doi:10.3969/j.issn.1003-3114.2020.03.001
引用格式:陈曙东,欧阳小叶.命名实体识别技术综述[J].无线电通信技术,2020,46(3):251-260.
[CHEN Shudong,OUYANG Xiaoye.Overview of Named Entity Recognition Technology[J].Radio Communications Technology,2020,46(3):251-260.]

命名实体识别技术综述

陈曙东[1,2],欧阳小叶[1,2]
(1.中国科学院微电子研究所,北京 100029;
2.中国科学院大学,北京 100049)

摘　要: 命名实体识别是自然语言处理中的热点研究方向之一,目的是识别文本中的命名实体并将其归纳到相应的实体类型中。首先阐述了命名实体识别任务的定义、目标和意义,分析提出了命名实体识别的主要难点在于领域命名实体识别局限性、命名实体表述多样性和歧义性、命名实体的复杂性和开放性;然后介绍了命名实体识别研究的发展进程,从最初的规则和字典方法到传统的统计学习方法再到现在的深度学习方法,不断地将新技术应用到命名实体识别研究中以提高性能;接着系统梳理了当下命名实体识别任务中的若干热门研究点,分别是匮乏资源下的命名实体识别、细粒度命名实体识别、嵌套命名实体识别以及命名实体链接;最后针对评判命名实体识别模型的好坏,总结了常用的若干数据集和实验测评指标,并给出了未来的研究建议。

关键词: 自然语言处理;命名实体识别;深度学习;神经网络;人工智能

中图分类号: TP389.1　　**文献标志码:** A　　**开放科学标识码(OSID):**
文章编号: 1003-3114(2020)03-0251-10

Overview of Named Entity Recognition Technology

CHEN Shudong[1,2],OUYANG Xiaoye[1,2]
(1.Institute of Microelectronics of the Chinese Academy of Sciences,Beijing 100029,China;
2.University of Chinese Academy of Sciences,Beijing 100049,China)

Abstract: Named entity recognition is one of the valuable research directions in natural language processing. The purpose is to identify named entities from text and classify them into corresponding entity types. Firstly, this paper explains the definition, goals, and meaning of named entity recognition tasks, and analyzes the main difficulties of named entity recognition, namely, the limitations of domain named entity recognition, the diversity and ambiguity of named entity expressions, the complexity and openness of named entities. Then, it introduces the development roadmap of named entity recognition tasks, originally from rules and dictionary then traditional statistical learning methods, and deep learning methods currently, continuously improving the performance. Next, we systematically organized several popular research points, including low resources named entity recognition, fine-grained named entity recognition, nested named entity recognition, and named entity linking. And several commonly used datasets and experimental evaluation indicators for named entity recognition tasks are summarized, and future research recommendations are given. Finally, in order to estimate the quality of named entity recognition models, several commonly used datasets and experimental evaluation indicators are summarized, followed by our future research opinion.

Key words: natural language processing; named entity recognition; deep learning; neural network; artificial intelligence

0　引言

命名实体识别技术(Named Entity Recognition,NER)是人工智能领域的核心基础技术之一。1956年由麦卡锡、明斯基、罗彻斯特和香农共同组织召开的用机器模拟人类智能的专题讨论会上指出,人工智能主要研究用人工的方法和技术模仿、延

收稿日期: 2020-03-18
基金项目: 国家自然科学基金项目(61876144);中国科学院B类先导科技专项培育项目(XDPB12-3)
Foundation Item: National Nature Science Foundation of China (61876144); Chinese Academy of Sciences Leading Science and Technology Special Cultivation Project Class B (XDPB12-3)

伸、扩展智能，最终实现机器智能，而人工智能的长期目标是实现达到人类智力水平的人工智能[1]。为实现人工智能的目标，建造一个可以支撑自然语言处理和理解的大规模全方位知识库非常重要，但是当前由于人类知识存在的庞杂性、多样性、开放性等特性，建造辅助人工智能建设的大规模全方位知识库依旧任重道远。命名实体识别技术可以检测出文本中的新实体和相应类型，并加入到现有知识库中，为推动人工智能发展提供可靠的知识和技术基础。

由此可见，文本中的实体包含了丰富的语义，是至关重要的语义单元，从原始文本中识别有意义的实体或实体指代项在自然语言理解中起着至关重要的作用。这个过程通常被称为命名实体识别，即在文本中标识命名实体并划分到相应的实体类型中，通常实体类型包括人名、地名、组织机构名、日期等。举例说明，“当地时间 14 日下午，叙利亚一架军用直升机在阿勒坡西部乡村被一枚恶意飞弹击中。”这句话中包含的实体有：日期实体“14 日下午”、组织机构实体“叙利亚”、地名实体“阿勒坡西部乡村”、装备实体“军用直升机”和“飞弹”。由此可见，实体识别是文本意义理解的基础。

1991 年 Rau 等学者[2]首次提出了命名实体识别任务，随后自 1996 年开始，命名实体识别任务被加入到信息抽取领域，它作为一个子任务被引入各类测评任务中，如 MUC－6，MUC－7，IEER－99，CoNLL-2002，CoNLL-2003 等[3]。这些任务大多针对英文数据集开展研究，英文数据集句子中的每个词都是通过空格自然分开便于研究，当下在一些常见的公开数据集中准确率、召回率、F1 值均可达 90%左右。而中文数据集中汉字排列紧密，中文句子由多个字符组成且单词之间没有空格，这一自身独特的语言特征增大了命名实体识别的难度，但亦有学者在开展此方面研究并取得了不错的成果[4-5]。除此之外，西班牙语、德语、蒙古语等语言研究也有学者开展[6-7]。在不同语言的命名实体识别任务上，主要区别在于更多考虑不同语言特征对模型进行调整，而基础的技术理念和手段大多相似。因此本文不针对不同语言进行分别探讨，而从全局角度分析命名实体识别的任务难点、技术进展和当下研究热点。

综上所述，命名实体识别技术是海量文本数据分析的关键技术，可以用于解决互联网文本数据的爆炸式信息过载问题，以及处理互联网中存在的海量虚假、冗余、噪声数据导致的有效信息查找和浏览问题。命名实体识别技术从最初的规则和字典方法到传统的统计学习方法再到现在的深度学习方法，为非结构化的文本分析处理提供了有效的技术手段。目前命名实体识别技术在多种自然语言处理任务中有着广泛应用，例如知识图谱构建[8]、机器翻译[9]、知识库构建[10-11]、自动问答[12]、网络搜索[13]等。

1 研究难点

当前，一些学术界学者认为命名实体识别在很多开放数据集上已经取得了很高的准确率，被认为是一个不具有研究价值的问题。然而，我们在非常多的自然语言处理实际应用中发现，命名实体识别依旧具有很大的挑战性，还远没有得到很好的解决。经调研，我们认为命名实体识别在以下几个方向上仍然具有很强的应用研究价值。

1.1 领域命名实体识别局限性

目前命令实体识别只是在有限的领域和有限的实体类型中取得了较好的成绩，如针对新闻语料中的人名、地名、组织机构名的识别。但这些技术无法很好地迁移到其他特定领域中，如军事、医疗、生物、小语种语言等。一方面，由于不同领域的数据往往具有领域独特特征，如医疗领域中实体包括疾病、症状、药品等，而新闻领域的模型并不适合；另一方面，由于领域资源匮乏造成标注数据集缺失，导致模型训练很难直接开展。因此，采用半监督学习、远监督学习、无监督学习方法实现资源的自动构建和补足，以及迁移学习等技术的应用都可作为解决该问题的核心研究方向。

1.2 命名实体表述多样性和歧义性

自然语言的多样性和歧义性给自然语言理解带来了很大挑战，在不同的文化、领域、背景下，命名实体的外延有差异，是命名实体识别技术需要解决的根本问题。获取大量文本数据后，由于知识表示粒度不同、置信度相异、缺乏规范性约束等问题，出现命名实体表述多样、指代不明确等现象。因此，需要充分理解上下文语义来深度挖掘实体语义进行识别。可以通过实体链接、融合对齐等方法，挖掘更多有效信息和证据，实现实体不同表示的对齐、消除歧义，从而克服命名实体表述多样性和歧义性。

1.3 命名实体的复杂性和开放性

传统的实体类型只关注一小部分类型，例如

“人名”“地名”“组织机构名”，而命名实体的复杂性体现在实际数据中实体的类型复杂多样，需要识别细粒度的实体类型，将命名实体分配到更具体的实体类型中。目前业界还没有形成可遵循的严格的命名规范。命名实体的开放性是指命名实体内容和类型并非永久不变，会随着时间变化发生各种演变，甚至最终失效。命名实体的开放性和复杂性给实体分析带来了巨大的挑战，也是亟待解决的核心关键问题。

2 命名实体识别研究进展

命名实体识别从早期基于词典和规则的方法，到传统机器学习的方法，后来采用基于深度学习的方法，一直到当下热门的注意力机制、图神经网络等研究方法，命名实体识别技术路线随着时间在不断发展，技术发展趋势如图 1 所示。

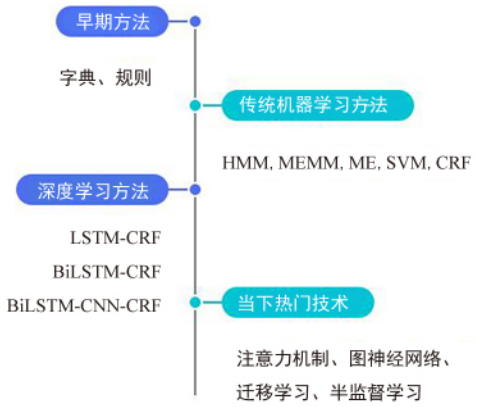

图 1 命名实体识别技术研究发展趋势

Fig.1 NER technology research development trend

2.1 基于规则和字典的方法

基于规则和字典的方法是最初代的命名实体识别使用的方法，这些方法多采用由语言学家通过人工方式，依据数据集特征构建的特定规则模板或者特殊词典。规则包括关键词、位置词、方位词、中心词、指示词、统计信息、标点符号等。词典是由特征词构成的词典和外部词典共同组成，外部词典指已有的常识词典。制定好规则和词典后，通常使用匹配的方式对文本进行处理以实现命名实体识别。

Rau 等学者[8]首次提出将人工编写的规则与启发式想法相结合的方法，实现了从文本中自动抽取公司名称类型的命名实体。这种基于规则的方法局限性非常明显，不仅需要消耗巨大的人力劳动，且不容易在其他实体类型或数据集扩展，无法适应数据的变化情况。

2.2 基于传统机器学习的方法

在基于机器学习的方法中，命名实体识别被当作是序列标注问题。与分类问题相比，序列标注问题中当前的预测标签不仅与当前的输入特征相关，还与之前的预测标签相关，即预测标签序列之间是有强相互依赖关系的。采用的传统机器学习方法主要包括：隐马尔可夫模型（Hidden Markov Model，HMM）、最大熵（Maximum Entropy，ME）[14]、最大熵马尔可夫模型（Maximum Entropy Markov Model，MEMM）[15]、支持向量机（Support Vector Machine，SVM）、条件随机场（Conditional Random Fields，CRF）[16]等。

在这 5 种学习方法中，ME 结构紧凑，具有较好的通用性，其主要缺点是训练时间复杂性非常高，甚至导致训练代价难以承受，另外由于需要明确的归一化计算，导致开销比较大。HMM 对转移概率和表现概率直接建模，统计共现概率。ME 和 SVM 在正确率上要 HMM 高一些，但是 HMM 在训练和识别时的速度要快一些。MEMM 对转移概率和表现概率建立联合概率，统计条件概率，但由于只在局部做归一化容易陷入局部最优。CRF 模型统计全局概率，在归一化时考虑数据在全局的分布，而不是仅仅在局部进行归一化，因此解决了 MEMM 中标记偏置的问题。在传统机器学习中，CRF 被看作是命名实体识别的主流模型，优点在于在对一个位置进行标注的过程中 CRF 可以利用内部及上下文特征信息。

还有学者通过调整方法的精确率和召回率对传统机器学习进行改进。Culotta 和 McCallum[17]计算从 CRF 模型提取的短语的置信度得分，将这些得分用于对实体识别进行排序和过滤。Carpenter[18]从 HMM 计算短语级别的条件概率，并尝试通过降低这些概率的阈值来增加对命名实体识别的召回率。对给定训练好的 CRF 模型，Minkov 等学者[19]通过微调特征的权重来判断是否是命名实体，更改权重可能会奖励或惩罚 CRF 解码过程中的实体识别。

2.3 基于深度学习的方法

随着深度学习的不断发展，命名实体识别的研究重点已转向深层神经网络（Deep Neural Network，DNN），该技术几乎不需要特征工程和领域知识[20-22]。Collobert 等学者[23]首次提出基于神经网络的命名实体识别方法，该方法中每个单词具有固定大小的窗口，但未能考虑长距离单词之间的有效信息。为了克服这一限制，Chiu 和 Nichols[24]提出

了一种双向 LSTM-CNNs 架构，该架构可自动检测单词和字符级别的特征。Ma 和 Hovy [25] 进一步将其扩展到 BiLSTM-CNNs-CRF 体系结构，其中添加了 CRF 模块以优化输出标签序列。Liu 等 [26] 提出了一种称为 LM-LSTM-CRF 的任务感知型神经语言模型，将字符感知型神经语言模型合并到一个多任务框架下，以提取字符级向量化表示。这些端到端模型具备从数据中自动学习的功能，可以很好地识别新实体。

部分学者将辅助信息和深度学习方法混合使用进行命名实体识别。Liu 等 [27] 在混合半马尔可夫条件随机场(Hybrid Semi-Markov Conditional Random Fields，HSCRFs) 的体系结构的基础上加入了 Gazetteers 地名词典，利用实体在地名词典的匹配结果作为命名实体识别的特征之一。一些研究尝试在标签级别跨数据集共享信息，Greenberg 等 [28] 提出了一个单一的 CRF 模型，使用异构标签集进行命名实体识别，此方法对平衡标签分布的领域数据集有实用性。Augenstein 等 [29] 使用标签向量化表示在任务之间进一步播信息。Beryozkin 等 [30] 建议使用给定的标签层次结构共同学习一个在所有标签集中共享其标签层的神经网络，取得了非常优异的性能。

近年来，在基于神经网络的结构上加入注意力机制、图神经网络、迁移学习、远监督学习等热门研究技术也是目前的主流研究方向，在下面研究热点中会穿插介绍。

3 研究热点

通过调研近三年来 ACL，AAAI，EMNLP，COLING，NAACL 等自然语言处理顶级会议中命名实体识别相关的论文，我们总结并选择了若干具有代表性的研究热点进行展开介绍，分别是匮乏资源命名实体识别、细粒度命名实体识别、嵌套命名实体识别、命名实体链接。

3.1 匮乏资源命名实体识别

命名实体识别通常需要大规模的标注数据集，例如标记句子中的每个单词，这样才能很好地训练模型。然而这种方法很难应用到标注数据少的领域，如生物、医学等领域。这是因为资源不足的情况下，模型无法充分学习隐藏的特征表示，传统的监督学习方法的性能会大大降低。

近来，越来越多的方法被提出用于解决低资源命名实体识别。一些学者采用迁移学习的方法，桥接富足资源和匮乏资源，命名实体识别的迁移学习方法可以分为两种：基于并行语料库的迁移学习和基于共享表示的迁移学习。利用并行语料库在高资源和低资源语言之间映射信息，Chen 和 Feng 等 [31-32] 提出同时识别和链接双语命名实体。Ni 和 Mayhew 等 [33] 创建了一个跨语言的命名实体识别系统，该系统通过将带注释的富足资源数据转换到匮乏资源上，很好地解决了匮乏资源问题。Zhou 等 [34] 采用双对抗网络探索高资源和低资源之间有效的特征融合，将对抗判别器和对抗训练集成在一个统一的框架中进行，实现了端到端的训练。

还有学者采用正样本-未标注样本学习方法(Positive-Unlabeled，PU)，仅使用未标注数据和部分不完善的命名实体字典来实现命名实体识别任务。Yang 等学者 [35] 采用 AdaSampling 方法，它最初将所有未标记的实例视为负实例，不断地迭代训练模型，最终将所有未标注的实例划分到相应的正负实例集中。Peng 等学者 [36] 实现了 PU 学习方法在命名实体识别中的应用，仅使用未标记的数据集和不完备的命名实体字典来执行命名实体识别任务，该方法无偏且一致地估算任务损失，并大大减少对字典大小的要求。

因此，针对资源匮乏领域标注数据的缺乏问题，基于迁移学习、对抗学习、远监督学习等方法被充分利用，解决资源匮乏领域的命名实体识别难题，降低人工标注工作量，也是最近研究的重点。

3.2 细粒度命名实体识别

为了智能地理解文本并提取大量信息，更精确地确定非结构化文本中提到的实体类型很有意义。通常这些实体类型在知识库的类型层次结构中可以形成类型路径 [37]，例如，牛顿可以按照如下类型的路径归类：物理学家/科学家/人。知识库中的类型通常为层次结构的组织形式，即类型层次。

大多数命名实体识别研究都集中在有限的实体类型上，MUC-7 [38] 只考虑了 3 类：人名、地名和组织机构名，CoNLL-03 [39] 增加了其他类，ACE [5] 引入了地缘政治、武器、车辆和设施 4 类实体，Ontonotes [40] 类型增加到 18 类，BBN [41] 有 29 种实体类型。Ling 和 Daniel [42] 定义了一个细粒度的112 个标签集，如图 2 所示，将标签问题表述为多类型多标签分类。

人物	医生	组织机构	恐怖组织				
演员	工程师	航空公司	政府机构				
建筑师	君主	公司	政府				
艺术家	音乐家	教育机构	政党				
运动员	政治家	兄弟会/姐妹会	教育部门				
作者	宗教领袖	体育联盟	军队				
教练	士兵	运动队	新闻机构				
导演	恐怖分子						
位置	水域	产品	照相机	艺术	写作		
城市	岛屿	发动机	手机	电影	新闻		
国家	山	飞机	电脑	演出	音乐		
省份	冰川	汽车	软件	事件	军事冲突		
县	星体	轮船	游戏	袭击	自然灾害		
铁路	公墓	飞船	仪器	竞选	体育事件		
路	公园	火车	武器	抗议	恐怖袭击		
桥							
建筑		时间		化学物质		网站	
机场		颜色		生物物质		广播网络	
坝		奖励		医学治疗		广播节目	
医院		教育程度		疾病		电视频道	
酒店		标题		症状		货币	
图书馆		法律		药物		股票交易	
发电站		种族		人体部位		算法	

图 2　定义的 112 个细粒度标签集

Fig.2　Defined 112 fine-grained label sets

学者们在该领域已经进行了许多研究，通常学习每个实体的分布式表示，并应用多标签分类模型进行类型推断。Neelakantan 和 Chang [43] 利用各种信息构造实体的特征表示，如实体的文字描述、属性和类型，之后，学习预测函数来推断实体是否为某类型的实例。Yaghoobzadeh 等 [44] 重点关注实体的名称和文本中的实体指代项，并为实体和类型对设计了两个评分模型。这些工作淡化了实体之间的内部关系，并单独为每个实体分配类型。Jin 等 [45] 以实体之间的内部关系为结构信息，构造实体图，进一步提出了一种网络嵌入框架学习实体之间的相关性。最近的研究表明以卷积方式同时包含节点特征和图结构信息，将实体特征丰富到图结构将获益颇多 [46-47]。此外，还有学者考虑到由于大多数知识库都不完整，缺乏实体类型信息，例如在 DBpedia 数据库中 36.53%的实体没有类型信息。因此对于每个未标记的实体，Jin 等 [48] 充分利用其文本描述、类型和属性来预测缺失的类型，将推断实体的细粒度类型问题转化成基于图的半监督分类问题，提出了使用分层多图卷积网络构造 3 种连通性矩阵，以捕获实体之间不同类型的语义相关性。

此外，实现知识库中命名实体的细粒度划分也是完善知识库的重要任务之一。细粒度命名实体识别现有方法大多是通过利用实体的固有特征（文本描述、属性和类型）或在文本中实体指代项来进行类型推断，最近有学者研究将知识库中的实体转换为实体图，并应用到基于图神经网络的算法模型中。

3.3　嵌套命名实体识别

通常要处理的命名实体是非嵌套实体，但是在实际应用中，嵌套实体非常多。大多数命名实体识别会忽略嵌套实体，无法在深层次文本理解中捕获更细粒度的语义信息。如图 3 所示，在“3 月 3 日，中国驻爱尔兰使馆提醒旅爱中国公民重视防控，稳妥合理加强防范。”句子中提到的中国驻爱尔兰使馆是一个嵌套实体，中国和爱尔兰均为地名，而中国驻爱尔兰使馆为组织机构名。普通的命名实体识别任务只会识别出其中的地名“中国”和“爱尔兰”，而忽略了整体的组织机构名。

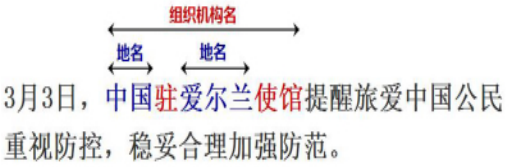

图 3　嵌套实体示例

Fig.3　Example of nested entity

学者们提出了多种用于嵌套命名实体识别的方法。Finkel 和 Manning [49] 基于 CRF 构建解析器，将每个命名实体作为解析树中的组成部分。Ju 等 [50] 动态堆叠多个扁平命名实体识别层，并基于内部命名实体识别提取外部实体。如果较短的实体被错误地识别，这类方法可能会遭受错误传播问题的困扰。嵌套命名实体识别的另一系列方法是基于超图的方法。Lu 和 Roth [51] 首次引入了超图，允许将边缘连接到不同类型的节点以表示嵌套实体。Muis 和 Lu [52] 使用多图表示法，并引入分隔符的概念用于嵌套实体检测。但是这样需要依靠手工提取的特征来识别嵌套实体，同时遭受结构歧义问题的困扰。Wang 和 Lu [53] 提出了一种使用神经网络获取分布式特征表示的神经分段超图模型。Katiyar 和 Cardie [54] 提出了一种基于超图的计算公式，并以贪

婪学习的方式使用 LSTM 神经网络学习嵌套结构。这些方法都存在超图的虚假结构问题，因为它们枚举了代表实体的节点、类型和边界的组合。Xia 等[6] 提出了 MGNER 架构，不仅可以识别句子中非重叠的命名实体，也可以识别嵌套实体，此外不同于传统的序列标注任务，它将命名实体识别任务分成两部分开展，首先识别实体，然后进行实体分类。

嵌套实体识别充分利用内部和外部实体的嵌套信息，从底层文本中捕获更细粒度的语义，实现更深层次的文本理解，研究意义重大。

3.4 命名实体链接

命名实体链接主要目标是进行实体消歧，从实体指代项对应的多个候选实体中选择意思最相近的一个实体。这些候选实体可能选自通用知识库，例如维基百科、百度百科[55]，也可能来自领域知识库，例如军事知识库、装备知识库。图 4 给出了一个实体链接的示例。短文本“美海军陆战队 F/A-18C 战斗机安装了生产型 AN/APG-83 雷达”，其中实体指代项是“生产型 AN/APG-83 雷达”，该实体指代项在知识库中可能存在多种表示和含义，而在此处短文本，其正确的含义为“AN/APG-83 可扩展敏捷波束雷达”。

图 4 实体链接示例

Fig.4 Example of named entity linking

实体链接的关键在于获取语句中更多的语义，通常使用两种方法。一种是通过外部语料库获取更多的辅助信息，另一种是对本地信息的深入了解以获取更多与实体指代项相关的信息[56]。Tan 等[57] 提出了一种候选实体选择方法，使用整个包含实体指代项的句子而不是单独的实体指代项来搜索知识库，以获得候选实体集，通过句子检索可以获取更多的语义信息，并获得更准确的结果。Lin 等[58] 寻找更多线索来选择候选实体，这些线索被视为种子实体指代项，用作实体指代项与候选实体的桥梁。Dai 等[59] 使用社交平台 Yelp 的特征信息，包括用户名、用户评论和网站评论，丰富了实体指代项相关的辅助信息，实现了实体指代项的歧义消除。因此，与实体指代项相关的辅助信息将通过实体指代项和候选实体的链接实现更精确的歧义消除。

另一些学者使用深度学习研究文本语义。Francis-Landau 等[60] 使用卷积神经网络学习文本的表示形式，然后获得候选实体向量和文本向量的余弦相似度得分。Ganea 和 Hofmann[61] 专注于文档级别的歧义消除，使用神经网络和注意力机制来深度表示实体指代项和候选实体之间的关系。Mueller 和 Durrett[62] 将句子左右分开，然后分别使用门控循环单元和注意力机制，获得关于实体指代项和候选实体的分数。Ouyang 等[4] 提出一种基于深度序列匹配网络的实体链接算法，综合考虑实体之间的内容相似度和结构相似性，从而帮助机器理解底层数据。目前，在实体链接中使用深度学习方法是一个热门的研究课题。

4 公开数据集和评价指标

4.1 数据集

常用的命名实体识别数据集有 CoNLL 2003，CoNLL 2002，ACE 2004，ACE 2005 等。数据集的具体介绍如下：

① CoNLL 2003 数据集[35] 包括 1 393 篇英语新闻文章和 909 篇德语新闻文章，英语语料库是免费的，德国语料库需要收费。英语语料取自路透社收集的共享任务数据集。数据集中标注了 4 种实体类型：PER，LOC，ORG，MISC。

② CoNLL 2002 数据集[63] 是从西班牙 EFE 新闻机构收集的西班牙共享任务数据集。数据集标注了 4 种实体类型：PER，LOC，ORG，MISC。

③ ACE 2004 多语种训练语料库[5] 版权属于语言数据联盟（Linguistic Data Consortium，LDC），ACE 2004多语言培训语料库包含用于 2004 年自动内容提取（ACE）技术评估的全套英语、阿拉伯语和中文培训数据。语言集由为实体和关系标注的各种类型的数据组成。

④ ACE 2005 多语种训练语料库[5] 版权属于 LDC，包含完整的英语、阿拉伯语和汉语训练数据，数据来源包括：微博、广播新闻、新闻组、广播对话等，可以用来做实体、关系、事件抽取等任务。

⑤ OntoNotes 5.0 数据集[37] 版权属于 LDC，由 1 745 K英语、900 K 中文和 300 K 阿拉伯语文本数据组成，OntoNotes 5.0 的数据来源也多种多样，来自电话对话、新闻通讯社、广播新闻、广播对话和博客等。实体被标注为 PERSON，ORGANIZATION，LOCATION 等 18 个类型。

⑥ MUC 7 数据集[34] 是发布的可以用于命名实体识别任务，版权属于 LDC，下载需要支付一定费用。数据取自北美新闻文本语料库的新闻标题，其

中包含 190 K 训练集、64 K 测试集。

⑦ Twitter 数据集是由 Zhang 等[64]提供，数据收集于 Twitter，训练集包含了 4 000 推特文章，3 257 条推特用户测试。该数据集不仅包含文本信息还包含了图片信息。

大部分数据集的发布官方都直接给出了训练集、验证集和测试集的划分。同时不同的数据集可能采用不同的标注方法，最常见的标注方法有 IOB，BIOES，Markup，IO，BMEWO 等，下面详细介绍几种常用的标注方法：

① IOB 标注法，是 CoNLL 2003 采用的标注法，I 表示内部，O 表示外部，B 表示开始。如若语料中某个词标注 B/I-XXX，B/I 表示这个词属于命名实体的开始或内部，即该词是命名实体的一部分，XXX 表示命名实体的类型。当词标注 O 则表示属于命名实体的外部，即它不是一个命名实体。

② BIOES 标注法，是在 IOB 方法上的扩展，具有更完备的标注规则。其中 B 表示这个词处于一个命名实体的开始，I 表示内部，O 表示外部，E 表示这个词处于一个实体的结束，S 表示这个词是单独形成一个命名实体。BIOES 是目前最通用的命名实体标注方法。

③ Markup 标注法，是 OntoNotes 数据集使用的标注方法，方式较简单。例如：ENAMEX TYPE = " ORG " > LondonENAMEX > is an international metropolis，它直接用标签把命名实体标注出来，然后通过 TYPE 字段设置相应的类型。

4.2 评价指标

目前，命名实体识别任务常采用的评价指标有精确率（Precision）、召回率（Recall）、$F1$ 值（$F1$-Measure）等。

精确率：对给定数据集，分类正确样本个数和总样本数的比值。即：

$$Precision = \frac{TP + TN}{TP + FN + FP + TN},$$

式中，TP 指将正预测为真，FN 指将正预测为假，FP 指将反预测为真，TN 指将反预测为假。

召回率：用来说明分类器中判定为真的正例占总正例的比率，即：

$$Recall = \frac{TP}{TP + FN};$$

$F1$ 值：是精确率和召回率的调和平均指标，是平衡准确率和召回率影响的综合指标。

$$\frac{1}{F1} = \frac{1}{Recall} + \frac{1}{Precision}。$$

5 结束语

命名实体识别是自然语言处理应用中的重要步骤，它不仅检测出实体边界，还检测出命名实体的类型，是文本意义理解的基础。本文指出了命名实体识别研究存在的难点，包括领域命名实体识别局限性、命名实体表述多样性和歧义性、命名实体的复杂性和开放性。还阐述了命名实体识别的研究进展，从早期基于规则和词典的方法，到传统机器学习的方法，到近年来基于深度学习的方法，神经网络与 CRF 模型相结合的 NN-CRF 模型依旧是目前命名实体识别的主流模型。同时，本文还介绍了当下的多个热门研究点，其中匮乏资源领域的命名实体识别在 NLP 领域应用有着非常巨大的价值，迁移学习、对抗学习、远监督学习方法以及图神经网络、注意力机制等新型技术都是未来研究的重点。

参 考 文 献

[1] CHINCHOR N，ROBINSON P.MUC-7Named Entity Task Definition [C] // Proceedings of the 7th Conference on Message Understanding，1997，29：1-21.

[2] RAU L F. ExtractingCompany Names from Text [C] // Proceedings of the Seventh IEEE Conference on Artificial Intelligence Application.IEEE，1991，1：29-32.

[3] SUN Z，WANG H. Overview on the Advance of the Research on Named Entity Recognition [J]. New Technology of Library and Information Service，2010，26（6）：42-47.

[4] OUYANG X，CHEN S，ZHAO H，et al. A Multi-Cross Matching Network for Chinese Named Entity Linking in Short Text [C] // Journal of Physics：Conference Series. IOP Publishing，2019，1325（1）：012069.

[5] ZHANG Y，YANG J.Chinese Ner Using Lattice Lstm [J]. arXiv preprint arXiv：1805.02023，2018.

[6] XIA C，ZHANG C，YANG T，et al.Multi-Grained Named Entity Recognition [J]. arXiv preprint arXiv：1906.08449，2019.

[7] NI J，FLORIAN R. Improving Multilingual Named Entity Recognition with Wikipedia Entity Type Mapping [J]. arXiv preprint arXiv：1707.02459，2017.

[8] XIE R，LIU Z，JIA J，et al. Representation Learning of Knowledge Graphs with Entity Descriptions [C] // Thirtieth AAAI Conference on Artificial

Intelligence,2016.

[9] BABYCH B,HARTLEY A.Improving Machine Translation Quality with Automatic Named Entity Recognition [C] // Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools,Improving MT Through Other Language Technology Tools,Resource and Tools for Building MT at EACL 2003,2003.

[10] RIEDEL S, YAO L, MCCALLUM A, et al. Relation Extraction with Matrix Factorization and Universal Schemas [C] // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2013: 74-84.

[11] SHEN W,WANG J,LUO P,et al.Linden: Linking Named Entities with Knowledge Base Via Semantic Knowledge [C] // Proceedings of the 21st International Conference on World Wide Web,2012: 449-458.

[12] BORDES A,USUNIER N,CHOPRA S,et al.Large-scale Simple Question Answering with Memory Networks [J]. arXiv preprint arXiv: 1506.02075,2015.

[13] ZHU J,UREN V,MOTTA E. ESpotter: Adaptive Named Entity Recognition for Web Browsing [C] // Biennial Conference on Professional Knowledge Management/Wissensmanagement.Springer,Berlin,Heidelberg,2005: 518-529.

[14] RATNAPARKHI A.A Maximum Entropy Model for Part-of-speech Tagging [C] // Conference on Empirical Methods in Natural Language Processing,1996: 133-142.

[15] MCCALLUM A,FREITAG D,PEREIRA F C N.Maximum Entropy Markov Models for Information Extraction and Segmentation [C] // Icml,2000,17: 591-598.

[16] LAFFERTY J,MCCALLUM A,PEREIRA F C N.Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence data [C] // Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001): 282-289.

[17] CULOTTA A,MCCALLUM A. Confidence Estimation for Information Extraction [C] // Proceedings of HLT-NAACL 2004: Short Papers,2004: 109-112.

[18] CARPENTER B.Ling Pipe for 99.99% Recall of Gene Mentions [C] // Proceedings of the Second BioCreative Challenge Evaluation Workshop. BioCreative, 2007, 23: 307-309.

[19] MINKOV E, WANG R C, TOMASIC A, et al. NER systems that Suit User's Preferences: Adjusting the Recall-precision Trade-off for Entity Extraction [C] // Proceedings of the Human Language Technology Conference of the NAACL,Companion Volume: Short Papers.2006: 93-96.

[20] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al.Neural Architectures for Named Entity Recognition [J]. arXiv preprint arXiv: 1603.01360,2016.

[21] ŽUKOV-GREGORIČ A, BACHRACH Y, COOPE S. Named Entity Recognition with Parallel Recurrent Neural Networks [C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),2018: 69-74.

[22] ZHOU J T, ZHANG H, JIN D, et al. Roseq: Robust Sequence Labeling [J]. IEEE Transactions on Neural Networks and Learning Systems,2019,PP(99): 1-11.

[23] COLLOBERT R,WESTON J,BOTTOU L,et al. Natural Language Processing (almost) from Scratch [J]. Journal of Machine Learning Research, 2011, 12(Aug): 2493-2537.

[24] CHIU J P C,NICHOLS E.Named Entity Recognition with Bidirectional LSTM-CNNs [J]. Transactions of the Association for Computational Linguistics,2016,4: 357-370.

[25] MA X,HOVY E. End-to-end Sequence Labeling Via Bi-directional Lstm-cnns-crf [J]. arXiv preprint arXiv: 1603.01354,2016.

[26] LIU L,SHANG J,REN X,et al.Empower Sequence Labeling with Task-aware Neural Language Model [C] // Thirty-Second AAAI Conference on Artificial Intelligence.2018.

[27] LIU T,YAO J G,LIN C Y. Towards Improving Neural Named Entity Recognition with Gazetteers [C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,2019: 5301-5307.

[28] GREENBERG N,BANSAL T,VERGA P,et al. Marginal Likelihood Training of Bilstm-crf for Biomedical Named Entity Recognition from Disjoint Label Sets [C] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,2018: 2824-2829.

[29] AUGENSTEIN I, RUDER S, SØGAARD A. Multi-task Learning of Pairwise Sequence Classification Tasks over Disparate Label Spaces [J]. arXiv preprint arXiv: 1802.09913,2018.

[30] BERYOZKIN G, DRORI Y, GILON O, et al. A Joint Named-Entity Recognizer for Heterogeneous Tag-sets Using a Tag Hierarchy [J]. arXiv preprint arXiv: 1905.09135,2019.

[31] CHEN Y,ZONG C,SU K Y.On Jointly Recognizing and Aligning Bilingual Named Entities [C] // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,2010: 631-639.

[32] FENG X,FENG X,QIN B,et al.Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge

Transfer [C] //IJCAI.2018: 4071-4077.

[33] MAYHEW S, TSAI C T, ROTH D. Cheap Translation for Cross - lingual Named Entity Recognition [C] // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 2536-2545.

[34] ZHOU J T, ZHANG H, JIN D, et al. Dual Adversarial Neural Transfer for Low-resource Named Entity Recognition [C] //Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 3461-3471.

[35] YANG P, LIU W, YANG J. Positive Unlabeled Learning Via Wrapper - based Adaptive Sampling [C] // IJCAI. 2017: 3273-3279.

[36] PENG M, XING X, ZHANG Q, et al. Distantly Supervised Named Entity Recognition using Positive - Unlabeled Learning [J] .arXiv preprint arXiv: 1906.01378, 2019.

[37] REN X, HE W, QU M, et al. Afet: Automatic Fine-grained Entity Typing by Hierarchical Partial - label Embedding [C] //Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 1369 -1378.

[38] CHINCHOR N A. Overview of Muc - 7/met - 2 [R]. Science Applications International Corp San Diego CA, 1998.

[39] SANG E F, DE MEULDER F. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition [J] .arXiv preprint cs/0306050, 2003.

[40] HOVY E, MARCUS M, PALMER M, et al. OntoNotes: the 90% solution [C] //Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, 2006: 57-60.

[41] WEISCHEDEL R, BRUNSTEIN A. BBN Pronoun Coreference and Entity Type Corpus [J] .Linguistic Data Consortium, Philadelphia, 2005, 112.

[42] LING X, WELD D S. Fine-grained Entity Recognition [C] //Twenty - Sixth AAAI Conference on Artificial Intelligence, 2012.

[43] NEELAKANTAN A, CHANG M W. Inferring Missing Entity Type Instances for Knowledge Base Completion: New dataset and methods [J] .arXiv preprint arXiv: 1504. 06658, 2015.

[44] YAGHOOBZADEH Y, SCHÜTZE H. Multi - level Representations for Fine - grained Typing of Knowledge Base Entities [J] .arXiv preprint arXiv: 1701.02025, 2017.

[45] JIN H, HOU L, LI J, et al. Attributed and Predictive Entity Embedding for Fine-grained Entity Typing in Knowledge Bases [C] //Proceedings of the 27th International Conference on Computational Linguistics, 2018: 282-292.

[46] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering [C] //Advances in Neural Information Processing Systems, 2016: 3844-3852.

[47] ATWOOD J, TOWSLEY D. Diffusion - convolutional Neural Networks [C] // Advances in Neural Information Processing Systems, 2016: 1993-2001.

[48] JIN H, HOU L, LI J, et al. Fine - Grained Entity Typing Via Hierarchical Multi Graph Convolutional Networks [C] //Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 4970-4979.

[49] FINKEL J R, MANNING C D. Nested Named Entity Recognition [C] // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1. Association for Computational Linguistics, 2009: 141-150.

[50] JU M, MIWA M, ANANIADOU S. A Neural Layered Model for Nested Named Entity Recognition [C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018: 1446-1459.

[51] LU W, ROTH D. Joint Mention Extraction and Classification with Mention Hypergraphs [C] // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 857-867.

[52] MUIS A O, LU W. Labeling Gaps Between Words: Recognizing Overlapping Mentions with Mention Separators [J] . arXiv preprint arXiv: 1810.09073, 2018.

[53] WANG B, LU W. Neural Segmental Hypergraphs for Overlapping Mention Recognition [J] . arXiv preprint arXiv: 1810.01817, 2018.

[54] KATIYAR A, CARDIE C. Nested Named Entity Recognition Revisited [C] //Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018: 861-871.

[55] LE P, TITOV I. Improving Entity Linking by Modeling Latent Relations between Mentions [J] . arXiv preprint arXiv: 1804.10637, 2018.

[56] GUPTA N, SINGH S, ROTH D. Entity Linking Via Joint Encoding of Types, Descriptions, and Context [C] // Pro-

ceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.2017: 2681-2690.

[57] TAN C,WEI F,REN P,et al. Entity Linking for Queries by Searching Wikipedia Sentences [J] .arXiv preprint arXiv: 1704.02788,2017.

[58] LIN Y,LIN C Y,JI H. List-only Entity Linking [C] // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) , 2017: 536-541.

[59] DAI H,SONG Y,QIU L,et al. Entity Linking within a Social Media Platform: A Case Study on Yelp [C] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,2018: 2023-2032.

[60] FRANCIS-LANDAU M, DURRETT G, KLEIN D. Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks [J] . arXiv preprint arXiv: 1604.00734,2016.

[61] GANEA O E,HOFMANN T.Deep Joint Entity Disambiguation with Local Neural Attention [J] . arXiv preprint arXiv: 1704.04920,2017.

[62] MUELLER D,DURRETT G. Effective Use of Context in Noisy Entity Linking [C] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,2018: 1024-1029.

[63] SANG E F,DE MEULDER F.Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition [J] .arXiv preprint cs/0306050,2003.

[64] ZHANG Q,FU J,LIU X,et al. AdaptiveCo-attention Network for Named Entity Recognition in Tweets [C] // Thirty-Second AAAI Conference on Artificial Intelligence,2018.

作者简介:

陈曙东　中国科学院微电子研究所研究员、博士生导师,中国科学院大学教授。目前围绕数据智能进行科学技术研究和智慧城市领域的社会发展民生工程建设。发表学术论文 40 余篇、拥有专利 30 多项和软件著作权 2 项,著有 7 本专著。

欧阳小叶　中国科学院微电子研究所博士生。主要研究方向: 知识图谱构建与应用。参与实验室多项知识图谱构建与应用项目,在国内外期刊会议发表论文 5 篇、专利 4 项。

“专家论坛”栏目,旨在刊登知名专家撰写的有关通信领域的热点技术、发展趋势等综述或研究类文章,达到启发引领行业发展的作用。