计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2022/16(09)-1954-15
doi: 10.3778/j.issn.1673-9418.2112109

命名实体识别方法研究综述

李冬梅[1,2],罗斯斯[1,2],张小平[3+],许　福[1,2]
1. 北京林业大学 信息学院,北京 100083
2. 国家林业和草原局林业智能信息处理工程技术研究中心,北京 100083
3. 中国中医科学院 中医药信息研究所,北京 100700
+ 通信作者 E-mail: xiao_ping_zhang@139.com

摘　要:在自然语言处理领域,命名实体识别是信息抽取的第一个关键环节。命名实体识别任务旨在从大量非结构化的文本中识别出命名实体并将其分类为预定义的类型,为关系抽取、文本摘要和机器翻译等自然语言处理任务提供基础支持。首先概述了命名实体识别的定义、研究难点和中文命名实体识别任务的特殊性,总结了命名实体识别任务中常用的中英文公共数据集和评估标准。然后根据命名实体识别的发展历程调研了现有的命名实体识别方法,主要为早期基于规则和词典的命名实体识别方法、基于统计机器学习的命名实体识别方法和基于深度学习的命名实体识别方法。归纳总结了每一种命名实体识别方法的关键思路、优缺点和具有代表性的模型,同时对各阶段的中文命名实体识别方法进行了总结。特别对最新的基于Transformer和基于提示学习的命名实体识别方法进行了综述,这两种细分类的方法是基于深度学习的命名实体识别方法中最先进的方法。最后总结了命名实体识别研究面临的挑战,并展望了未来的研究方向。

关键词:自然语言处理;命名实体识别;机器学习;深度学习;关系抽取
文献标志码:A　**中图分类号:**TP391

Review on Named Entity Recognition

LI Dongmei[1,2], LUO Sisi[1,2], ZHANG Xiaoping[3+], XU Fu[1,2]
1. School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China
2. Engineering Research Center for Forestry-Oriented Intelligent Information Processing, National Forestry and Grassland Administration, Beijing 100083, China
3. Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China

Abstract: In the field of natural language processing, named entity recognition is the first key step of information extraction. Named entity recognition task aims to recognize named entities from a large number of unstructured texts and classify them into predefined types. Named entity recognition provides basic support for many natural language processing tasks such as relationship extraction, text summarization, machine translation, etc. This paper first introduces the definition of named entity recognition, research difficulties, particularity of Chinese named entity recognition, and summarizes the common Chinese and English public datasets and evaluation criteria in named entity recognition tasks. Then, according to the development history of named entity recognition, the existing named entity recognition methods are investigated, which are the early named entity recognition methods based on

基金项目:中央级公益性科研院所基本科研业务费专项资金(ZZ140319-W);国家自然科学基金(61772078)。
This work was supported by the Fundamental Research Funds for the Central Public Welfare Research Institutes (ZZ140319-W), and the National Natural Science Foundation of China (61772078).
收稿日期:2021-12-29　**修回日期:**2022-04-29

rules and dictionaries, the named entity recognition methods based on statistic and machine learning, and the named entity recognition methods based on deep learning. This paper summarizes the key ideas, advantages and disadvantages and representative models of each named entity recognition method, and summarizes the Chinese named entity recognition methods in each stage. In particular, the latest named entity recognition based on Transformer and based on prompt learning are reviewed, which are state-of-the-art in deep learning-based named entity recognition methods. Finally, the challenges and future research trends of named entity recognition are discussed.

Key words: natural language processing; named entity recognition; machine learning; deep learning; relation extraction

命名实体识别(named entity recognition，NER)是指识别出文本中具有特定意义的命名实体并将其分类为预先定义的实体类型，如人名、地名、机构名、时间、货币等。在大数据时代，如何精准并高效地从海量无结构或半结构数据中获取到关键信息，这是自然语言处理(natural language processing，NLP)任务的重要基础。命名实体通常包含丰富的语义，与数据中的关键信息有着密切的联系，NER任务可以用于解决互联网文本数据的爆炸式信息过载问题，能有效获取到关键信息，并广泛应用于关系抽取、机器翻译以及知识图谱构建等领域。

NER历经了MUC(message understanding conference)[1]、MET(multilingual entity task)[2]、CoNLL(conference on computational natural language learning)[3]、ACE(automatic content extraction)[4]等，众多研究者不断深入研究，其理论和方法愈加完善。研究方法从最初需要人工设计规则，到后来借助传统机器学习中的模型方法，目前已经发展到利用各种深度学习。研究领域从一般领域到特定领域，研究语言从单一语言发展到多种语言，各种NER模型的性能随着发展也在不断提升。

本文调研了NER发展史上有代表性的综述论文，孙镇等[5]综述了NER的方法，包括对基于规则和词典的方法以及基于统计的方法的介绍。Li等[6]详细总结和分析了NER的深度学习方法。李猛等[7]从迁移学习的角度，总结了NER的迁移方法。赵山等[8]调研了在不同神经网络架构下最具代表性的晶格结构的中文NER模型。以上综述都是对NER的传统方法或者深度学习的部分方法的阐述，没有详细地包含基于规则和词典的NER方法、基于统计机器学习的NER方法和基于深度学习的NER方法这三者的介绍，且并未针对最新的基于提示学习的方法进行总结。本文首先从基于规则和词典、基于统计机器学习和基于深度学习的NER方法这三方面对目前NER研究工作进行系统性梳理，归纳总结了每一种NER方法的关键思路、优缺点和具有代表性的模型。同时对基于提示学习的NER方法进行了比较分析。其次，扩充了中文NER的介绍，给出了中文NER的特殊性，总结归纳中文NER特有的数据集，对各个阶段的主流方法均单独进行了综述。

1 NER概述

1.1 NER定义

最初在1991年第7届IEEE人工智能应用会议上，Rau[9]发表了一篇“从文本中抽取公司名称”的论文，提出了一种从文本中提取公司名的方法，在文中需要识别的命名实体仅为公司名称。在1996年MUC-6会议上，命名实体被定义为“实体的唯一标识符”，需要识别的命名实体包含：人名、地名、机构名、时间、日期、货币和百分比[1]。

NER是对文本中的命名实体进行定位和分类的过程。对给定文本的标注序列 $S=<w_1, w_2, \cdots, w_n>$，经过NER过程后得到三元组列表，如 $<I_s, I_e, t>$，每一个三元组都包含一个实体的信息。在三元组 $<I_s, I_e, t>$ 中，$I_s \in [1, n]$，$I_e \in [1, n]$，分别指代实体的开始索引和结束索引，t 是预定义类别集合中的实体类型[6]。图1给出了一个标注序列的样例，在经过NER系统后得到了3个三元组，根据三元组判断得到：Zhang San是Person类实体，Beijing和China是Location类实体。

1.2 NER的研究难点

目前，针对NER的研究仍存在一些通用难点。

(1)未登录词。随着时间的推移和各领域发展，会产生大量新实体，这些新产生的实体并没有一个统一的命名规则，传统的方法不再适用，此时要求NER模型具有较强的上下文推理能力。

(2)嵌套实体。嵌套实体是指该实体中存在其他命名实体。这类实体不仅需要识别外层实体，还

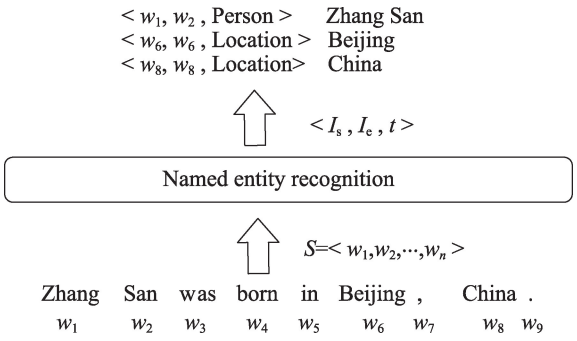


图1 NER任务的实例

Fig.1 Example of NER task

要识别内层实体,对模型来说具有很大的挑战,这也是目前NER的一个研究热点。

(3)文本歧义。文本在某处为命名实体,而在另一处为普通名词,或者为不同的实体类型,即文本类型是不明确的。因此,需要在NER之前进行额外的命名实体消歧任务。

(4)非正式文本。随着社交软件的流行,如微博等社交媒体中含有大量的语料,但这些语料有着简短、口语化、包含谐音等特点,这使得NER任务更加难于处理,可以使用注意力机制和迁移学习结合深度学习完成对非正式文本的识别。

1.3 中文NER的特殊性

面向中文的NER起步较晚,而且中文与英文等其他语言相差较大,由于其自身的语言特性,中文领域的NER主要存在以下3个特殊性。

(1)中文词语的边界不明确。中文的单元词汇边界模糊,缺少英文文本中空格这样明确的分隔符,也没有明显的词形变换特征,因此容易造成许多边界歧义,从而加大了NER的难度。

(2)中文NER需要同中文分词和语法分析相结合。只有准确的中文分词和语法分析才能正确划分出命名实体,才能提升NER的性能,这也额外增加了中文NER的难度。

(3)中文存在多义性、句式复杂表达灵活、多省略等特点。在不同领域的同一词语所表示的含义并不相同,且同一语义也可能存在多种表达。此外,互联网的迅速发展,尤其是网络文本中的文字描述更加个性化和随意化,这都使得实体的识别更加困难。

1.4 NER常用数据集

常用于NER的英文数据集有:MUC-6、MUC-7、CoNLL2002-2003、ACE2004-2005、GENIA、OntoNotes5.0、BC5CDR、NCBI Disease、Few-NERD等。中文数据集有:1998年人民日报数据集、MSRA、OntoNotes5.0、BosonNLP NER、Weibo NER、Chinese Resume、CCKS2017-2020、CLUENER2020等。以上数据集总结如表1所示。

1.5 评估标准

在NER领域,通常使用准确率(precision)、召回率(recall)和F1值作为评估指标。其中,准确率是对于给定的测试数据集,分类器正确识别的实体样本数与提取出来的全部实体样本数之比;召回率则是对于给定的测试数据集,分类器正确识别测试集中的全部实体的百分比;而F1值则是准确率和召回率的调和平均值,可以对系统的性能进行综合性评价。准确率、召回率和F1值的计算公式如下:

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{3}$$

其中,TP 表示将正例预测为正;FP 表示将负例预测为正;FN 表示将正例预测为负。

2 NER的方法

根据NER的发展历程,主流的NER方法可以分为3类:基于规则和词典的方法、基于统计机器学习的方法和基于深度学习的方法。这3类方法根据处理特点又细分为若干种不同的子方法。图2给出了NER方法的详细分类,后面的内容围绕该分类方法分别进行详细阐述。

2.1 基于规则和词典的NER方法

早期的NER方法主要运用由语言学专家根据语言知识特性手工构造的规则模板,通过匹配的方式实现命名实体的识别。针对不同的数据集通常需要构造特定的规则,一般根据特定统计信息、标点符号、关键字、指示词和方向词、位置词、中心词等特征来构造。Krupka[10]提出了一个用于英文NER的SRA系统,系统包括NameTag和HASTEN两个子系统,HASTEN根据文本的语义信息来构造生成人名和地名规则模板,进一步来识别。Shaalan等[11]利用文本的上下文特征构造规则,并同时增加地名词典来识别专业名词。

对于中文NER,最初的研究聚焦于专业名词的研究,张小衡等[12]根据机构名称的结构规律和形态标记等特点进一步总结规则,从600多万的三地语料库

表1 NER数据集总结

Table 1 Summary of NER datasets

数据集	时间	语言	语料来源	实体类型数量	URL
MUC-6 MUC-7	1996 1997	英文	新闻	共7类：人名、地名、机构名、日期、时间、货币和百分比	https://catalog.ldc.upenn.edu/LDC-2001T02
1998年人民日报数据集	1998	中文	人民日报	共3类：人名、地名和机构名	https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/renMinRiBao
CoNLL2002-2003	2002—2003	英文、德文	新闻	共4类：人名、地名、机构名和其他实体	https://www.clips.uantwerpen.be/conll2002/ner/
GENIA	2004	英文	生物学和临床文本	共36个细粒度实体类型	http://www.geniaproject.org/home
ACE2004-2005	2004—2005	英文、阿拉伯文、中文	新闻、博客	共7类：人名、地名、机构名、地理政治、设施、交通工具和武器	https://www.ldc.upenn.edu/collaborations/past-projects/ace
MSRA	2006	中文	新闻	共3类：人名、地名和机构名	http://www.sighan.org/bakeoff2006/
OntoNotes5.0	2013	英文、阿拉伯文、中文	新闻、博客、宗教文本	共18类：人名、地名、机构名、地理政治、设施、产品、事件实体等	https://catalog.ldc.upenn.edu/LDC-2013T19
BosonNLP NER	2014	中文	网络文本	共6类：人名、地名、机构名、时间、公司名和产品实体	http://bosonnlp.com
NCBI Disease	2014	英文	PubMed摘要	共790个细粒度实体类型	https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/
BC5CDR	2015	英文	PubMed摘要	共3类：化学物质、疾病名和化学-疾病相互作用	http://bioc.sourceforge.net/
CCKS2017	2017	中文	电子病历	共5类：症状体征、检查检验、疾病诊断、治疗和身体部位	https://www.biendata.xyz/competition/ CCKS2017_2/
Weibo NER	2018	中文	博客	共4类：人名、地名、机构名和地理-政治	https://github.com/hltcoe/golden-horse
Chinese resume	2018	中文	简历	共8类：人名、地名、机构名、国家、教育机构、职业、种族和职称	https://github.com/jiesutd/LatticeLSTM
CCKS2018	2018	中文	电子病历	共5类：解剖部位、症状描述、独立症状、药物和手术	https://www.biendata.xyz/competition/CCKS2018_1/
CCKS2019-2020	2019—2020	中文	电子病历	共6类：疾病和诊断、检查、检验、药物、手术以及解剖部位	https://www.biendata.xyz/competition/ccks_2020_2_1/
CLUENER2020	2020	中文	新闻	共10类：人名、地名、机构名、公司、政府等	https://github.com/CLUEbenchmark/CLUENER2020
Few-NERD	2021	英文	维基百科	共8个粗粒度、66个细粒度	https://ningding97.github.io/fewnerd/

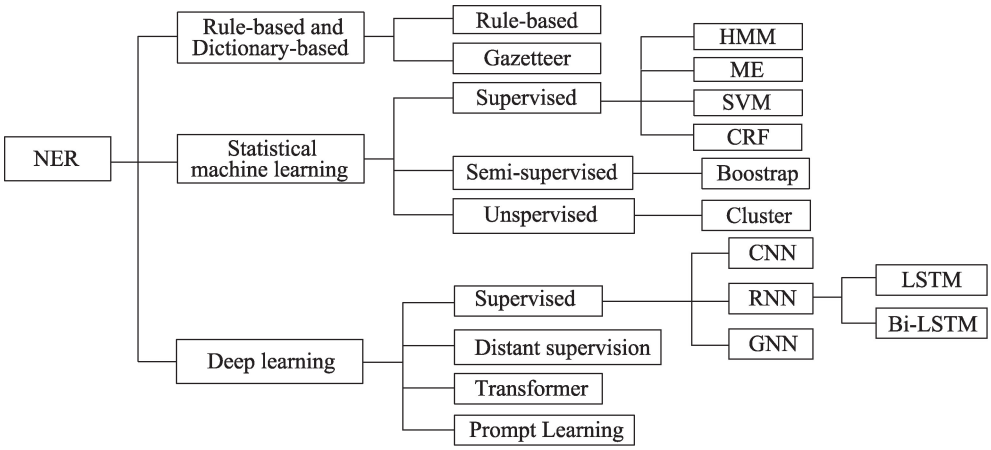


图2 NER方法的分类

Fig.2 Classification of NER methods

中识别高校名称实体,正确率达到了97.3%。王宁等[13]从专业名词识别的角度,充分考虑金融领域的特征,利用规则的方法专门针对公司名的识别问题进行了研究。该方法分析研究了金融新闻文本,总结了公司名的结构特征以及上下文信息,归纳形成知识库,并采取两次扫描的策略进行识别。在共1 336篇真实金融新闻的数据集上进行实验,其中在封闭测试环境中的准确率和召回率分别为97.13%和89.13%,在开放测试环境中分别为62.18%和62.11%。

表2对上述提及到的方法进行了总结。基于规则和词典的方法可以利用相关语言特性或特定领域知识来制定规则,在特定的语料库中该类方法具有较好的识别效果,但是该方法存在规则制定成本高、规则无法移植到其他语料等局限性。因此在其他大型语料中单纯依靠这种方法较难获得有效的识别结果。

2.2 基于统计机器学习的NER方法

随着机器学习在NLP领域的兴起,研究者们借助机器学习的方法研究NER。这种方法可以在一定程度上克服基于规则和词典的NER方法的局限性,该类方法可以归纳为三种:有监督学习、半监督学习和无监督学习。

2.2.1 有监督学习

有监督学习的NER方法是将NER任务转换成分类问题,通过机器学习方法将已标记的语料构造为特征向量,以此建立分类模型来识别实体。基于特征的有监督学习的NER方法一般流程包括:(1)获取实验原始数据。(2)对原始数据预处理。(3)根据数据的文本信息,选择合适的特征。(4)给不同的特征设置不同的权重并选择合适的分类器训练特征向量,得到NER模型。(5)利用NER模型进行实体识别。(6)对结果进行评估。

采用有监督机器学习的分类模型包括:HMM(hidden Markov models)、MEM(maximum entropy models)、SVM(support vector machines)和CRF(conditional random fields)等模型。

(1)HMM

基于HMM的NER方法利用维特比算法将可能的目标序列分配给每个单词序列,能够捕捉现象的局部性,进而提高了实体识别性能。Bikel等[14]基于大小写、数字符号、句子首词等特征,利用HMM来计算某一单词为某一实体类型的概率。但该模型仍然无法捕捉到远距离信息,还存在一些无法识别的实体。Zhou等[15]提出一种基于HMM的组块标记器的NER方法,在Bikel的基础上扩充了内部语义特征、内部地名词典特征以及外部上下文特征,对HMM的传统公式做了改进,以便能融合更多的上下文信息来确定当前预测类型。

对于中文NER,张华平等[16]借助HMM提出了基于角色标注的中国人名自动识别方法。该方法采取HMM对分词结果进行角色标注,通过对最佳角色序列的最大匹配来识别和分类命名实体,该方法解决了不具备明显特征的姓名的丢失、内部成词以及上下文成词的人名难召回的问题。俞鸿魁等[17]提出一种基于层叠HMM的中文NER模型,该模型由三级HMM构成。在分词后低层的HMM识别普通无嵌套的人名、地名和机构名等,高层的HMM识别嵌套的人名、地名和机构名。

(2)MEM

基于MEM的NER方法的主要思想是在已知部分知识的前提下选择熵最大的概率分布,从而来确定某一实体的类型,MEM能够较好地融合多种特征信息进行分类。Borthwick等[18]最早将MEM用于英文NER任务,综合考虑了首字母大小写、句子的结尾信息以及文本是否为标题等多种特征信息。Bender等[19]在Borthwick的基础上进行改进,模型结构依次为输入序列、预处理、全局搜索、后处理和序列标注。

对于中文NER,周雅倩等[20]最早将MEM应用在中文名词短语的识别上,将短语识别问题转化为标注问题。利用预定义的特征模板从语料中抽取候选特征,然后根据候选特征集识别名词短语。但该模型未能将更多的语义、词语共现等信息融合在模型

表2 基于规则和词典的主流NER方法总结

Table 2 Summary of mainstream NER methods based on rules and dictionaries

方法	时间	支持语言	数据集	F1值/%	方法关键字
Krupka[10]	1995	英文	MUC-6	96.42	人名和地名规则
Shaalan等[11]	2009	阿拉伯文	ACE、政府提供以及网上数据构成的数据集	92.26	地名词典+规则
张小衡等[12]	1997	中文	香港理工大学的三地现代汉语数据集	97.30	机构名规则
王宁等[13]	2002	中文	互联网金融新闻构成的数据集	89.13	公司名规则

中。因此，张玥杰等[21]提出一种融合多特征的MEM中文NER模型，该模型能集成局部与全局多种特征，将规则和机器学习的方法相结合，分别构建了局部特征模板和全局特征模板，同时引入启发式知识解决效率和空间问题。

（3）SVM

SVM是定义为特征空间上的间隔最大的线性分类器。首先通过高维特征空间的转化使分类问题转换成线性可分问题，然后基于结构风险最小理论构建最优分割超平面，使得分类器得到全局最优化。该模型在NER任务上被广泛使用，Isozaki等[22]提出了一种基于SVM的特征选择方法以及有效的训练方法，能增加系统训练的速度。为了验证SVM在不同领域的表现效果，Takeuchi等[23]在MUC-6评测语料与分子生物学领域语料使用SVM进行实体识别，发现SVM在生物领域的NER具有良好的表现。

对于中文NER，李丽双等[24]提出一种基于SVM的中文地名的自动识别的方法，结合地名的特点信息作为向量的特征。此外，面对训练数据不足的难点，陈霄等[25]针对中文组织机构名的识别任务，提出了一种基于SVM的分布递增式学习的方法，利用主动学习的策略对训练样本进行选择，逐步增加分类器训练样本的规模，进一步提高分类器的识别精度。

（4）CRF

CRF模型统计了全局概率，不仅在局部进行归一化，且考虑了数据在全局的分布情况。CRF具有表达长距离依赖性和交叠性的优势，能有效融入上下文信息以及领域知识，可以解决标注偏置问题。即使CRF具有时间复杂度高导致的训练难度大等问题，但仍十分广泛地被用于NER。McCallum等[26]提出了一种基于CRF的特征归纳的NER方法，与传统方法相比，自动归纳特征既提高了准确性，又显著减少了特征数量。Krishnan等[27]提出了一种利用非局部依赖且基于两个耦合的CRF分类器的方法。第一层CRF利用局部信息提取特征，第二层CRF将局部信息和从第一层CRF的输出中提取的特征结合，在整个文档中使用特征去捕捉非局部的依赖信息。

对于中文NER，冯元勇等[28]在CRF框架中引入了小规模的常用尾字特征来降低特征集的规模，在提高模型训练速度同时保证识别准确率。燕杨等[29]针对中文电子病历的NER问题，提出一种层叠CRF，该模型在第二层中使用包含实体和词性等特征的特征集，对疾病名称和临床症状两类命名实体进行识别。与无自定义组合特征的层叠CRF相比，该模型的F1值提高了约3个百分点，和单层CRF相比，F1值提高了约7个百分点。

综上所述，以上几种有监督机器学习NER方法各有所长，也各有所短。研究者充分利用各种算法的优势，进一步提升实体识别的性能。上述几种方法的相关比较如表3所示。

2.2.2 半监督学习

有监督学习的方法需要专家手工标注大量训练数据，为了解决这一问题，学者开始研究利用少量的标注语料进行NER任务，因此，半监督的NER方法应运而生。该方法通过使用少量标记和大量无标记的语料库进行NER的研究。半监督学习NER的一般流程：（1）人工构造初始种子集合。（2）根据命名实体上下文信息生成相关联的模式。（3）将生成的模式和测试数据匹配，标识出新的命名实体，生成新的模式，便于促进循环。（4）将新识别的命名实体添加到实体集合中。流程图如图3所示。

半监督学习的NER方法主要采用自举的方法，该方法利用少量的标注数据进行训练，从而取得良好的实验结果。如Teixeira等[30]提出一种基于CRF的自举训练方法，首先基于词典对50 000条新闻标注人名，并使用标注好人名的数据作为训练集建立基于CRF的分类模型。然后使用CRF分类模型对初始

表3 基于有监督机器学习NER比较

Table 3 Comparison of NER methods for supervised machine learning

模型	原理	优点	缺点	代表文献
HMM	对转移概率和表现概率直接建模，统计共现概率	时间复杂度低	准确率比MEM略低	[14-17]
MEM	对转移概率和表现概率建立联合概率，统计的是条件概率	准确率比HMM高	时间复杂度高	[18-21]
SVM	特征空间上的间隔最大的线性分类器	利用内积核函数代替向高维空间的非线性映射	大规模样本训练和多分类效果差	[22-25]
CRF	统计全局概率，不仅在局部归一化，考虑数据在全局的分布	考虑数据的全局分布，解决了标注偏置问题	时间复杂度高	[26-29]

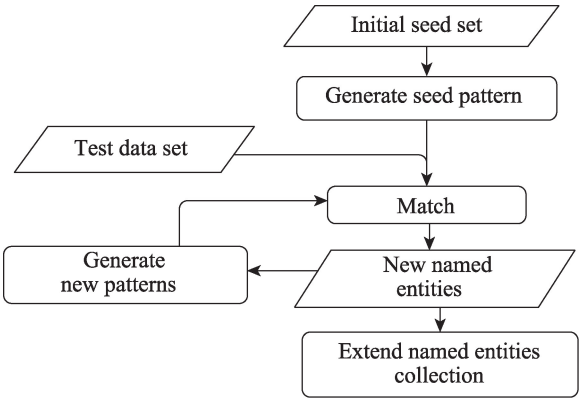


图3 半监督学习的NER一般流程

Fig.3 General process of NER based on semi-supervised learning

种子语料库额外标注，并将其用于训练新的分类模型。该模型经过7次自举方法的迭代后，在HAREM数据集上进行实验表现良好。此外，Thenmalar等[31]不仅在英文语料中使用半监督的自举方法，还增加了泰米尔文语料进一步验证该方法的可行性。该方法利用少量训练数据中命名实体、单词和上下文特征来定义模式，分别对英文和泰米尔文进行NER，两种语言的平均F1值为75%。

对于中文NER，针对结构复杂的产品名的识别任务，黄诗琳等[32]提出一种半监督学习方法，提取不同产品实体的结构特征和相互关系，构建一种三层半监督学习框架。首层结合规则和词典选取数据集中的候选数据；第二层利用相似度算法，把与种子集上下文相似的候选词加入正例中，这一步骤能解决数据稀疏问题；第三层是一个CRF的分类器用于识别相似度较低的实体。但因产品名的表达方式多样化，该方法与一般的NER方法相比，性能还存在一定的差距。在医学NER任务上，Long等[33]提出一个基于自举的NER方法，在自举训练过程中将命名实体特征集表示为类特征向量，候选命名实体的上下文信息表示为示例特征向量，这两种特征向量的相似程度决定了候选实体是否为命名实体。此外，针对少数民族语言的NER任务，王路路等[34]以CRF为基本框架，通过引入词法特征、词典特征以及基于词向量的无监督学习特征，对比不同特征对识别结果的影响，进而得到最优模型。

2.2.3 无监督学习

为了解决跨域和跨语言标注文本的不足，学者们提出了NER的无监督学习技术。无监督学习是不需要使用标注数据的算法，该方法使用未标注的数据来做出决策。无监督学习旨在考虑数据的结构和分布特征，从而发现更多关于数据的学习。

早期，Etzioni等[35]提出了一个名为KnowwitAll的无监督NER系统，该系统以无监督和可扩展的方式自动地从网页中提取大量命名实体。Nadeau等[36]在Etzioni等的基础上进一步研究，该系统可以自动构建地名词典以及消解命名实体歧义，将构建的地名词典与常用的地名词典相结合。Han等[37]提出一个基于聚类主动学习的生物医学NER系统，该聚类方法通过使用底层分类器在文档中查找候选命名实体来进行聚类，因而更能反映命名实体的分布。

综上所述，无监督学习的NER方法既能解决有监督学习中需要大量带标注的训练数据的问题，也不需要少量标注的种子数据，但是这种方法需要提前确定聚类阈值并且性能较低，仍需进一步改善聚类方法。

对基于有监督、半监督、无监督的三种方法进行了比较，如表4所示，并对基于统计机器学习的各种主流NER模型进行了总结，如表5所示。

2.3 基于深度学习的NER方法

基于深度学习的方法对处理NER等序列标注任务的处理流程是类似的。首先，将序列通过Word2Vec等编码方式转换成分布式表示，随后将句子的特征表示输入到编码器中，利用神经网络自动提取特征，最后使用CRF等解码器来预测序列中词对应的标签。早期，研究者大多对基于有监督和远程监督两种深度学习的NER方法进行深入研究。预训练模型BERT（bidirectional encoder representation from transformers）[38]自2018年提出以来，也备受研究

表4 基于有监督、半监督、无监督的NER比较

Table 4 Comparison of NER methods for supervised, semi-supervised and unsupervised

方法	实现方法	域泛化能力	优点	缺点	改进方法
有监督学习	分类	最弱	充分利用先验知识，针对特定的域	需要大量的标注数据，可移植性差	增加特征，增加标记语料
半监督学习	分类	较强	需要少量的语料	需要大量的分析和后期处理	扩展模式，减少噪音
无监督学习	聚类	最强	不需要标注语料库，用于大规模未标注语料	需要提前确定聚类阈值，性能较低	扩展特征，改善聚类

表5 基于统计机器学习的主流NER模型总结

Table 5 Summary of mainstream NER models for statistical machine learning

方法	时间	支持语言	数据集	F1值/%	方法关键字
Borthwick等[18]	1998	英文	MUC-7	92.05	MEM、知识库
Bikel等[14]	1999	英文	MUC-6	94.92	HMM
Zhou等[15]	2002	英文	MUC-6、MUC-7	96.60、94.10	HMM、特征扩充
Isozaki等[22]	2002	英文	General数据集	90.03	SVM
Bender等[19]	2003	英文	CoNLL2003	89.26	MEM
McCallum等[26]	2003	英文	CoNLL2003	84.04	单层CRF
张华平等[16]	2004	中文	人民日报数据集	人名95.40	HMM、角色标注
俞鸿魁等[17]	2006	中文	人民日报数据集	均值91.20	层叠HMM
Krishnan等[27]	2006	英文	CoNLL2003	87.24	双层CRF
Nadeau等[36]	2006	英文	MUC-7	69.33	无监督学习、地名词典
张玥杰等[21]	2008	中文	SIGHAN 2008	87.92	MEM、规则
李丽双等[24]	2007	中文	人民日报数据集	90.12	SVM
陈霄等[25]	2008	中文	人民日报数据集	84.18	SVM
冯元勇等[28]	2008	中文	863简体NER评测数据集	88.76	单层CRF、尾字特征
Teixeira等[30]	2011	英文	HAREM数据集	68.00	自举方法、CRF
黄诗琳等[32]	2013	中文	新闻和网页文档构成的数据集	78.20	自举方法、规则和词典、CRF
燕杨等[29]	2014	中文	临床医院65分电子病历数据集	97.02	层叠CRF
Thenmalar等[31]	2015	英文	CoNLL2003	82.57	自举方法
Long等[33]	2014	中文	医生相关的文本构成的数据集	均值94.75	自举方法
王路路等[34]	2018	维吾尔文	新疆多语种信息技术实验室提供的数据集	87.43	半监督学习、CRF
Han等[37]	2016	英文	BioCreative提供的数据集	81.40	聚类方法、主动学习

者关注。最近，基于提示学习的方法也在NER任务上得到了初步尝试，并取得了成功。

基于深度学习的NER方法一般流程如图4所示，共分为4步[6]：(1)Sequence，预处理后的输入序列。(2)Word embedding，将输入序列转换成固定长度的向量表示。(3)Context encoder，将词嵌入进行语义编码。(4)Tag decoder，进一步进行标签解码。

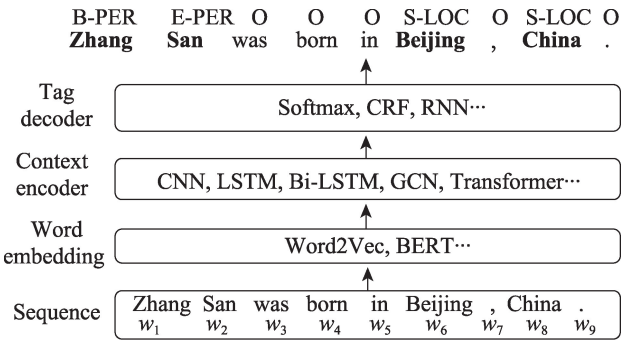


图4 基于深度学习的NER一般流程

Fig.4 General process of NER based on deep learning

2.3.1 基于有监督深度学习的方法

基于有监督深度学习的方法目前主要分为CNN(convolutional neural network)、RNN(recurrent neural network)和GNN(graph neural network)等。

(1)CNN

早期，CNN在计算机视觉领域取得突破性成果，后来也逐渐在NLP领域被广泛使用。2011年，Collobert等[39]提出了一种基于CNN的NLP模型，能处理包含NER等多种任务。该模型不需要利用人工输入特征，而是基于大量未标记的训练数据来学习内部表示，在输入时减少特征的预处理，使用以端到端方式训练的多层神经网络体系结构。在Collobert等的基础上，Yao等[40]将CNN应用到生物医学NER上，模型具有多层结构，每层根据底层生成的特征提取特征。该模型具有良好准确率，但并未充分利用CPU并行性，其计算效率不高，因此，Strubell等[41]提出了一种迭代扩张卷积神经网络(ID-CNNs)，与传统的CNN相比，该模型具有更好的上下文和结构化预测能力并能大幅缩短训练时间。

对于中文NER，2015年Wu等[42]利用卷积层生成由多个全局隐藏节点表示的全局特征，然后利用局部特征和全局特征以识别临床文本中的命名实体。

Wu等[43]提出了一种CNN-LSTM-CRF,以获取短距离和长距离内容依赖,同时提出将NER和分词任务联合学习以挖掘这两个任务之间的内在联系,增强中文NER模型识别实体边界的能力,但该模型无法捕捉全局的上下文信息。因此,Kong等[44]提出一种融合多层次CNN和注意力机制的中文临床NER方法。该方法既能捕捉短距离和长距离的上下文信息,且注意力机制还能获取全局上下文信息,进一步解决了LSTM在句子较长时无法捕捉全局信息的问题。但该方法目前对稀有命名实体仍然存在难以识别的问题,因此,Gui等[45]将词典信息融合到CNN结构中,解决稀有实体识别的问题。

综上所述,CNN最大的特点是可以并行化,每个时间状态不受上一时间状态的影响,但其无法很好地提取序列信息。随着RNN的深入研究,CNN和RNN常常混合使用。

(2)RNN

RNN是基于深度学习的NER方法中的主流模型,RNN将语言视为序列数据,能很好地处理序列数据,解决了CNN无法记忆上下文信息的问题。Huang等[46]在Collobert等基础上,提出了多种基于LSTM的序列标注模型,包括LSTM、Bi-LSTM和Bi-LSTM-CRF等。首次将Bi-LSTM-CRF模型用于NER,该模型不仅可以同时利用上下文的信息,而且可以使用句子作为输入。Gregoric等[47]在同一输入端采用多个独立的Bi-LSTM单元,通过使用模型间正则化来促进LSTM单元之间的多样性,能够减少模型的参数。Li等[48]提出一个模块化交互网络模型用于NER,能同时利用段级信息和词级依赖。Xu等[49]提出一种有监督多头自注意网络的NER模型,利用自我注意力机制获取句子中词与词之间的关系,并引入一个多任务学习框架来捕捉实体边界检测和实体分类之间的依赖关系。

对于中文NER,Zhang等[50]首次提出了基于混合字符和词典的Lattice-LSTM模型,通过门控单元,将词汇信息嵌入到每个字符中,从而利用上下文中有用的词汇提升NER效果。但是由于词汇的长度和数量无法确定,Lattice-LSTM存在无法批量训练而导致模型训练较慢的问题。为了解决该问题,Liu等[51]提出了基于单词的LSTM(WC-LSTM)。该方法在输入的向量中融入最优词汇的信息,在正向LSTM中融入基于该字开头的词汇信息,在反向LSTM中融入基于该字结尾的词汇信息。Ma等[52]也在Lattice-LSTM模型基础上做了改进,不修改LSTM的内部结构,只在输入层进行词与所有匹配到的词汇信息的融合,该方法还可以应用到不同的序列模型框架中,如CNN和Transformer。

(3)GNN

近年来,GCN(graph convolutional network)和GGNN(gated graph neural network)在NER任务中得到广泛的关注。Cetoli等[53]率先在NER任务中使用图GCN来解决实体识别问题,在传统的Bi-LSTM-CRF模型的Bi-LSTM层和CRF层中间额外添加一层GCN层。Bi-GCN层利用句子的句法依存关系构图,通过GCN将节点信息传递给最近的节点,通过将N层图堆叠在一起,该网络结构可以传播最多相距N跳的节点特征。

在中文领域,为了解决在NER过程中使用词典的最长匹配和最短匹配带来的问题,Ding等[54]提出了一种基于GNN并结合地名词典的NER方法,其目的使模型自动学习词典的特征。该模型首先根据地名词典构图,然后依次通过GGNN层、LSTM层和CRF层进行实体的识别。Gui等[55]通过引入一个具有全局语义的基于词典的GNN模型来获取全局信息。此外,Tang等[56]进一步研究了如何将词汇信息整合到基于字符的方法中,提出一种基于单词-字符图卷积网络(WC-GCN),通过使用交叉GCN块同时处理两个有向无环图,并引入全局GCN块来学习全局上下文的节点表示。

2.3.2 基于远程监督深度学习的方法

基于远程监督深度学习的方法主要利用外部词典或知识库对无标注数据进行标注,可以解决有监督学习需要大量已标注数据这一问题,其常采用的方式包括词典匹配和词典匹配与神经网络相融合两种。Peng等[57]仅借助未标记数据和命名实体词典,提出了一种新的PU(positive-unlabeled)远程监督NER模型,该模型不需要利用词典标记句子中的每个实体,能大幅度降低对词典质量的要求。此外,Yang等[58]提出了一个基于部分标注学习和强化学习的远程监督的NER模型,不仅可以通过远程监督自动获取到大规模的训练数据,而且通过使用部分标注学习和强化标注学习,解决了远程监督方法产生的不完全标注和噪音标注的问题。

对于中文NER,Zhang等[59]利用远程监督的方法识别时间,提出了一种利用中文知识图谱和百度百科生成的数据集进行模型训练的方法,该方法不需要像手动标注数据,且对不同类型的文本的适应性

良好。此外，边俐菁[60]基于深度学习和远程监督的方法针对产品进行实体识别，利用爬虫整理得到的词典高质量地标注数据，按照词典完全匹配、完全匹配+规则、核心词汇+词性扩展+规则这三种方式进行实体识别，该方法能大大减少手工标注语料库的工作量。

远程监督的方法相对于有监督的方法极大地减少了人工成本，但远程监督的方法会产生不完全标注和噪音标注，导致自动标注获得的数据集准确率较低，会影响整个NER模型的性能。

2.3.3 基于Transformer的方法

基于Transformer方法典型代表是BERT类的预训练模型。Souza等[61]在NER任务上提出一种BERT-CRF模型，将BERT的传输能力与CRF的结构化预测相结合。Naseem等[62]提出一种针对生物医学NER的预训练语言模型BioALBERT，该模型在ALBERT中使用自我监督损失，能较好学习上下文相关的信息。Yang等[63]提出了一种分层的Transformer模型，应用于嵌套的NER。实体表征学习结合了以自下而上和自上而下的方式聚集的相邻序列的上下文信息。

对于中文NER，李妮等[64]提出了基于BERT-IDCNN-CRF的中文NER模型，该模型通过BERT预训练模型得到字的上下文表示，再将字向量序列输入IDCNN-CRF模型中进行训练。Li等[65]为解决大规模标记的临床数据匮乏问题，在未标记的中国临床电子病历文本上利用BERT模型进行预训练，从而利用未标记的领域特定知识，同时将词典特征整合到模型中，利用汉字字根特征进一步提高模型的性能。Wu等[66]在Li等的基础上，提出了一个基于RoBERTa和字根特征的模型，使用RoBERTa学习医学特征，同时利用Bi-LSTM提取偏旁部首特征和RoBERTa学习到医学特征向量做拼接，解码层使用CRF进行标签解码。Yao等[67]针对制造文本进行细粒度实体识别，提出一种基于ALBERT-AttBiLSTM-CRF和迁移学习的模型，使用更轻量级的预训练模型ALBERT对原始数据进行词嵌入，Bi-LSTM提取词嵌入的特征并获取上下文的信息，解码层使用CRF进行标签解码。

2.3.4 基于提示学习的方法

随着NLP技术的发展，近两年有研究者在低资源任务中使用提示学习的方法来获得良好的任务效果。提示学习通常不需要改变预训练语言模型的结构和参数，而是通过向输入中添加一些提示信息，并修改下游任务来适应预训练模型，进而获得更好的任务效果的一种方法。Brown等[68]首次在文本分类任务中使用提示学习的方法进行了小样本学习任务。在低资源的NER任务中，没有大规模的训练语料，大量依赖训练数据的模型都无法取得较好的效果。因此在低资源的NER任务中使用提示学习是现在的一种新思路。Cui等[69]提出一种基于模板的NER方法，再利用标注实体填充的预定义模板提示对预训练模型BART（bidirectional and auto-regressive Transformers）微调，该方法解决了小样本NER的问题。Chen等[70]受提示学习的启发，提出一种轻量级的低资源提示引导型注意生成框架，将连续的提示输入到自我注意层中，来重新调节注意力并调整预先训练的权重。基于模板提示的方法需要枚举所有可能的候选实体，存在较高的计算复杂度问题，因此，Ma等[71]提出一种在小样本场景下无模板的提示微调方法，放弃模板构建的枚举思路，采用预训练任务中的掩码预测任务的方式，将NER任务转化成将实体位置的词预测为标签词的任务。该方法能减少预训练和微调之间的差距并且解码速度比基线方法快1 930.12倍。此外，Liu等[72]提出一种带有问答的提示学习NER方法，将NER问题转换成问答任务。该方法在低资源的场景下具有更高的性能和更强的鲁棒性。总的来说，提示学习在低资源场景的NER任务上得到了初步尝试，未来会有更多复杂的方法来增强提示，并应用于低资源场景下的许多任务中。

综上所述，本文针对基于有监督深度学习、基于远程监督深度学习、基于Transformer和基于提示学习的四种方法进行了比较分析，具体如表6所示。此外，本文还总结了一些经典的基于深度学习的NER模型，具体如表7所示。

3 NER的研究趋势

目前，NER技术日渐成熟，但依然需要研究人员投入大量精力进行不断探索，通过对现有NER研究工作进行总结，在以后的研究中可以从下面几个方面展开相关的研究。

（1）多任务联合学习。传统的pipeline模型有一定的局限性，例如，NER任务中的实体标注错误，将会进一步导致后续任务的标注错误；同时，多个任务之间会有一定信息共享，但是pipeline模型是无法利用这些潜在的信息的。多任务的联合学习，能解决

表6 基于深度学习的NER方法比较

Table 6 Comparison of NER methods for deep learning

方法	代表模型	优点	缺点
有监督深度学习方法	CNN	数据处理并行化，对高维数据处理无压力	无法很好地提取序列信息
	RNN	解决了CNN无法记忆上下文信息的问题	会出现梯度消失和梯度爆炸现象
	LSTM	引入了输入门、输出门和遗忘门，解决了RNN长期依赖的问题，有效地学习长期依赖信息	梯度问题未完全解决
	GNN	利用图数据结构的数据处理模型，可以更高效地挖掘实体之间的联系	模型灵活性和扩展性差
远程监督深度学习方法	—	一定程度解决了需要大规模已标注数据问题	会产生不完全标注和噪音标注
Transformer方法	BERT类	采用掩码语言模型对双向的Transformer进行预训练，以生成深层的双向语言表征	需要大量GPU和训练数据
提示学习的方法	—	重新定义语言模型的输入，减少预训练模型和下游任务的差距，在低资源场景性能良好	计算复杂度高，需要人工设计提示

表7 基于深度学习的主流NER模型总结

Table 7 Summary of mainstream NER models for deep learning

方法	时间	支持语言	数据集	F1值/%	方法关键字
Collobert等[39]	2011	英文	RCV1	89.59	CNN-CRF、Gazetteer
Huang等[46]	2015	英文	CoNLL2003	90.10	Bi-LSTM-CRF
Strubell等[41]	2017	英文	CoNLL2003、OntoNotes5.0	90.65、86.84	ID-CNNs、RUG
Cetoli等[53]	2018	英文	OntoNotes5.0	83.60	Bi-LSTM-GCN-CRF
Gregoric等[47]	2018	英文	CoNLL2003	91.48	Multiple independent bidirectional LSTM units、Softmax
Zhang等[50]	2018	中文	OntoNotes4.0、MSRA、Weibo NER、Chinese resume	73.88、93.18、58.79、94.46	Lattice-LSTM
Yang等[56]	2020	中文	E-commerce-NER、NEWS NER	61.45、79.22	Distantly supervised、Bi-LSTM-MLPCRF
Gui等[45]	2019	中文	OntoNotes4.0、MSRA、Weibo NER、Chinese resume	74.45、93.71、59.92、95.11	Lexicon rethinking、CNN
Liu等[51]	2019	中文	OntoNotes4.0、MSRA、Weibo NER、Chinese resume	74.43、93.74、59.84、95.21	WC-LSTM
Ding等[54]	2019	中文	OntoNotes4.0、MSRA、Weibo NER、E-commerce-NER	76.00、94.40、59.50、75.20	GGNN-LSTM-CRF
Gui等[55]	2019	中文	OntoNotes4.0、MSRA、Weibo NER、Chinese resume	74.89、93.46、60.21、95.37	Lexicon-based、GCN
Wu等[43]	2019	中文	Bakeoff-3、Bakeoff-4	89.42、90.18	CNN-LSTM-CRF、Joint training
Peng等[57]	2019	英文	CoNLL2003、CoNLL2002、Twitter	82.94、75.85、59.36	Positive-unlabeled learning、Distantly supervised
Tang等[56]	2020	中文	OntoNotes4.0、MSRA、Weibo NER、Chinese resume	75.87、94.40、63.63、95.53	Word-character representation、GCN
Ma等[52]	2020	中文	OntoNotes4.0、MSRA、Weibo NER、Chinese resume	82.81、95.42、70.50、96.11	SoftLexicon、Bi-LSTM
李妮等[64]	2020	中文	MSRA	94.42	BERT-IDCNN-CRF
Li等[65]	2020	中文	CCKS2017、CCKS2018	91.60、89.56	BERT-Bi-LSTM-CRF
Kong等[44]	2021	中文	CCKS2017、CCKS2019	90.49、85.13	Multi-Level CNN、Attention Mechanism
Li等[48]	2021	英文	CoNLL2003	92.53	Modularized interaction network、RNN-BiLSTM-CRF
Xu等[49]	2021	英文	ACE2004、ACE2005、GENIA	86.30、85.40、79.60	Multi-head self-attention、BERT
Naseem等[62]	2021	英文	NCBI Disease、BC5CDR	97.18、97.78	BioALBERT
Yang等[63]	2021	英文	ACE2004、ACE2005、GENIA	87.88、87.04、79.08	Hierarchical transformer
Wu等[66]	2021	中文	CCKS2017、CCKS2019	93.26、82.87	RoBERTa、Radical-level feature
Cui等[69]	2021	英文	CoNLL2003	92.55	BART、Multi-template
Chen等[70]	2021	英文	CoNLL2003	93.90	Prompt-guided attention
Ma等[71]	2021	英文	CoNLL2003、OntoNotes5.0	74.80、72.99	Template-free、Prompt tuning

pipeline模型局限，使得多任务学习之间相互影响，提高学习的性能，利用这种方法来进一步研究NER仍是未来的一个研究热点。

（2）基于提示学习的低资源NER研究。在近些年的研究中，NER任务在广度上已经延伸到跨领域、跨任务和跨语言等任务中。在一般领域，大多数最先进的NER模型需要依赖大量已标记数据进行训练，这使得它们难以扩展到新的、资源较少的语言中。随着提示学习在低资源NER任务上的成功应用[69-72]，这种方法能在低资源和高资源之间架起桥梁，从而实现知识转移。因此，探索更优的提示学习方法来提升低资源的NER模型性能是该领域的重要研究方向。

（3）中文嵌套NER的研究。由于中文构词规则，中文信息文本中的实体嵌套更为明显，此外中文词语没有明显的边界，使得中文的嵌套NER具有一定挑战。近年来，随着深度学习的发展，中文嵌套NER方法出现新思路，如金彦亮等[73]提出一种基于分层标注的中文嵌套NER的方法，能充分捕捉嵌套实体之前的边界信息，有效地提高中文嵌套NER的效果。因此，将各种神经网络、BERT、注意力机制等方法融合用于中文嵌套NER仍然值得研究。

参考文献：

[1] GRISHMAN R, SUNDHEIM B. Message understanding conference-6: a brief history[C]//Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Aug 5-9, 1996. Stroudsburg: ACL, 1996: 466-471.

[2] MERCHANT R, OKUROWSKI M E, CHINCHOR N. The multilingual entity task (MET) overview[C]//Proceedings of the Tipster Text Program Phase II, Vienna, May 6-8, 1996. Stroudsburg: ACL, 1996: 445-447.

[3] SANG E F T K, DE MEULDER F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition[C]//Proceedings of the 7th Conference on Natural Language Learning, Held in Cooperation with HLT-NAACL 2003, Edmonton, May 31- Jun 1, 2003. Stroudsburg: ACL, 2003: 142-147.

[4] DODDINGTON G R, MITCHELL A, PRZYBOCKI M A, et al. The automatic content extraction (ACE) program - tasks, data, and evaluation[C]//Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, May 26-28, 2004. Stroudsburg: ACL, 2004: 837-840.

[5] 孙镇, 王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术, 2010(6): 42-47.
SUN Z, WANG H L. Overview on the advance of the research on named entity recognition[J]. New Technology of Library and Information Service, 2010(6): 42-47.

[6] LI J, SUN A X, HAN J L, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 50-70.

[7] 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.

[8] 赵山, 罗睿, 蔡志平. 中文命名实体识别综述[J]. 计算机科学与探索, 2022, 16(2): 296-304.
ZHAO S, LUO R, CAI Z P. Survey of Chinese named entity recognition[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(2): 296-304.

[9] RAU L F. Extracting company names from text[C]//Proceedings of the 7th IEEE Conference on Artificial Intelligence Application, Miami, Feb 24, 1991. Washington: IEEE Computer Society, 1991: 29-32.

[10] KRUPKA G R. SRA: description of the SRA system as used for MUC-6[C]//Proceedings of the 6th Conference on Message Understanding, Columbia, Nov 6-8, 1995. Stroudsburg: ACL, 1995: 221-235.

[11] SHAALAN K, RAZA H. NERA: named entity recognition for Arabic[J]. Journal of the American Society for Information Science and Technology, 2009, 60(8): 1652-1663.

[12] 张小衡, 王玲玲. 中文机构名称的识别与分析[J]. 中文信息学报, 1997(4): 22-33.
ZHANG X H, WANG L L. Identification and analysis of chinese organization and institution names[J]. Journal of Chinese Information Processing, 1997(4): 22-33.

[13] 王宁, 葛瑞芳, 苑春法, 等. 中文金融新闻中公司名的识别[J]. 中文信息学报, 2002, 16(2): 1-6.
WANG N, GE R F, YUAN C F, et al. Company name identification in Chinese financial domain[J]. Journal of Chinese Information Processing, 2002, 16(2): 1-6.

[14] BIKEL D M, SCHWARTZ R, WEISCHEDEL R M. An algorithm that learns what's in a name[J]. Machine Learning, 1999, 34: 211-231.

[15] ZHOU G D, SU J. Named entity recognition using an HMM-based chunk tagger[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Jul 6-12, 2002. Stroudsburg: ACL, 2002: 473-480.

[16] 张华平, 刘群. 基于角色标注的中国人名自动识别研究[J]. 计算机学报, 2004, 27(1): 85-91.
ZHANG H P, LIU Q. Automatic recognition of Chinese personal name based on role tagging[J]. Chinese Journal of Computers, 2004, 27(1): 85-91.

[17] 俞鸿魁, 张华平, 刘群, 等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2): 87-94.
YU H K, ZHANG H P, LIU Q, et al. Chinese named entity

identification using cascaded hidden Markov model[J]. Journal on Communications, 2006, 27(2): 87-94.

[18] BORTHWICK A, STERLING J, AGICHTEIN E, et al. NYU: description of the MENE named entity system as used in MUC-7[C]//Proceedings of the 7th Message Understanding Conference, Virginia, Apr 29-May 1, 1998. Stroudsburg: ACL, 1998: 1-7.

[19] BENDER O, OCH F J, NEY H. Maximum entropy models for named entity recognition[C]//Proceedings of the 7th Conference on Natural Language Learning, Edmonton, May 31-Jun 1, 2003. Stroudsburg: ACL, 2003: 148-151.

[20] 周雅倩, 郭以昆, 黄萱菁, 等. 基于最大熵方法的中英文基本名词短语识别[J]. 计算机研究与发展, 2003, 40(3): 440-446.
ZHOU Y Q, GUO Y K, HUANG X J, et al. Chinese and English base NP recognition based on a maximum entropy model[J]. Journal of Computer Research and Development, 2003, 40(3): 440-446.

[21] 张玥杰, 徐智婷, 薛向阳. 融合多特征的最大熵汉语命名实体识别模型[J]. 计算机研究与发展, 2008, 45(6): 1004-1010.
ZHANG Y J, XU Z T, XUE X Y. Fusion of multiple features for Chinese named entity recognition based on maximum entropy model[J]. Journal of Computer Research and Development, 2008, 45(6): 1004-1010.

[22] ISOZAKI H, KAZAWA H. Efficient support vector classifiers for named entity recognition[C]//Proceedings of the 19th International Conference on Computational Linguistics, Taipei, China, Aug 24-Sep 1, 2002. Stroudsburg: ACL, 2002: 1-7.

[23] TAKEUCHI K, COLLIER N. Use of support vector machines in extended named entity recognition[C]//Proceedings of the 6th Conference on Natural Language Learning, Taipei, China, Aug 24-Sep 1, 2002. Stroudsburg: ACL, 2002: 184-190.

[24] 李丽双, 黄德根, 陈春荣, 等. 基于支持向量机的中文文本中地名识别[J]. 大连理工大学学报, 2007, 47(3): 433-438.
LI L S, HUANG D G, CHEN C R, et al. Identification of location names from Chinese texts based on support vector machine[J]. Journal of Dalian University of Technology, 2007, 47(3): 433-438.

[25] 陈霄, 刘慧, 陈玉泉. 基于支持向量机方法的中文组织机构名的识别[J]. 计算机应用研究, 2008, 25(2): 362-364.
CHEN X, LIU H, CHEN Y Q. Chinese organization names recognition based on SVM[J]. Application Research of Computers, 2008, 25(2): 362-364.

[26] MCCALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]//Proceedings of the 7th Conference on Natural Language Learning, Edmonton, May 31-Jun 1, 2003. Stroudsburg: ACL, 2003: 188-191.

[27] KRISHNAN V, MANNING C D. An effective two-stage model for exploiting non-local dependencies in named entity recognition[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Jul 17-21, 2006. Stroudsburg: ACL, 2006: 1121-1128.

[28] 冯元勇, 孙乐, 张大鲲, 等. 基于小规模尾字特征的中文命名实体识别研究[J]. 电子学报, 2008, 36(9): 1833-1838.
FENG Y Y, SUN L, ZHANG D K, et al. Study on the Chinese named entity recognition using small scale character tail hints[J]. Acta Electronica Sinica, 2008, 36(9): 1833-1838.

[29] 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(工学版), 2014, 44(6): 1843-1848.
YAN Y, WEN D W, WANG Y J, et al. Named entity recognition in Chinese medical records based on cascaded conditional random field[J]. Journal of Jilin University(Engineering and Technology Edition), 2014, 44(6): 1843-1848.

[30] TEIXEIRA J, SARMENTO L, OLIVEIRA E C. A bootstrapping approach for training a NER with conditional random fields[C]//LNCS 7026: Proceedings of the 15th Portuguese Conference on Artificial Intelligence, Lisbon, Oct 10-13, 2011. Berlin, Heidelberg: Springer, 2011: 664-678.

[31] THENMALAR S, BALAJI J, GEETHA T. Semi-supervised bootstrapping approach for named entity recognition[J]. arXiv:1511.06833, 2015.

[32] 黄诗琳, 郑小林, 陈德人. 针对产品命名实体识别的半监督学习方法[J]. 北京邮电大学学报, 2013, 36(2): 20-23.
HUANG S L, ZHENG X L, CHEN D R. A semi-supervised learning method for product named entity recognition[J]. Journal of Beijing University of Posts and Telecommunications, 2013, 36(2): 20-23.

[33] LONG L Y, YAN J Z, FANG L Y, et al. The identification of Chinese named entity in the field of medicine based on bootstrapping method[C]//Proceedings of the 2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems, Beijing, Sep 28-29, 2014. Piscataway: IEEE, 2014: 1-6.

[34] 王路路, 艾山·吾买尔, 买合木提·买买提, 等. 基于CRF和半监督学习的维吾尔文命名实体识别[J]. 中文信息学报, 2018, 32(11): 16-26.
WANG L L, AISHAN W, MAIHEMUTI M, et al. A semi-supervised approach to Uyghur named entity recognition based on CRF[J]. Journal of Chinese Information Processing, 2018, 32(11): 16-26.

[35] ETZIONI O, CAFARELLA M, DOWNEY D, et al. Unsupervised named-entity extraction from the web: an experimental study[J]. Artificial Intelligence, 2005, 165(1): 91-134.

[36] NADEAU D, TURNEY P D, MATWIN S. Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity[C]//LNCS 4013: Proceedings of the 19th Conference of the Canadian Society for Computational Stu-

dies of Intelligence, Quebec, Jun 7-9, 2006. Berlin, Heidelberg: Springer, 2006: 266-277.

[37] HAN X, KWOH C K, KIM J J. Clustering based active learning for biomedical named entity recognition[C]//Proceedings of the 2016 International Joint Conference on Neural Networks, Vancouver, Jul 24-29, 2016. Piscataway: IEEE, 2016: 1253-1260.

[38] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Jun 2-7, 2019. Stroudsburg: ACL, 2019: 4171-4186.

[39] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12: 2493-2537.

[40] YAO L, LIU H, LIU Y, et al. Biomedical named entity recognition based on deep neutral network[J]. International Journal of Hybrid Information Technology, 2015, 8(8): 279-288.

[41] 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, Sep 9-11, 2017. Stroudsburg: ACL, 2017: 2670-2680.

[42] WU Y H, JIANG M, LEI J B, et al. Named entity recognition in Chinese clinical text using deep neural network [C]//Proceedings of the 15th World Congress on Health and Biomedical Informatics, São Paulo, Aug 19-23, 2015: 624-628.

[43] WU F Z, LIU J X, WU C H, et al. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation[C]//Proceedings of the World Wide Web Conference, San Francisco, May 13-17, 2019. New York: ACM, 2019: 3342-3348.

[44] KONG J, ZHANG L X, JIANG M, et al. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2021, 116:103737-103743.

[45] GUI T, MA R T, ZHANG Q, et al. CNN-based Chinese NER with lexicon rethinking[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, Aug 10-16, 2019: 4982-4988.

[46] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv:1508.01991, 2015.

[47] GREGORIC A Z, BACHRACH Y, COOPE S. Named entity recognition with parallel recurrent neural networks[C]//Proceedings of the 56th International Conference on Computational Linguistics, Melbourne, Jul 15-20, 2018. Stroudsburg: ACL, 2018: 69-74.

[48] LI F, WANG Z, HUI S C, et al. Modularized interaction network for named entity recognition[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Aug 1-6, 2021. Stroudsburg: ACL, 2021: 200-209.

[49] XU Y X, HUANG H Y, FENG C, et al. A supervised multi-head self-attention network for nested named entity recognition[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, the 33rd Conference on Innovative Applications of Artificial Intelligence, the 11th Symposium on Educational Advances in Artificial Intelligence, Feb 2-9, 2021. Menlo Park: AAAI, 2021: 14185-14193.

[50] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Jul 15-20, 2018. Stroudsburg: ACL, 2018: 1554-1564.

[51] LIU W, XU T G, XU Q H, et al. An encoding strategy based word-character LSTM for Chinese NER[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Jun 2-7, 2019. Stroudsburg: ACL, 2020: 5951-5960.

[52] MA R T, PENG M L, ZHANG Q, et al. Simplify the usage of lexicon in Chinese NER[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul 5-10, 2020. Stroudsburg: ACL, 2019: 2379-2389.

[53] CETOLI A, BRAGAGLIA S, O'HARNEY A D, et al. Graph convolutional networks for named entity recognition[C]//Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories, Prague, Jan 23-24, 2018. Stroudsburg: ACL, 2018: 37-45.

[54] DING R X, XIE P J, ZHANG X Y, et al. A neural multi-digraph model for Chinese NER with gazetteers[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, Jul 28-Aug 2, 2019. Stroudsburg: ACL, 2019: 1462-1467.

[55] GUI T, ZOU Y C, ZHANG Q, et al. A lexicon-based graph neural network for Chinese NER[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hongkong, China, Nov 3-7, 2019. Stroudsburg: ACL, 2019: 1040-1050.

[56] TANG Z, WAN B, YANG L. Word-character graph convolution network for Chinese named entity recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1520-1532.

[57] PENG M L, XING X Y, ZHANG Q, et al. Distantly supervised named entity recognition using positive-unlabeled learning[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, Jul 28-Aug 2, 2019. Stroudsburg: ACL, 2019: 2409-2419.

[58] YANG Y S, CHEN W L, LI Z H, et al. Distantly supervised

NER with partial annotation learning and reinforcement learning[C]//Proceedings of the 27th International Conference on Computational Linguistics, Florence, Jul 28-Aug 2, 2019. Stroudsburg: ACL, 2019: 2159-2169.

[59] ZHANG H L, LIU L T, CHENG S Z, et al. Distant supervision for Chinese temporal tagging[C]//Proceedings of the 3rd China Conference on Knowledge Graph and Semantic Computing, Tianjin, Aug 14-17, 2018. Cham: Springer, 2018: 14-27.

[60] 边俐菁. 基于深度学习和远程监督的产品实体识别及其领域迁移研究[D]. 上海: 上海财经大学, 2020.
BIAN L J. Research on product entity recognition and domain migration based on deep learning and remote supervision[D]. Shanghai: Shanghai University of Finance and Economics, 2020.

[61] SOUZA F, NOGUEIRA R, LOTUFO R. Portuguese named entity recognition using BERT-CRF[J]. arXiv:1909.10649, 2019.

[62] NASEEM U, KHUSHI M, REDDY V, et al. BioALBERT: a simple and effective pre-trained language model for biomedical named entity recognition[C]//Proceedings of the 2021 International Joint Conference on Neural Networks, Shenzhen, Jul 18-22, 2021. Piscataway: IEEE, 2021: 1-7.

[63] YANG Z W, MA J, CHEN H C, et al. HiTRANS: a hierarchical transformer network for nested named entity recognition[C]//Proceedings of the Findings of the Association for Computational Linguistics, Punta Cana, Nov 16-20, 2021. Stroudsburg: ACL, 2021: 124-132.

[64] 李妮, 关焕梅, 杨飘, 等. 基于BERT-IDCNN-CRF的中文命名实体识别方法[J]. 山东大学学报(理学版), 2020, 55(1): 102-109.
LI N, GUAN H M, YANG P, et al. BERT-IDCNN-CRF for named entity recognition in Chinese[J]. Journal of Shandong University (Natural Science), 2020, 55(1): 102-109.

[65] 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: 103422-103428.

[66] WU Y, HUANG J, XU C, et al. Research on named entity recognition of electronic medical records based on RoBERTa and radical-level feature[J]. Wireless Communications and Mobile Computing, 2021: 2489754.

[67] YAO L G, HUANG H S, WANG K W, et al. Fine-grained mechanical Chinese named entity recognition based on ALBERT-AttBiLSTM-CRF and transfer learning[J]. Symmetry, 2020, 12(12): 1986-2006.

[68] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems 33, Dec 6-12, 2020: 1877-1901.

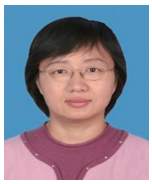
[69] CUI L Y, WU Y, LIU J, et al. Template based named entity recognition using BART[C]//Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Bangkok, Aug 1-6, 2021. Stroudsburg: ACL, 2021: 1835-1845.

[70] CHEN X, ZHANG N, LI L, et al. Lightner: a lightweight generative framework with prompt-guided attention for low-resource NER[J]. arXiv:2109.00720, 2021.

[71] MA R, ZHOU X, GUI T, et al. Template-free prompt tuning for few-shot NER[J]. arXiv:2109.13532, 2021.

[72] LIU A T, XIAO W, ZHU H, et al. QaNER: prompting question answering models for few-shot named entity recognition[J]. arXiv:2203.01543, 2022.

[73] 金彦亮, 谢晋飞, 吴迪嘉. 基于分层标注的中文嵌套命名实体识别[J]. 上海大学学报(自然科学版), 2021, 27(3): 1-9.
JIN Y L, XIE J F, WU D J. Chinese nested named entity recognition based on hierarchical tagging[J]. Journal of Shanghai University (Natural Science Edition), 2021, 27(3): 1-9.



李冬梅(1972—),女,博士,教授,主要研究方向为自然语言处理、知识图谱。
LI Dongmei, born in 1972, Ph.D., professor. Her research interests include natural language processing and knowledge graph.

罗斯斯(1993—),女,硕士研究生,主要研究方向为自然语言处理、知识图谱。
LUO Sisi, born in 1993, M.S. candidate. Her research interests include natural language processing and knowledge graph.

张小平(1969—),女,博士,正高级工程师,主要研究方向为数据挖掘、人工智能。
ZHANG Xiaoping, born in 1969, Ph.D., professorate senior engineer. Her research interests include data mining and artificial intelligence.

许福(1979—),男,博士,教授,主要研究方向为遥感信息处理、智慧林业、智慧园林。
XU Fu, born in 1979, Ph.D., professor. His research interests include remote sensing information processing, smart forestry and smart garden.