• 医学信息研究 •

生物医学命名实体识别研究现状及中文生物医学命名实体识别难点与意义综述*

潘璀然 施 维 薛 均 王青华 王 理 董建成

（南通大学医学院医学信息学系 南通 226001）

〔摘要〕 介绍国内外生物医学命名实体识别的研究现状，详细阐述生物医学命名实体识别的技术方法，包括基于词典和规则的方法、基于机器学习的方法、混合方法和神经网络方法以及相关测评组织和标准，总结中文生物医学命名实体识别难点和意义。

〔关键词〕 中文；生物医学；命名实体识别

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2018.03.012

Study Status of Biomedical Named Entity Recognition as Well as Difficulties and Significance of Chinese Biomedical Named Entity Recognition

PAN Cui－ran, SHI Wei, XUE Jun, WANG Qing－hua, WANG Li, DONG Jian－cheng, School of Medicine of Nantong University, Nantong 226001, China

〔**Abstract**〕 The paper introduces the study situation of biomedical named entity recognites in and outside of China, elaborating on the technical method of biomedical named entity recognition, including method based on dictionaries and rules, method based on machine learning, mixed method, neural network method and related evaluation organizations and criterion, and summarizes the difficulties and significance of Chinese biomedical named entity recognition.

〔**Keywords**〕 Chinese; Biomedicine; Named entity recognition

1 引言

当前生物医学发展迅速，大量的生物医学知识主要以非结构化的形式存在于各种形式的文本中。这些文本文件中包含丰富的生物医学知识，可为科研和教学提供大量专业的数据和知识。据统计，Medline 数据库中文献平均以每年 60 万篇的速度在增长，国际数据公司的一项市场调查显示，在 2009－2020 年间数字信息将会以 44 倍的速度增长，但是人员维护和投资却只以 1.4 倍的速度增长，远远超过数据库更新速度，应对巨大的比率差异是一个很大的挑战，因此需要引入一些智能化的方式来自动获取信息[1]。

生物医学非结构化知识转换的任务之一就是生物医学文本中命名实体的识别。其主要任务是从生物医学文本中找到并提取出基因（蛋白质）、疾病、药物等特定类型的名称。近几年生物医学命名实体

〔收稿日期〕 2017－11－10

〔作者简介〕 潘璀然，硕士研究生；通讯作者：王理，副教授。

〔基金项目〕 江苏省研究生科研与实践创新计划项目（项目编号：KYCX17－1932）；国家自然科学基金资助项目（项目编号：81501559、81701793）。

识别领域的研究发展迅速，以“命名实体识别”和“Named Entity Recognition”为关键词搜索文献可发现，在英文领域对生物医学命名实体的研究偏多，研究中文生物医学命名实体识别的文献相对偏少，且大多研究中文人名、地名或组织名。本研究首先回顾生物医学命名实体识别的发展历史，详细介绍命名实体识别涉及的主要研究方法以及相关测评标准和组织，然后总结中文命名实体识别的难点和意义，最后对未来发展趋势进行展望。

2 国内外研究现状

2.1 国外

国外对命名实体识别研究进行较早。1996 年命名实体在 MUC－6 会议上被首次提出，用来指代具有唯一标识符的实体。在 1996－2008 年间对 NER 工具的评估在 MUC、CoNLL（Computational Natural Language Learning）和 ACE（Automatic Content Extraction）会议上都作为任务之一在进行。早期的命名实体识别工作主要集中在 3 类名词：人名、地名和组织名。Fleischman 和 Hovy 在人名再分类的问题上提出一种监督学习方法，考虑实体周围的上下文环境以及从 WordNet 和 Topic Signatures 获取的语义信息，融入一种强化算法[2]。此外，对特定领域的命名实体识别日益受到研究者的关注。在生物医学领域，对基因、蛋白质疾病名等的识别已取得了不错的效果。Murugesan 等[3]提出 BCC－NER（bidirectional, contextual clues named entity tagger for gene/protein mention recognition）方法，分为文本预处理（特征提取），模型训练（CRF，使用 MIRA 算法对前向和后向模型整合）和后期处理 3 个模块来对 BioCreative II GM 语料库训练和测试，获得较好的准确率。

2.2 国内

在中文信息处理领域，国家 863 计划智能计算机专家组从 1995 年起，组织中文信息处理与智能人机接口测评，随后国内对命名实体识别的研究也逐渐重视起来。在已有研究成果中，与国外情况相似，很多是针对人名、地名或者组织名的研究成果。闫萍[4]采用统计与规则结合的方法，通过对姓氏在真实文本中作为真实姓名的概率进行统计分析等工作，实现对人名的自动识别。近几年对生物医学命名实体识别的研究已逐渐成为该领域不断探索的焦点。目前国内对生物医学的研究主要集中在基因，蛋白质和疾病名等的实体识别上。Tang 等[5]调查了基于机器学习的 BNER 系统的 3 种不同类型的词表示特征，即基于聚类的表示、分布表示和词向量表示，通过对 BioCreAtIvE II GM 和 JNLPBA 语料库的评估表明，3 种特征都有利于 BNER 系统效果的提高。何林娜等[6]旨在从生物医学文献中提取药物名称，提出基于 FCG 的半监督方法结合 CRF 自动识别文献中的药物名称。由于国内缺少公开的中文生物医学测评语料，因此目前国内学者对生物医学的研究大多基于英文语料。

3 研究方法

3.1 基于词典和规则的方法

最早使用的方法是基于词典的方法，该方法主要是运用已有的标准术语词典和匹配算法来识别文本中出现的术语。在生物医学领域，最著名的术语词典为 ICD－10、UMLS、RxNorm 和 SNOMED CT 等，因此早期的生物医学命名实体识别大多采用词典匹配的方法，形成医疗领域 3 个代表性的通用工具：MedLEE、MedKAT 和 cTAKES[7]。但该方法的识别效果很大程度依赖于词典质量和匹配算法，而生物医学领域专业术语众多且新的命名实体不断出现，词典质量的更新面临挑战，因此单纯依赖传统的词典匹配方法效果难以提高，通常与其他方法结合使用。夏光辉[8]根据 UMLS 转换构建基因实体词典，结合机器学习方法条件随机场对 GENIA3.02 语料进行基因命名实体识别，在获得较高识别率的同时，还能降低时间复杂度。基于规则的方法多需要手动或启发式的制定规则模板来识别文本中的命名实体。程志刚[9]运用开源软件 GATE 框架定义规则，自动识别出语料中较为规范的时间词和数词等实体，然后采用条件随机场针对不同的语料分别设计模板并择优筛选，识别人名、地名和组织名，识

别效果良好。基于规则的方法取得较好的性能，但该方法缺乏可移植性和鲁棒性，针对新的领域需要建立新的规则，需要大量的领域专家，语言学家和时间成本。基于复杂规则的系统通常精确率高，但规则也越特殊，召回率越低[10]。

3.2 基于机器学习的方法

3.2.1 概述 机器学习的方法通常被做为序列标注问题来研究，序列标注是指对序列中的每个符号赋予一个特定的标签，输入是一些词序列，输出是实体加预测结果。机器学习方法主要解决两个问题，分别是实体边界的确立，以及实体类型的预测标注。如门诊行 CT 检查，显示脑萎缩。其中 CT 是检查实体，脑萎缩是症状实体。对每个实体给出特定的标签来表明实体的开始中间和结束等词位信息。机器学习方法大致分为 3 类：有监督学习方法、半监督学习方法和无监督学习方法。

3.2.2 有监督学习方法 即根据已标注语料库来训练模型，从而得到 1 个最优模型，再利用该模型将输入映射为相应的输出，对输出进行简单的判断从而实现分类的目的。有监督学习将命名实体识别作为序列标注问题。常用的序列标注模型有：隐马尔科夫模型（Hidden Markov Models，HMM），最大熵（Maxi mum Entropy，ME）和条件随机场（Conditional Random Fields，CRF）等。隐马尔科夫模型利用上下文信息和 Viterbi 算法求解命名实体类别序列，在训练和识别时速度较快。但 HMM 条件是假设可观测变量之间相互独立，这在实际应用时并不现实。魏尊强、舒红平和王亚强[11]通过分析中医临床记录的特点，对序列标注方法进行改进，对中医症状名称识别，通过实验对比，改进后 HMM 算法在性能评估上优于未改进算法。最大熵模型利用信息熵的定义，从符合约束条件的模型中择优选择使信息熵达到最大的模型，还解决模型中的参数平滑问题。但该模型算法收敛速度慢，训练时间代价较大，数据稀疏问题严重，且引起标记偏置问题。条件随机场是由 Lafferty 在 2001 年提出的一种判别式模型，是主要应用于标注和切分有序数据的条件概率模型。CRF 克服 HMM 的独立性假设条件和 ME 的标记偏置问题，可以充分利用上下文信息。同时 CRF 也具有训练代价大，复杂度高等缺点，但基于 CRF 的众多优点，其成为当下命名实体识别研究的主流技术方法，常用于分词、词性标注、命名实体识别等任务。燕杨、文敦伟等[12]提出一种层叠条件随机场方法用于实体识别，在第 1 层中实现对基本疾病名和身体部位的识别，将识别结果传递到第 2 层条件随机场，该方法相比无自定义组合特征的层叠条件随机场模型，F 值提高 3%，相比单层条件随机场模型，F 值提高 7%。江林刚[13]针对条件随机场在训练大规模数据时传统单机性能低下的问题，提出基于第 2 代 Hadoop 平台的条件随机场模型训练并行优化算法：CRFs - L - MapReduce 和 CRFs - V - Spark，均能有效提高识别率。

3.2.3 半监督学习方法 随着互联网的快速发展，收集大量标注文本相对困难，需耗费的人力和物力成本过大，因此如何使用未标记或少量的标记数据来提高 NLP 效果成为目前该领域的热点之一。半监督学习即给定少量的标注集作为种子用于学习，系统会自动学习实体的上下文环境，不断循环发现新的上下文和实体。Thenmalar、Balaji 和 Geetha 从少量训练数据开始，用识别出的实体、词向量和上下文特征来定义模型，将各类型实体定义的模型分别作为种子模型，分别进行训练和测试，取得 75% 的平均 *F* 值。何红磊[14]从大规模未标注语料中训练词向量，基于词向量的聚类和布朗聚类 3 种词表示特征，而后组合到 CRF 和 SVM 的特征子集中进行半监督学习，分别进行组合实验，最终 *F* 值提高了 1.59%。

3.2.4 无监督学习方法 最典型的就是聚类。无监督学习的数据集没有任何的标记或者有相同的标记，聚类的目的在于把内容或格式类似的数据聚在一起，即通过相似的上下文将不同的命名实体聚到一起。J Brooke、A Hammond 等[15]在无标注数据的情况下，首先在切分好的语料上进行布朗聚类（Brown Clustering），把结果当做 Bootstrap 的种子进行训练，同时加入文本级上下文分类器（Text - level Context Classifier）的概念，得到分类模型，随后改善短语名词分类不精准的缺点。目前国内对无监

督学习的研究较国外少一些，读者可参考其他领域的学习方法进行跨领域研究。

3.3 混合方法

目前生物医学命名实体识别研究大多使用几种技术方法的结合，借以弥补相互的缺点。近几年，国内关于生物医学命名实体识别的研究主要集中于对基因和蛋白质等实体研究。范文婷[16]针对 JNLPBA2004 任务，提出基于组合分类器和多代理策略的两阶段生物医学命名实体识别方法。构建 6 个单个分类器，运用两层叠加方法分别组合，然后应用多代理框架对组合结果分类，取得 76.06% 的 *F* 值。针对 BioCreative II GN 任务，提出多阶段基因标准化系统，主要包括预处理、词典查询、歧义消解和过滤 4 个步骤。杨娅[17]针对疾病名命名实体，提出词典与条件随机场相结合的实体识别方法。首先使用 PharmGKB 的资源构建疾病词典，结合条件随机场模型进行疾病名识别，然后运用全称 - 缩写词的上下文线索优化结果，取得 83.82% 的 *F* 值。曲春燕[18]针对中文电子病历进行研究，首先制定标注规范构建标注语料，然后采用最大熵、条件随机场和结构化支持向量机 3 种模型并融合组合分类算法，构建多种集合分类器，系统性能最优达到 92.97%。

3.4 神经网络方法

随着深度学习的兴起，基于机器学习的优缺点，为降低人工消耗和训练代价，研究者们将神经网络应用于自然语言处理领域，获得不少成果。2006 年 Hinton 在《科学》杂志上发表文章，讨论了在人工神经网络（Artificial Neural Network，ANN）训练中能有效减少数据维度和训练复杂度的方法，总结出深度神经网络（Deep Neural Network，DNN）具有更好的特征学习能力[19]。Wu 等[20]提出使用最小特征工程方法识别中文电子病历中的实体。结合 DNN 共比较 3 种方法，实验证明结合 DNN 的方法效果最好，性能提升得益于词向量捕获的语义信息。金留可[21]在 RNN 的基础上进行算法改进，在此基础上结合 LSTM 构建双向 LSTM（BLSTM）递归神经网络，而后融入句子向量，构建 ST - BLSTM 系统，在 BioCreative II GM 语料上取得 88.61% 的 *F* 值。在药物识别方面，DDIExtraction2011（DDI2011）和 DDIExtraction2013（DDI2013）引入药名实体识别任务，参赛方法大多依赖手工制定的特征工程和特定领域知识，因此 Zeng 等[22]提出自动探索词向量和特征的方法，LSTM 结合 CRF，融合从大量文本训练出的词向量和字符表示两种词表示方法，该方法优于 DDI2013 挑战中的最好系统。命名实体方法汇总，见表 1。由方法总结可看出，有 13 篇论文的语料是英文语料，只有 4 篇是基于中文生物医学的语料，在生物医学领域中文的开放语料比较少，相关研究较难开展，致使中文生物医学的发展前进缓慢。

表 1 命名实体方法汇总

方法	作者	实验资源	评价	语料语种
词典和 CRF	夏光辉	UMLS，GENIA 3.02	80.57%（*F*）	英文
规则和 CRF	程志刚	GATE，人民日报	人名 82.51%（*F*），地名 85.60%（*F*），机构名 75.20%（*F*）	中文
序列标注，HMM	魏尊强等	中医临床记录数据集	79.40%（*F*）	中文
层叠条件随机场	燕杨等	ICTCLAS 分词器，ICD - 10，ICD - 9 - CM，脑血管科室电子病历	97.02%（*F*）	中文
Hadoop，CRF	江林刚	Hadoop - 2.2.0，Spark - 1.0.0，MEDLINE 文献，JNLPBA 2004	识别效率分别提高 4.4 ~ 7.4 倍和 5 ~ 9 倍	英文
半监督	Thenmalar 等	IEER，CoNLL 2003，FIRE 语料库的 Tamil 语数据集	平均 75%（*F*）	英文，素米尔语
词表示，CRF，SVM	何红磊	BioCreative II GM	88.51%（*F*）	英文
Brown clustering，无监督	J Brooke 等	Project Gutenberg corpus	79.2%（*F*）	英文

续表 1

CRF，SVM，ME	范文婷	JNLPBA2004，BioCreative II GN	76.06%（*F*）	英文
词典，CRF	杨娅	PharmGKB，NCBI，MEDLINE，MEDIC	83.82%（*F*）	英文
ME，CRF，SSVM	曲春燕	自建标注规范，中文电子病历	92.97%（*F*）	中文
CRF，DNN	Yonghui Wu 等	中文电子病历	92.80%（*F*）	中文
RNN，LSTM	金留可	BioCreative II GM	88.61%（*F*）	英文
LSTM－CRF	Donghuo Zeng 等	DDI2013	79.26%（*F*）	英文

4 评测标准与评测组织

4.1 评测标准

目前命名实体识别的一般评测标准是精确率（precision），召回率（recall）和 *F* 测度（F－score）。*P* 是指系统正确识别的数量占识别出的实体总量的比例：

$$P = \frac{TP}{TP + FP}$$

R 是指系统正确识别出的命名实体数量占标准结果中命名实体总数的比例：

$$R = \frac{TP}{TP + FN}$$

F 值为 *P* 和 *R* 的调和平均值，*P* 和 *R* 是互相影响，所以在尽可能要求两者都高的情况下，一般使用 *F* 值评估，且已成为该领域默认的统一评估方法：

$$F = \frac{2PR}{P + R}$$

4.2 生物医学命名实体识别会议

4.2.1 概述 生物医学命名实体识别是一个跨学科的交叉领域，在很多领域都召开该主题的研讨会，如自然语言处理、生物信息学、机器学习等。目前国际上关于生物医学命名实体识别的国际会议主要有以下几项。

4.2.2 JNLPBA Joint Workshop on Natural Language Processing in Biomedicine and its Applications 是由 NLPBA 和 BioNLP 联合举办的测评会议。JNLPBA2004 的任务是从生物医学文献中识别出蛋白质、DNA、RNA、细胞系和细胞类型 5 类实体，评测指标是准确率、召回率和 *F* 值。JNLPBA 提供 GENIA V3.02 语料库，该语料库是目前在生物医学命名实体识别领域应用较为广泛，规模较大的标注语料库之一。

4.2.3 BioCreative Critical Assessment of Information Extraction systems in Biology 是由西班牙国家癌症研究中心、美国 MITRE 公司、美国生物技术信息中心、NCBI、CNIO 5 个机构联合举办的国际生物医学文本挖掘会议。

4.2.4 I2B2 从 2006 年开始组织有关电子病历信息挖掘的会议并构建相关语料库，其主要测评任务是命名实体识别和实体关系抽取。I2B2 2006 任务是去隐私和识别患者吸烟状态；I2B2 2008 任务是根据出院小结识别患者的肥胖相挑战；I2B2 2009 任务是识别病历文本中药品的相关属性，如药名，剂量等；I2B2 2010 关注病历中最主要的医疗实体，包括疾病名、症状、检查和治疗方法等；I2B2 2011 的任务是共指消解，即抽取实体间的等价关系；I2B2 2012 则是识别抽取与患者病情和治疗相关的时间，便于构建患者病情的时间线。

5 中文生物医学实体识别难点与意义

5.1 难点

由于中文的语言特点，中文文本缺少像英文文本中空格之类的词分隔符，因此中文命名实体识别通常先进行分词、词性标注等预处理。如药名、基因、蛋白质等命名实体命名复杂，缩略词形式多变，没有明确的规律可以参考，且中文命名实体在不同语境、领域的含义不同。基于词典的方法对词典等标准库的依赖性比较大，但国内缺少像 UMLS 和 RxNorm 此类标准库，研究者建立的实体库一般

规模较小且标准程度不高，在识别率和可信度方面有待考察。因此目前命名实体识别的一个研究趋势是大规模中文标准库的建立。基于统计的方法对语料库依赖性比较大，目前可用的基于基因/蛋白质识别的英文语料有 BioCreative、Yapex、GENIA、GENETAG 和 AIMed 等，在电子病历方面有 I2B2 测评会议提供的语料。在中文方面缺乏公开的标注语料，研究人员可以使用的生物医学文本仅有电子病历和生物医学文献，需要自行标注，制作训练语料，但缺乏统一的标准，因此难以进行统一的比较。且由于国内对患者隐私的保护政策，医院电子病历也较难获得。

5.2 意义

生物医学领域主要的语料是电子病历和生物医学文献。电子病历是医务人员在医疗活动过程中对患者医疗活动的记录，通过分析电子病历能挖掘出大量与患者相关的医疗知识[23]。病历的及时传输和获取，有助于对各种传染病和突发疾病的发现与预防，命名实体的共享传输接口比电子病历整体的接口容易实现且适用性高。生物医学文献是生物医学研究者的重要参考资料，从中可挖掘出未被发现的生物医学知识。从电子病历和生物医学文献中识别命名实体及抽取实体间的关系，还可用于临床决策支持（Clinical Decision Support，CDS），服务于医疗专业人员。临床决策支持旨在帮助卫生专业人员做出临床决策，即将患者的特征数据与支持系统中的知识库相匹配，目的是生成患者的特异性评估，供临床医生参考。上述从文献和电子病历中抽取的实体和关系是为 CDS 知识库的建立提供数据。由于国内缺少生物医学方面的标准库和大规模的语料库，所以中文文本挖掘技术的进展多少会受到影响。Bennett 利用 RxNorm 来捕获实时电子病历中患者的用药史并表示为 RxNorm 格式，方便传输和共享，便于医生在后续的治疗中根据用药史来调整治疗方案，减少医疗事故的发生[24]。国内缺少像 RxNorm 此类标准库，而标准库的建立又与实体识别紧密相关。只有准确高效地识别出生物医学文献或电子病历等文本中的基因、疾病名等命名实体，才能更好地进行基因或疾病名的标准化。

6 结语

本研究概述命名实体识别的发展历史，总结命名实体识别技术方法的发展和趋势以及在中文生物医学命名实体识别领域的技术难点和意义。中文生物医学命名实体识别在近几年的发展中已取得一定的成果，相关测评会议的召开有力地推动该领域研究的发展，但将研究成果真正投入应用还有一定难度，且在提高识别准确率方面仍面临着很大的挑战。因此认为以下几点研究趋势值得关注：（1）建立大规模、质量高的中文语料库。当前主流的命名实体识别方法是机器学习法，该方法需要利用标注语料训练模型，且系统性能的高低也依赖于语料的规模和质量。面对已有语料库完整性不足、标注标准不统一等问题，如何利用小规模的标注语料构建大规模的语料库是一个可参考的方向[25]。（2）构建中文生物医学标准库。电子病历和生物医学文献中充斥着大量的医学专业术语，标准库的建立有助于在实体识别时通过映射消除命名实体歧义性。如 UMLS、RxNorm、MeSH 等标准化词典，通过建立术语的标准化格式，支持计算机的高效检索，便于不同系统中的生物医学数据在一个适当的层面上实现数据共享。（3）关注神经网络、深度学习在命名实体识别方面的研究。面对缺乏标注语料、机器学习训练代价的问题，可把神经网络和深度学习应用于中文生物医学命名实体识别。Hinton 教授指出深度学习可通过模仿人脑的多层抽象机制实现对数据的抽象表达，可进行自主的特征学习，从而减少人工干预。近几年投入应用的神经网络模型有 CNN[26]、RNN[27]、LSTM[28]等模型，已取得一定的成果。近期在基于神经网络的命名实体识别研究中，主要集中在两个方面：一是利用 Attention Mechanism 来提高模型效果，二是针对少量标注数据的 Semi - supervised 研究[29]。

参考文献

1 Marrero M, Urbano J, Sánchez - Cuadrado S, et al. Named

Entity Recognition: fallacies, challenges and opportunities [J]. Computer Standards & Interfaces, 2013, 35 (5): 482 – 489.

2 Fleischman M, Hovy E. Fine Grained Classification of Named Entities [C]. Taipei: International Conference on Computational Linguistics. Association for Computational Linguistics, 2002: 1 – 7.

3 Murugesan G, Abdulkadhar S, Bhasuran B, et al. BCC – NER: bidirectional, contextual clues named entity tagger for gene/protein mention recognition [J]. Eurasip Journal on Bioinformatics & Systems Biology, 2017, 2017 (1): 7.

4 闫萍. 基于统计与规则相结合的命名实体识别研究 [D]. 郑州: 河南工业大学, 2012.

5 Tang B, Cao H, Wang X, et al. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks [J]. Biomed Research International, 2014, 2014 (2): 23 – 29.

6 何林娜, 杨志豪, 林鸿飞, 等. 基于特征耦合泛化的药名实体识别 [J]. 中文信息学报, 2014, 28 (2): 72 – 77.

7 Savova G K, Masanz J J, Ogren P V, et al. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications [J]. Journal of the American Medical Informatics Association Jamia, 2010, 17 (5): 507.

8 夏光辉. 基于词典与机器学习的基因命名实体识别机制研究 [D]. 北京: 北京协和医学院, 2013.

9 程志刚. 基于规则和条件随机场的中文命名实体识别方法研究 [D]. 武汉: 华中师范大学, 2015.

10 张向喆, 王明辉, 赵洪波, 等. 生物医学文本中命名实体识别研究 [J]. 上海交通大学学报 (农业科学版), 2010, 28 (2): 132 – 139.

11 魏尊强, 舒红平, 王亚强. 基于序列标注的中医症状名识别技术研究 [J]. 山东工业技术, 2015 (8): 237 – 238.

12 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 [J]. 吉林大学学报 (工), 2014, 44 (6): 1843 – 1848.

13 江林刚. 基于生物医学文献数据的命名实体识别并行算法研究 [D]. 长沙: 湖南大学, 2015.

14 何红磊. 基于词表示方法的生物医学命名实体识别 [D]. 大连: 大连理工大学, 2015.

15 Brooke J, Hammond A, Baldwin T. Bootstrapped Text – level Named Entity Recognition for Literature [C]. Berlin: Meeting of the Association for Computational Linguistics, 2016: 344 – 350.

16 范文婷. 生物医学领域的命名实体识别和标准化 [D]. 大连: 大连理工大学, 2013.

17 杨娅. 生物医学文本中的疾病实体识别和标准化研究 [D]. 大连: 大连理工大学, 2015.

18 曲春燕. 中文电子病历命名实体识别研究 [D]. 哈尔滨: 哈尔滨工业大学, 2015.

19 王国昱. 基于深度学习的中文命名实体识别研究 [D]. 北京: 北京工业大学, 2015.

20 Wu Y, Jiang M, Lei J, et al. Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network [J]. Studies in Health Technology & Informatics, 2015, (216): 624 – 628.

21 金留可. 基于递归神经网络的生物医学命名实体识别 [D]. 大连: 大连理工大学, 2016.

22 Zeng D, Sun C, Lin L, et al. LSTM – CRF for Drug – Named Entity Recognition [J]. Entropy, 2017, 19 (6): 283.

23 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.

24 Bennett C C. Utilizing RxNorm to Support Practical Computing Applications: capturing medication history in live electronic health records [J]. Journal of Biomedical Informatics, 2012, 45 (4): 634.

25 郑强, 刘齐军, 王正华, 等. 生物医学命名实体识别的研究与进展 [J]. 计算机应用研究, 2010, 27 (3): 811 – 815.

26 Collobert R, Weston J, Karlen M, et al. Natural Language Processing (Almost) from Scratch [J]. Journal of Machine Learning Research, 2011, 12 (1): 2493 – 2537.

27 Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition [C]. San Diego: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 260 – 270.

28 黄积杨. 基于双向 LSTMN 神经网络的中文分词研究分析 [D]. 南京: 南京大学, 2016.

29 robert_ ai. 神经网络结构在命名实体识别 (NER) 中的应用 [EB/OL]. [2017 – 11 – 01]. http://www.cnblogs.com/robert – dlut/p/6847401.html.