IEEE Trans Know Dat Engi*, 2005, 17(11): 1529-1541.

[20] Ren X, EI-kishky A, Wang C, *et al*. ClusType: effective entity recognition and typing by relation phrase-based clustering[C/OL]. 2015, 995-1004[2019-05-05]. https://dl.acm.org/doi/10.1145/2783258.2783362.

[21] Brooke J, Hammond A, Baldwin T. Bootstrapped text-level named entity recognition for literature[C/OL]. 2016, 344-350[2019-05-05]. https://www.researchgate.net/publication/306093485_Bootstrapped_Text-level_Named_Entity_Recognition_for_Literature. Doi: 10.18653/v1/P16-2056.

[22] Wei Q K, Chen T, Xu R F, *et al*. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks[J/OL]. *Database(Oxford)*, 2016, 2016: baw140[2019-05-05]. https://pubmed.ncbi.nlm.nih.gov/27777244/ . Doi: 10.1093/database/baw140.

[23] Gurulingappa H, Rajput A M, Roberts A, *et al*. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports[J]. *J Biomed Inform*, 2012, 45(5): 885-892.

[24] Mulligen E M V, Fourrier-Reglat A, Gurwitz D, *et al*. The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships [J]. *J Biomed Inform*, 2012, 45(5): 879-884.

[25] Lei J B, Tang B Z, Lu X Q, *et al*. A comprehensive study of named entity recognition in chinese clinical text[J]. *J Am Med Inform Assoc*, 2014, 21(5): 808-814.

[26] Wu Y H, Jiang M, Lei J B, *et al*. Named entity recognition in chinese clinical text using deep neural network[J/OL]. *Stud Health Technol Inform*, 2015, 216: 624-628[2019-05-05]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4624324/. Doi:10.3233/978-1-61499-564-7-624.

[27] Gupta S, Pawar S, Ramrakhiyani N, *et al*. Semi-supervised recurrent neural network for adverse drug reaction mention extraction[EB/OL]. (2017-09-06)[2019-05-05]. https://arxiv.org/abs/1709.01687?context=cs.

[28] Ji B, Liu R, Li S S, *et al*. A hybrid approach for named entity recognition in chinese electronic medical record[J]. *Bmc Med Inform Deci Mak*, 2019, 19(Suppl 2): 149-158.

[29] Hemati W, Mehler A. LSTMVoter: chemical named entity recognition

using a conglomerate of sequence labeling tools[J/OL]. *J Cheminform*, 2019, 11: 7[2019-05-05]. https://link.springer.com/article/10.1186/s13321-018-0327-2.

[30] Luo L, Yang Z H, Yang P, *et al*. An attention-based bilstm-crf approach to document-level chemical named entity recognition[J]. *Bioinform*, 2018, 34(8): 1381-1388.

[31] Luo L, Yang Z H, Yang P, *et al*. A neural network approach to chemical and gene/protein entity recognition in patents[J/OL]. *J Cheminform*, 2018, 10(1): 65[2019-05-05]. https://link.springer.com/article/10.1186/s13321-018-0318-3.

[32] Gridach M. Character-level neural network for biomedical named entity recognition[J/OL]. *J Biomed Inform*, 2017, 70: 85-91[2019-05-05]. https://www.sciencedirect.com/science/article/pii/S1532046417300977. Doi: 10.1016/j.jbi.2017.05.002.

[33] Xu K, Yang Z, Kang P, *et al*. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition[J/OL]. *Comput bio med*, 2019, 108: 122-132[2019-05-05]. https://linkinghub.elsevier.com/retrieve/pii/S0010-4825(19)30110-6. Doi: 10.1016/j.compbiomed.2019.04.002.

[34] Liu X, Zhou Y J, Wang Z R. Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network[J/OL]. *J Vis Commu Im Rep*, 2019, 60: 1-15[2019-05-05]. https://www.onacademic.com/detail/journal_1000041590816099_2d60.html. Doi: 10.1016/j.jvcir.2019.02.001.

[35] Wunnava S, Qin X, Kakar T, *et al*. Adverse drug event detection from electronic health records using hierarchical recurrent neural networks with dual-level embedding[J]. *Drug Safety*, 2019, 42(1): 113-122.

[36] Dai H J, Touray M, Jonnagaddala J, *et al*. Feature engineering for recognizing adverse drug reactions from twitter posts[J/OL]. *Information (Switzerland)*, 2016, 7(2): 27[2019-05-05]. https://tmu.pure.elsevier.com/en/publications/feature-engineering-for-recognizing-adverse-drug-reactions-from-t. Doi: 10.3390/info7020027.

[37] Luo L, Yang Z H, Yang P, *et al*. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J]. *Bioinformatics*, 2018, 34(8) :1381-1388.

[38] Mikolov T, Chen K, Corrado G S, *et al*. Efficient estimation of word representations in vector space[EB/OL]. (2013-9-7)[2019-05-05]. http://arxiv.org/abs/1301.3781v3.

[39] Ling W, Luís T, Marujo L, *et al*. Finding function in form: compositional character models for open vocabulary word representation[EB/OL]. (2016-05-23)[2019-05-05]. https://arxiv.org/abs/1508.02096.

[40] 杨培，杨志豪，罗凌，等. 基于注意机制的化学药物命名实体识别 [J]. 计算机研究与发展，2018, 55(7): 1548-1556.

[41] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[EB/OL]. (2016-7-19)[2019-05-05]. https://arxiv.org/abs/1511.08308.

[42] Gao J F, Li M, Huang C N, *et al*. Chinese word segmentation and named entity recognition: a pragmatic approach[J]. *Comput Ling*, 2005, 31(4): 531-574.

[43] 武文雅，陈钰枫，徐金安，等. 中文实体关系抽取研究综述 [J]. 计算机与现代化，2018, (8): 21-27.

[44] 冯艳红，于红，孙庚，等. 基于 BLSTM 的命名实体识别方法 [J]. 计算机科学，2018, 45(2): 261-268.

[45] 张海楠，伍大勇，刘悦，等. 基于深度神经网络的中文命名实体识别 [J]. 中文信息学报，2017, 31(4): 28-35.

[46] Wu Y H, Jiang M, Lei J B, *et al*. Named entity recognition in chinese clinical text using deep neural network[J/OL]. *Stud Health Technol Inform*, 2015, 216: 624-628[2019-05-05]. https://www.researchgate.net/publication/294444331_Named_Entity_Recognition_in_Chinese_Clinical_Text. Doi: 10.3233/978-1-61499-564-7-624.

[47] 夏宇彬，郑建立，赵逸凡，等. 基于深度学习的电子病历命名实体识别 [J/OL]. 电子科技，2018, 31(11): 31-34, 37.

[48] 张艺品，关贝，吕荫润，等. 深度学习基础上的中医实体抽取方法研究 [J]. 医学信息学杂志，2019, 40(2): 58-63.

[49] 高甦，金佩，张德政. 基于深度学习的中医典籍命名实体识别研究 [J]. 情报工程，2019, 5(1): 113-123.

[专家介绍] 廖俊，博士，中国药科大学理学院副教授，硕士生导师，美国密西根大学药学院访问学者。现任中国药科大学高性能计算中心主任，主要从事药学信息学、人工智能辅助病理诊断研究。目前致力于将深度学习运用于药品不良反应、中药物质基础及作用机制、数字病理诊断研究，积累了较为丰富的人工智能及医药大数据研究经验。