计算机科学
COMPUTER SCIENCE

命名实体识别任务综述

高翔，王石，朱俊武，梁明轩，李阳，焦志翔

引用本文

高翔，王石，朱俊武，梁明轩，李阳，焦志翔命名实体识别任务综述[J]．计算机科学, 2023, 50(6A): 220200119-8.

GAO Xiang, WANG Shi, ZHU Junwu, LIANG Mingxuan, LI Yang, JIAO Zhixiang. Overview of Named Entity Recognition Tasks [J]. Computer Science, 2023, 50(6A): 220200119-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220200119

命名实体识别任务综述

高 翔[1,2] 王 石[2] 朱俊武[1] 梁明轩[1,2] 李 阳[1,2] 焦志翔[1,2]
1 扬州大学信息工程学院 江苏 扬州 225000
2 中国科学院计算技术研究所 北京 100190
(SteveGao66@163.com)

摘 要 命名实体识别作为自然语言处理中一项十分基础的任务,为其他许多下游任务的高效完成奠定了基础。其目的是从一段用自然语言描述的文本中识别出相应的实体并标注其类型,以此为其他相关任务作出数据标注的准备。首先介绍了命名实体识别任务的发展历程以及在对应背景下相关研究用到的重点方法,包括自诞生初期用到的基于规则和字典的方法以及后期发展衍生出的基于统计学、深度学习的方法。其次总结了一些该领域比较主流的研究方向,包括低资源条件下的命名实体识别、嵌套命名实体识别以及跨语言的命名实体识别等,这些方向都是近期该任务的热门研究趋势,包含了该任务目前最为流行的研究方法。最后总结了研究中的相关经验,展望了该任务未来的发展方向及难点。

关键词:命名实体识别;嵌套命名实体识别;深度学习;低资源;跨语言
中图法分类号 TP391

Overview of Named Entity Recognition Tasks

GAO Xiang[1,2],WANG Shi[2],ZHU Junwu[1],LIANG Mingxuan[1,2],LI Yang[1,2] and JIAO Zhixiang[1,2]
1 College of Information Engineering,Yangzhou University,Yangzhou,Jiangsu 225000,China
2 Institute of Computing Technology,Chinese Academy of Sciences,Beijing 100190,China

Abstract Named entity recognition,as a very basic task in natural language processing,lays the foundation for the efficient completion of many other downstream tasks. Its purpose is to identify the corresponding entity from a text described in natural language and label its type,so as to make preparations for data labeling for other related tasks. This paper first introduces the development process of named entity recognition tasks and the key methods used in related research in the corresponding context,including the rule-based and dictionary-based methods used in the early days of the birth,and the statistics and deep learning derived from the later development. Secondly,it summarizes some of the more mainstream research directions in this field,including named entity recognition under low-resource conditions,nested named entity recognition,and cross-language named entity recognition. These directions are the hot research trends of this task recently,including the most popular research method of this task at present. Finally,the relevant experience in the research is summarized,and the future development direction and difficulties of the task are prospected.

Keywords Named entity recognition,Nested named entity recognition,Deep learning,Low-resource,Cross-language

1 引言

如今,如何高效地处理文本中包含的信息作为一个困难却有意义的问题,引起了学术界以及工业界的广泛研究。在这其中,命名实体识别(Named Entity Recognition,NER)任务作为一个基础性的研究课题,相关研究方法近来年已层出不穷。该任务属于自然语言处理(Natural Language Processing,NLP)领域重要的基础任务之一,也是人工智能领域的一项核心技术,对句法分析[1]、文本分类[2]、文本相似度计算[3]、语言模型(例如 Baidu ernie)、机器翻译[4]、自动问答[5]、知识图谱[6]等众多 NLP 下游任务均具有重要的支撑作用。因此,该任务在 NLP 领域中具有良好的研究意义。

目前,该任务普遍被建模为一个序列标注模型,通过序列标注的方法辨别文本中的命名实体并其对应的类型按照各种标注方法给予标注。此任务由 Rau 等[7]在 1991 年首次提出,随后在信息抽取等领域得到了广泛应用,并且作为一个序列标注子任务被应用到测评任务中。

具体到各种语言而言,由于各种语言文字构造的差异,

基金项目:国家自然科学基金(61702234);国家 242 信息安全计划项目(2021A008);北京市科技新星计划交叉学科合作课题(Z191100001119014);国家重点研发计划重点专项(2017YFC1700300,2017YFB1002300)
This work was supported by the National Natural Science Foundation of China(61702234),National 242 Information Security Program(2021A008),Beijing NOVA Program(Z191100001119014),National Key Research and Development Program of China(2017YFB1002300,2017YFC1700300).
通信作者:王石(wangshi@ict.ac.cn)

命名实体识别任务的难度也会有所不同。比如，对于中文而言，由于中文数据集中各个汉字之间的排列非常紧密，没有空格，因此中文领域命名实体识别的难度大大增加；而英文则恰恰相反，由于其单词构造的特殊性，每个单词之间存在空格，因此对于命名实体识别任务而言，其难度比较中文简单[8]。除此之外，对于其他的小语种领域，虽然用来做训练的数据集可能很少，但也有大批学者做出了相应的研究[9-10]，这其中涉及到的低资源条件下的 NER 方法将在后文中详细介绍。总体而言，不同语种的命名实体识别任务的主要区别在于对不同语言做出模型上的调整，技术理念和手段大致相似。

本文第 2 节主要简单介绍命名实体识别任务的相关概念和研究意义与难点；第 3 节详细介绍其从诞生到如今的发展历程与研究方法，其中基于条件随机场的方法在该任务中应用最为广泛；第 4 节讨论近期命名实体识别任务的趋势，包括应用注意力机制和迁移学习的方法、嵌套命名实体识别、低资源下的命名实体识别、跨语言命名实体识别等，同时介绍在这些方向上一些相关学者展开的研究和具体方法；第 5 节简单介绍一些数据集、标注方法及评价指标。

2 命名实体识别任务

2.1 定义与目标

命名实体识别任务的定义从根本上说就是把相关实体从一段用自然语言表述的文本中找出，接着标注出实体的相关信息，包括类型和位置[11]。一般来说，我们通常将这种任务归类为序列标注问题中的一种。而它的目标主要是尽可能高效并准确地识别出文本中实体的类型，例如人名(PER)、地名(LOC)、机构名(ORG)等[12]。

2.2 意义及难点

如今，对 NLP 的研究进入高速发展时期，伴随着文本挖掘、关系抽取等基本任务的广泛应用，文本语义知识变得愈发重要。新兴的研究领域，如语义分析、自动问答、意见挖掘等，均需要丰富的语义知识作为支撑。而实体作为文本中重要的语义知识，对它的识别和分类已成为一项重要的基础性研究问题。计算机科学中的机器学习、计算语言学中的语义分析、图书情报中的本体构建等领域，都对该问题进行了广泛的研究。目前，NER 任务在 NLP 领域的许多场景有着广泛应用，比如构建领域知识库[13]、机器翻译、情感分析[14]等。然而由于命名实体本身的随意性、复杂性、多变性等特点，该问题还远没有达到可以完全解决的地步。纵观命名实体识别任务高速发展的这些年，与时俱进的技术与层出不穷的讨论将该研究问题不断推新，这使得虽然自提出至今已有二十多年，该项任务仍然是一个重要但具有挑战性的研究课题。

3 研究进展

命名实体识别任务自诞生以来所用到的研究方法与时俱进，最初被广泛应用的是基于词典和规则的方法，该方法具有一定的局限性；伴随着机器学习的发展，基于统计学的方法进入大众视野。近年来，深度学习的加入使得该任务发展到一个新阶段，其发展历程大致如图 1 所示。

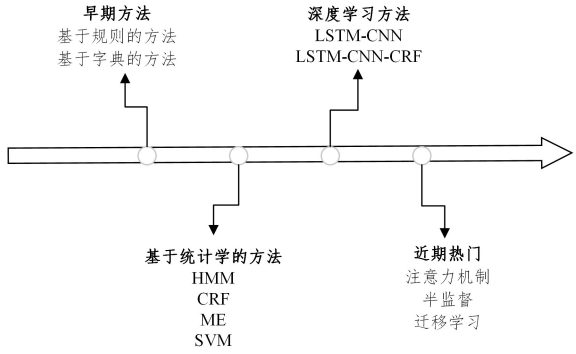

图 1 命名实体识别技术研究进展

Fig. 1 Research progress of named entity recognition technology

3.1 基于规则和字典的 NER

基于规则的和字典方法的 NER 主要是利用字符串和模式相匹配。当所选用的规则将文本信息较好地反映出来时，该方法通常可以展现出不错的效果。但同时，这种方法也存在着比较明显的不足，例如所选用的规则通常依赖于特定的语言和领域，不易覆盖到所有的语言现象，系统的可移植性较差等。

Petasis 等[15]针对该方法的不足，提出了一种可以较好维护基于规则命名实体识别和分类系统的方法。其中的基本思想是使用一个机器学习构建的独立系统提高基于规则系统的性能。另外一个系统通过基于规则系统生成的数据进行训练。两个系统产生的分歧作为更新基于规则系统的信号。

Chiticariu 等[16]为了研究基于规则的方法是否仍然在 NER 上可行，在通用代数提取系统 SystemT 上设计了一种名为 NERL 的高级语言。当应用到 NER 任务时，该语言针可以进行微调。实验结果表明这些定制的注释器匹配可能优于使用机器学习技术实现的最佳结果。

3.2 基于统计学方法

该类的方法主要包括：隐马尔可夫模型(Hidden Markov Model，HMM)、条件随机场(Conditional Random Field，CRF)、最大熵(The Maximum Entropy Principle，ME)等。

3.2.1 基于条件随机场的 NER

用条件随机场做命名实体识别任务具有比较明显的优点，比如全局最优等。但与此同时，它也有着明显的缺点，比如训练时间较长，局部标签依赖，低效的维特比解码速度等。

McCallum 和 Li[17]二人最先在命名实体识别任务中加入了 CRF，他们通过大量词汇测试，将具有特征归纳的 CRF 应用于印地语命名实体识别任务并取得了不错的效果，缺点是训练时间过长。

在此之后，越来越多的学者开始采用基于 CRF 的方法进行 NER 任务的研究。Konkol 等[18]基于 CRF 构造了一个捷克语命名实体识别系统，该系统被构造为输出平面或分层命名实体，从而可以使用捷克语的所有已知系统进行评估。

Xu 等[19]在条件随机场的基础上建立了中文分词、命名实体识别和词性标注 3 个系统。该系统采用了基本特征以及大量语言特征。对于分割任务，根据每个字符的置信度调整 BIO 标签。

Lin 等[20]提出了一种具有多通道信息和条件随机场的综合词表示结合到传统 BiLSTM 中的方法，该方法不是单独

使用原始的预训练词嵌入作为最终的词表示,而是为输入句子中的每个词构建一个全面的词表示。这种全面的词表示包含字符级子词信息、原始预训练词嵌入和多个句法特征。通过 LSTM 层,使得每个单词都有一个隐藏状态。隐藏状态被最终的 CRF 层视为单词的特征向量,可以从中解码输入句子的最终预测标签序列。

3.2.2　基于最大熵模型的 NER

最大熵模型的通俗理解就是按照模型熵最大的原则来挑选出好的模型。此模型的优点是结构紧凑,适用范围广;缺点是训练时所需要的时间复杂度很高。

Mikheev 等[21]最先将最大熵模型应用于 NER 任务中,该系统结合了基于规则的语法与最大熵模型,分为 Sure-fire Rules,Partial Match 1,Rule Relaxation,Partial Match 2 和 Title Assignment 5 个步骤。

在随后的研究中,不少学者利用改进的最大熵模型进行 NER 任务的研究。Borthwick 等[22]围绕最大实体框架构造了一种新型识别系统,通过在最大熵理论框架内工作并利用灵活的基于对象的体系结构,使得其能够在做出标记决策时将多样化的知识源考虑入内。纯粹的统计系统不包含手工生成的模式,并获得与最佳统计系统相当的结果。当与其他手工编码系统结合时,该系统获得的分数超过了当时其他系统的最高可比分数。

Tsai 等[23]通过结合基于规则和机器学习的方法构造了一个被称为 Mencius 的中文命名实体识别器。该系统构造了一个被称为 InfoMap 的模板匹配工具进入最大熵框架。InfoMap 将位置或组织名称表示为模板。输入的字符串首先通过 InfoMap 与一个或多个位置或组织模板匹配,然后传递给 Mencius,这里它被分配了特征值以进一步区分它属于哪个命名实体类别。

3.2.3　基于隐马尔可夫模型的 NER

隐马尔可夫模型是一个概率生成模型,具有严格的时间顺序[24],由一个不可观测的状态序列和一个可观测的观测序列组成。

具体到 NER 任务的应用而言,Bikel 等[25]是首个应用 HMM 模型的学者,其提出了一种使用标准隐马尔可夫模型的变体在文本中查找名称和其他非递归实体的统计学习方法。模型中有一个遍历 HMM,包含 8 个内部状态(名称类,包括 NOT-A-NAME 类)和 2 个特殊状态(START-和 END-OF-SENTENCE 状态)。

Li 等[26]提出了一种条件隐马尔可夫模型(Conditional Hidden Markov Model,CHMM)。在此前常见的马尔可夫模型上,该模型利用通过预训练的语言模型具有较强的上下文表示能力这一特征,增强了能力。CHMM 是用于多源标签去噪的 HMM 变体。它将真实实体标签建模为隐藏变量,并从观察到的噪声标签中推断出它们;同时还使用替代训练方法进一步完善 CHMM。

Liu 等[27]提出了一种基于分层隐马尔可夫模型的中文自由文本产品——NER 方法。该模型由内部状态 IS、生产状态 PS 和每个级别的结束状态 ES 3 个级别组成。在统一的统计框架内,该方法能够通过利用各种语言特征和知识源,为全局优化做出相对合理的概率决策。

3.3　基于深度学习的 NER 方法

近年来,基于神经网络而衍生出的深度学习被广泛应用于各种各样的场景。这种方法不再像之前的方法那样需要大量的手工特征工程以及领域专家提供的领域知识。

使用深度学习来处理命名实体识别任务一般采用 DL-CRF 模型,将之前的深层神经网络接入 CRF 层,从而能够预测出句子级别的标签。其中,DL 包括 LSTM,CNN 和 RNN 等。比如,Peng 和 Mark[28]提出了一种将 LSTM 和 CRF 相结合的模型联合训练命名实体识别和分词任务,该工作比之前发布的相关结果提高了近 5% 的 F 值。Liu 等[29]通过在混合半马尔可夫 CRF 架构上添加了一个简单的模块,将分段神经命名实体识别的性能大大提高。Lample 等[30]使用一种 BiLSTM-CRF 的结构和另一种基于转换的结构,依赖于从未注释的语料库中学习到的无监督单词表示,在 4 种语言的命名实体识别任务中展现出了优异的效果。此外,基于卷积神经网络[31]和混合神经网络[32]的方法在命名实体识别任务上的应用也非常广泛,并且都展示出了比较好的效果。

4　近期 NER 方法的趋势

随着技术的发展,NER 任务展现出新的发展趋势,包括加入注意力机制和迁移学习、低资源环境下的 NER、嵌套 NER、跨语言 NER 等。

4.1　注意力机制在 NER 中的应用

注意力(Attention)机制最早出现在机器翻译中。之后,随着 Google mind 将 Attention 机制应用于循环神经网络中并以此来解决图像分类的问题[33],Attention 机制在命名实体识别任务中得到了普遍应用。用该机制处理 NER 任务可以有效捕获完全由字符向量序列携带的上下文信息,大大提高了 NER 任务的性能。

Bahdanau 等[34]首次将 Attention 机制用于 NER 任务并取得了优异的效果。随后,越来越多的学者开始使用该方法处理该任务。

Liao 等[35]通过引入多头自注意力机制解决了军事命名实体识别需要大量专业领域知识的问题。该模型中的注意力层首先查询所有键的点积以获得字符向量间的相似度分数,然后获得更稳定的梯度并利用 softmax 函数计算权重系数,最后通过加权求和得到注意力的输出。计算过程如式(1)所示:

$$\begin{cases} head_i = \text{attention}(HW_i^Q, HW_i^K, HW_i^V) \\ multi\_head(Q,K,V) = \text{contact}(head_1, head_2, \cdots, head_i) \end{cases} \tag{1}$$

其中,W_i 是可训练的参数,*multi_head* 表示对 3 个向量进行多头注意力计算,*contact* 表示对各个头注意力进行拼接计算。

Wu 等[36]在多任务神经模型中加入了注意力和多任务机制,在 NER 任务中取得了优异的性能。该模型中注意力层从句子中提取多个信息组件,其将注意力序列和 BiLSTM 输出序列连接起来作为 CRF 层的输入,通过该层生成实体标签序列;在分类任务中使用了结构化的自注意力方法;使用的注意力层与 NER 任务中的相同,并且将多跳注意力计算得到的标注矩阵与 BiLSTM 输出相乘以生成用于分类的句子嵌入。

4.2 应用迁移学习方法

迁移学习(Transfer Learning)的主要思想是将相关领域所学到的知识结构迁移到目标领域,以达到目标领域的改进或者任务学习的效果[37]。Lin 等[38]利用迁移学习的方法提出了一种多语言多任务架构,开发了一种具有少量标注数据的多语言多任务模型,该模型在命名实体识别任务上相对于基线模型获得了 4.3%～50.5%的 F 值收益。

Qu 等[39]使用迁移学习来学习领域特定的 NER 模型,提出 TransInit。该方法包括 3 个步骤:1)将线性 CRF 层通过比较大的源领域预料库上进行训练;2)为了学习到源域类型和目标域类型之间的相关性,使用具有两层的神经网络进行训练学习;3)通过神经网络训练目标域命名实体类型的 CRF 层。之前迁移学习的研究中只假设域不匹配,在该模型中作者还加入了假设标签不匹配这一规则,这也是文章的创新所在。

Francis 等[40]通过实验证明了在来自源域的标记数据上训练的 NER 模型可以用作基础模型,然后使用少量标记数据进行微调,以识别目标域中的不同命名实体类。其使用从先前训练的模型中学习到的特征或权重形式的知识来为相关的目标任务训练更新的模型。

总的来说,目前的研究中有两种不同的迁移学习策略。第一种是使用预训练模型作为特征提取器,这种方法可以利用预训练网格来提取其他任务的特征;另一种方法是微调预训练模型,该策略适用于相同的标签集和不同标签集的传输。

4.3 低资源环境下的 NER

4.3.1 定义

所谓的低资源环境,指的是训练数据较少的环境。就 NER 任务而言,用来做英语 NER 任务的训练数据比较充足。然而,对于一些小语种而言,可以用来做 NER 任务的数据可能极其缺乏,这使得模型对于一些隐藏特征无法充分学习,导致学习方法的效率降低。因此,如何用较少的资源来训练神经命名实体识别模型,成为了需要解决却又具有一定挑战性的问题。

4.3.2 方法

目前,很多学者针对该问题提出了解决方法。主流方法是使用迁移学习来连接高资源和低资源的数据集。其中一种是通过领域自适应方法来实现。例如,Kruengkrai 等[41]通过考虑桥接 Ruder 等[42]的多任务学习(Multi-Task-Learning,MTL)和 Devlin 等[43]的预训练模型来改进低资源环境下的命名实体识别。与之前的多任务学习不同的是,在这篇文章中的工作目标是通过使用辅助任务(句子分类)创建预训练表示和正则化器来提高主要任务(NER)的性能。主要方法是使用在神经网络的底层共享 Word embedding 和 BiLSTM,Attention 机制和隐藏层实现不同语言中的共享,然后在顶层对不同的语言使用特定于任务的不同 CRF 层和线性层,以此实现低资源环境的 NER 任务。Yu 等[44]提出了一种新颖的鲁棒性和自适应方法 RDANER,该方法只使用廉价且容易获得的资源,在 3 个数据集上实现了最佳性能。该方法中的 RDANER 由两个过程组成:LM 微调和 bootstrapping。

还有的学者通过数据增强的方法来解决此类问题。Yaseen 和 Langer[45]采用反向翻译为低资源命名实体识别生成高质量和语言多样化的合成数据,对来自材料科学(MaSciP)和生物医学领域(S800)的两个数据集进行了实验,证明了所提出方法在资源极度匮乏的情况下对 NER 任务的有效性。Zhu 等[46]提出了一种基于预训练语言模型(Pre-Trained-Language-Model,PLM)和课程学习策略的新型数据增强方法,具体来说,使用 PLM 通过预测不同的掩码来生成不同的训练实例,并设计特定于任务的课程学习策略以减轻噪声的影响。在 3 个数据集上的测试表明,该模型分别比之前的基线模型提高了 3.46%,2.58%和 0.99%的 F1 分值。

总的来说,对于低资源环境下的命名实体识别问题,最近的研究提出了一些较为有效的方法,例如基于迁移学习、对抗性领域自适应方法、远程监督学习和数据增强方法等被广泛应用于此任务中。这些方法在很大程度上降低了对人工标注数据量的要求,降低了成本。

4.4 嵌套命名实体识别

4.4.1 相关背景

一般情况下,在命名实体识别任务中,处理的对象往往是一些非嵌套的实体。然而,嵌套的实体在 NER 任务中也随处可见。比如,在"北京大学是一所世界高水平大学"这句话中,"北京大学"是一个组织机构名,同时,"北京"又是一个地名,如图 2 所示。

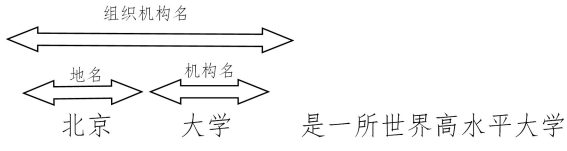

图 2 嵌套的实体实例

Fig. 2 Nested entity instance

如果一个模型可以准确地将"北京大学"标记成一个机构名称,并在此前提下同时可以将"北京"标记成一个地名,那么每次在碰到相同的文字结构时,就有理由相信该模型可以将[地点][机构名]的排列方式准确标记出来,这样就达到了提高模型准确率的效果。

4.4.2 具体方法

已经有大批学者在嵌套命名实体识别任务上展开了研究,也提出了许多具体的方法。Wang 和 Li 等[47]提出了一个名为 HIT 的新型嵌套 NER 模型,包括一个显式边界标记和一个边界内标记之间的紧密内部连接,在 NER 的 3 个数据集上实现了先进的性能。Wang 等[48]提出了一种名为 Pyramid 的新型分层模型和一个逆向 Pyramid 以允许层之间双向交互,此方法也可用于重叠命名实体识别的更一般任务。Long 等[49]考虑了子枚举序列的高复杂性使得领先的基于区域的模型面临效率和有效性挑战的问题,提出了一种分层区域学习框架,以自动生成具有近乎线性复杂性的候选区域的树层次结构,并将结构信息合并到区域表示中,以进行更好的分类。在基准数据集 ACE-2005,GENIA 和 JNLPBA 上的实验证明了,与目前最先进的基线模型相比,此方法具有同等甚至更好的结果。

然而,目前的大部分方法忽略了不同实体类型下单词之间的语义相关性。对于此问题,Xu 等[50]考虑到句子中的单词在不同实体类型下表达的含义有所不同,将 NER 任务视为一个对词对进行多类分类的问题,并由此设计了一个具有

多头注意力机制的有监督神经网络模型。在该模型中,每个注意力头代表一种具体的实体类型。该模型可以通过具体类型下头尾两个部分的相关强度来预测跨度类型。此外,作者还将多任务学习框架应用于此,由此将实体边界检测和分类进行融合,对于这两个任务之间的依赖性,该模型也可以进行捕获。Fu 等[51]将嵌套 NER 任务视为部分观察树,通过这个观察树进行选区解析,同时使用 TreeCRF 对该任务建模;并提出了 Masked Inside 算法,加快了训练和推理速度。

总的来说,对于嵌套命名实体识别任务,目前的方法包括基于状态转换的方法、基于超图的方法、基于阅读理解的方法等。它是一种特殊的命名实体形式,在 NER 任务中应用广泛。此任务充分利用了外部与内部命名实体的嵌套信息,对于 NER 任务而言具有良好的研究意义。

4.5　跨语言命名实体识别

4.5.1　相关背景

命名实体识别任务本质上是多语言的,任何给定语言的注释都可能受到限制,这使得学者们开始考虑多语言命名实体识别任务。其中一种简单的想法是用从一种以上语言中提取的带注释数据训练一个模型。但是这种方法并不普遍适用。因为这种方法尽管可以访问更多的训练数据,但使用从多种语言中提取的带注释数据对 NER 模型的简单训练始终不如仅在单语言数据上训练的模型。

4.5.2　具体方法

对于此领域的问题,已经有大批学者展开了相关研究。Mueller 等[52]探索了多语言命名实体识别模型中多语言迁移的来源,并检查了多语言模型与其单语模型相比的权重结构。通过探索发现多语言模型有效地跨语言共享许多参数,并且微调可能会利用大量的这些参数。Lin 等[38]设计了一个使用两层参数共享组合各种传输模型的多语言多任务架构,针对不同的传输方案采取不同的参数共享策略。Xie 等[53]考虑到了不同语言单词和词序的差异问题,提出了一种基于双语词嵌入查找翻译的方法来改进跨语言的词汇项映射;并且使用自注意力机制提高了词序差异的鲁棒性。此外,Yu 等[54]从不同的思路展开研究,通过对在英文命名实体识别任务中广泛应用的字符级模式(CLM)在区分名称标记和非名称标记的二元任务中能力的研究,并在其他语言数据集上添加 CLM 模型来提高该数据集上 NER 任务的性能,证明了 CLM 提供了一个简单而强大的模型来捕获两种标记间的差异。Mayhew 等[55]利用词典将一或多种训练数据较多的语言中可用的注释数据翻译成另一种目标语言,并在这种语言中学习标准的单语 NER 模型,此方法是一种可移植性较高的跨语言命名实体识别方法。Hamdi 等[56]提出了名为 UNER 的数据集,这是一个用于 NER 任务的多语言和分层并行的语料库。创建的三步过程为:从维基百科文章中提取实体,将它们链接到 DBpedia 并将 DBpedia 类集映射到 UNER 标签,最后处理程序,可显著增加最终结果中已识别实体的数量。

总的来说,对于跨语言命名实体识别问题,现有的研究大多利用了迁移学习的知识,利用参数共享来实现不同语言之间命名实体的移植。还有的研究使用了自注意力机制,目的是使得不同语言的词序方面具有一定的灵活性。跨语言的命名实体识别任务的研究可以使得原本在资源较多的语言上训练的模型在数据不足的其他语言上具有通用性,具有良好的研究意义。

5　相关数据集和评价指标

5.1　常用数据集

对于 NER 任务的实验而言,比较常用的数据集包括:CoNLL-2003OntoNotes5.0,ACE2004,ACE2005 和 MSRA 等。

(1)CoNLL-2003[57]:该命名实体数据集主要由英语和德语两种语言组成。其中,英语语料的主要来源是路透社 RCV1 语料库的新闻专线,包含 4 种类型的命名实体:位置(LOC)、组织(ORG)、人物(PER)和杂项(MISC)。

(2)OntoNotes5.0[58]:该数据集中包含了 1 745 000 条英语数据,900 000 条中文数据和 300 000 条阿拉伯语数据,来源包括广播对话、广播新闻、新闻专线、杂志、电话对话和网络文本等。数据集包含用于 CoNLL-2012 任务的 18 种命名实体,由 11 种类型和 7 个值组成。

(3)ACE2004[59]:该数据集的版权所有者是语言数据联盟,它包含了用于对 2004 年自动内容提取(ACE)技术进行评估的英语、中文和阿拉伯语数据,包含了 7 个实体类型,分别是“PER”“ORG”“LOC”“GEP”“VEH”“WEA”和“FAC”。可以用在 NER,关系抽取等任务上。

(4)ACE2005[59]:该数据集的版权也属于 LDC,包含的语言和实体类型也与之相同。它的数据来源主要包括新闻广播、微博和广播对话等。

(5) MSRA[60]:该数据集是一个中文数据基准数据集,包含 5 万多条中文 NER 标注数据。数据来源于新闻领域,当中的命名实体包含 3 种类型。曾被用于 SIGNAN backoff 2006 上的共享任务。

5.2　标注方法

目前,命名实体识别任务常用的标注方法包括 IOB 标注法、BIOES 标注法和 Makeup 标注法。

(1)IOB 标注法:此方法是一种在 CoNLL-2003 任务上被广泛采取的标注方法。其中,I(Inside)表示中间;O(Outside)表示外部;B(Begin)表示开头。比如,“Steve Jods is my name”这句话被标注成“Steve B-PER Jobs I-PER is O my O name O”。

(2)BIOES 标注法:此方法由前一种方法扩展而来,但更加完备。其中 B,I,O 表示的意义不变,E 表示这个词的位置在一个句子的结束,S 表示这个单词本身就可以组成一个实体。此方法也是在命名实体识别任务中应用最广泛的方法。

(3)Makeup 标注法:这是一种应用于 OntoNotes 上的标注方法,相对来说比较简单。举例如下:

〈ENAMES TYPE = ” ORG ”〉 Universal Studios 〈/ENAMEX〉 recently opened.

此方法用标签框出命名实体,接着在 TYPE 上设置实体的类型。

5.3　评价指标

对于命名实体识别任务而言,目前比较通用的评价指标包括正确率、精确率、召回率和 F1 值。

正确率(Accuracy):在归类中所有预测正确的样本占总样本的比值,即

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN} \tag{2}$$

其中，TP 指正确匹配的数目，TN 指正确非匹配的数目，FN 指将本身正类预测为负类的数量，FP 指将本身负类预测为正类的数量。

精确率(Precision)：将样本中正类预测正确的数量占所有预测为正类数量的比例，即

$$Precision=\frac{TP}{TP+FP} \tag{3}$$

召回率(Recall)：样本中将正类预测正确的数量占所有真正类的比例，即

$$Recall=\frac{TP}{TP+FN} \tag{4}$$

F1 值：综合考虑了精确率和召回率，即：

$$F1=\frac{2Precision*Recall}{Precision+Recall} \tag{5}$$

结束语 总的来说，作为自然语言处理中一项重要的基础任务，命名实体识别为其他许多相关任务奠定了良好的基础。目前，该任务的主流模型仍然是基于 LSTM-CRF 的神经网络模型。最近，随着注意力机制的迅速崛起，不少研究将自注意力机制引入到模型中并且取得了不错的效果。目前该任务的难点主要在于：有限的命名实体只在有限的领域中，命名实体表达方式的多样性等。此外，由于嵌套实体的特殊性，如何处理嵌套实体的识别也成为该任务面临的一大难点。本文介绍了命名实体识别任务的研究意义以及它的发展历程，从基于规则到基于统计再到现如今火热的基于深度学习的方法和目前该领域的几个研究趋势，其中最为重要的是在资源匮乏下的命名实体识别研究，该研究对于解决困扰 NER 的难题意义重大。展望命名实体识别未来的发展，用迁移学习的方法处理匮乏资源下的 ner，采用图神经网络改进下游模型从而解决嵌套实体识别的难题，采用远程监督学习的方法解决训练时间过长的难题，将成为新的发展趋势。

参 考 文 献

[1] HAHNE A, FRIEDERICI A D. Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes[J]. Journal of Cognitive Neuroscience, 1999, 11(2): 194-205.

[2] SHANG W, HUANG H, ZHU H, et al. A novel feature selection algorithm for text categorization[J]. Expert Systems with Applications, 2007, 33(1): 1-5.

[3] YANG Y, WU Y N, LI J. Text similarity calculation based on potential feature words[J]. Computer Engineering and Design, 2011, 32(2): 572-575.

[4] 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.

[5] SORICUT R, BRILL E. Automatic question answering: Beyond the factoid[C]// Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. 2004: 57-64.

[6] ZHANG J, XIE J, HOU W, et al. Mapping the knowledge structure of research on patient adherence: knowledge domain visualization based co-word analysis and social network analysis[J]. PloS One, 2012, 7(4): e34497.

[7] RAU L F. Extracting company names from text [C]// Proceedings The Seventh IEEE Conference on Artificial Intelligence Application. IEEE Computer Society, 1991: 29, 30, 31, 32-29, 30, 31, 32.

[8] 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: 012069.

[9] XIA C, ZHANG C, YANG T, et al. Multi-grained named entity recognition[C]// 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019. Association for Computational Linguistics(ACL). 2020: 1430-1440.

[10] NI J, FLORIAN R. Improving Multilingual Named Entity Recognition with Wikipedia Entity Type Mapping[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 1275-1284.

[11] ZHANG X Y, WANG T, CHEN H W. Research on Named Entity Recognition[J]. Computer Science, 2005(4): 44-48.

[12] CHEN S D, OUYANG X Y. Overview of named entity recognition technology[J]. Radio Communication Technology, 2020, 46(3): 251-260.

[13] SHEN W, WANG J, LUO P, et al. Linden: linking named entities with knowledge base via semantic knowledge[C]// Proceedings of the 21st International Cconference on World Wide Web. 2012: 449-458.

[14] HUNG J C, CHANG J W. Multi-level transfer learning for improving the performance of deep neural networks: Theory and practice from the tasks of facial emotion recognition and named entity recognition [J]. Applied Soft Computing, 2021, 109: 107491.

[15] PETASIS G, VICHOT F, WOLINSKI F, et al. Using machine learning to maintain rule-based named-entity recognition and classification systems [C]// Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. 2001: 426-433.

[16] CHITICARIU L, KRISHNAMURTHY R, LI Y, et al. Domain adaptation of rule-based annotators for named-entity recognition tasks[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010: 1002-1012.

[17] LI W, MCCALLUM A. Rapid development of Hindi named entity recognition using conditional random fields and feature induction[J]. ACM Transactions on Asian Language Information Processing(TALIP), 2003, 2(3): 290-294.

[18] KONKOL M, KONOPÍK M. CRF-based Czech named entity recognizer and consolidation of Czech NER research[C]// International Conference on Text, Speech and Dialogue. Berlin: Springer, 2013: 153-160.

[19] XU Z, QIAN X, ZHANG Y, et al. CRF-based hybrid model for word segmentation, NER and even POS tagging[C]// Proceed-

ings of the Sixth SIGHAN Workshop on Chinese Language Processing. 2008.

[20] LIN B Y,XU F F,LUO Z,et al. Multi-channel bilstm-crf model for emerging named entity recognition in social media[C]// Proceedings of the 3rd Workshop on Noisy User-generated Text. 2017:160-165.

[21] MIKHEEV A,MOENS M,GROVER C. Named entity recognition without gazetteers[C]// Ninth Conference of the European Chapter of the Association for Computational Linguistics. 1999:1-8.

[22] BORTHWICK A,STERLING J,AGICHTEIN E,et al. Exploiting diverse knowledge sources via maximum entropy in named entity recognition[C]// Sixth Workshop on Very Large Corpora. 1998.

[23] TSAI R T H,WU S H,LEE C W,et al. Mencius:A Chinese named entity recognizer using the maximum entropy-based hybrid model[C]// International Journal of Computational Linguistics & Chinese Language Processing. 2004:65-82.

[24] KROGH A,LARSSON B,VON HEIJNE G,et al. Predicting transmembrane protein topology with a hidden Markov model:application to complete genomes[J]. Journal of Molecular Biology,2001,305(3):567-580.

[25] BIKEL,DANIEL M. Nymble:a High-Performance Learning Name-finder[C]// Fifth Conference on Applied Natural Language Processing. 1997:194-201.

[26] LI Y,SHETTY P,LIU L,et al. BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. 2021.

[27] LIU F,ZHAO J,LV B,et al. Product named entity recognition based on hierarchical hidden Markov model[C]// Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing. 2005.

[28] PENG N,DREDZE M. Improving named entity recognition for Chinese social media with word segmentation representation learning[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016:149-155.

[29] LIU T,YAO J G,LINC 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.

[30] LAMPLE G,BALLESTEROS M,SUBRAMANIAN S,et al. Neural Architectures for Named Entity Recognition[C]// HLT-NAACL. 2016.

[31] DONG X,QIAN L,GUAN Y,et al. A multiclass classification method based on deep learning for named entity recognition in electronic medical records[C]// 2016 New York Scientific Data Summit(NYSDS). IEEE,2016:1-10.

[32] SHAO Y,HARDMEIER C,NIVRE J. Multilingual named entity recognition using hybrid neural networks[C]// The Sixth Swedish Language Technology Conference(SLTC). 2016.

[33] MNIH V,HEESS N,GRAVES A. Recurrent models of visual attention[C]// Advances in Neural Information Processing Systems. 2014:2204-2212.

[34] BAHDANAU D,CHO K,BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. arXiv:1409.0473,2014.

[35] LIAO F,MA L,PEI J,et al. Combined Self-Attention Mechanism for Chinese Named Entity Recognition in Military[J]. Future Internet,2019,11(8):180.

[36] WU C,LUO G,GUO C,et al. An attention-based multi-task model for named entity recognition and intent analysis of Chinese online medical questions[J]. Journal of Biomedical Informatics,2020,108:103511.

[37] EGAN T M,YANG B,BARTLETT K R. The effects of organizational learning culture and job satisfaction on motivation to transfer learning and turnover intention[J]. Human Resource Development Quarterly,2004,15(3):279-301.

[38] LIN Y,YANG S,STOYANOV V,et al. A multi-lingual multi-task architecture for low-resource sequence labeling[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(Long Papers). 2018:799-809.

[39] QU L,FERRARO G,ZHOU L,et al. Named Entity Recognition for Novel Types by Transfer Learning[C]// EMNLP. 2016.

[40] FRANCIS S,VAN LANDEGHEM J,MOENSM F. Transfer learning for named entity recognition in financial and biomedical documents[J]. Information,2019,10(8):248.

[41] KRUENGKRAI C,NGUYEN T H,ALJUNIED S M,et al. Improving Low-Resource Named Entity Recognition using Joint Sentence and Token Labeling[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:5898-5905.

[42] RUDER S. An overview of multi-task learning in deep neural networks[J]. arXiv:1706.05098,2017.

[43] DEVLIN J,CHANG M W,LEE K,et al. Bert:Pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810.04805,2018.

[44] YU H,MAO X L,CHI Z,et al. A Robust and Domain-Adaptive Approach for Low-Resource Named Entity Recognition[C]// 2020 IEEE International Conference on Knowledge Graph(ICKG). IEEE,2020:297-304.

[45] YASEEN U,LANGER S. Data Augmentation for Low-Resource Named Entity Recognition Using Backtranslation[J]. arXiv:2108.11703,2021.

[46] ZHU W,LIU J,XU J,et al. Improving Low-Resource Named Entity Recognition via Label-Aware Data Augmentation and Curriculum Denoising[C]// China National Conference on Chinese Computational Linguistics. Cham:Springer,2021:355-370.

[47] WANG Y,LI Y,TONG H,et al. HIT:nested named entity recognition via head-tail pair and token interaction[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2020:6027-6036.

[48] WANG J,SHOU L,CHEN K,et al. Pyramid:A layered model for nested named entity recognition[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:5918-5928.

[49] LONG X,NIU S,LI Y. Hierarchical Region Learning for Nested Named Entity Recognition[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing:Findings. 2020:4788-4793.

[50] XU Y, HUANG H, FENG C, et al. A Supervised Multi-Head Self-Attention Network for Nested Named Entity Recognition [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 14185-14193.

[51] FU Y, TAN C, CHEN M, et al. Nested Named Entity Recognition with Partially-Observed TreeCRFs[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 12839-12847.

[52] MUELLER D, ANDREWS N, DREDZE M. Sources of Transfer in Multilingual Named Entity Recognition[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8093-8104.

[53] XIE J, YANG Z, NEUBIG G, et al. Neural Cross-Lingual Named Entity Recognition with Minimal Resources[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 369-379.

[54] YU X, MAYHEW S, SAMMONS M, et al. On the Strength of Character Language Models for Multilingual Named Entity Recognition[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3073-3077.

[55] 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.

[56] HAMDI A, LINHARES PONTES E, BOROS E, et al. A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers[C]// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 2328-2334.

[57] SANG E F T K, DE MEULDER F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition[J]. Development, 1837, 922: 1341.

[58] PRADHAN S, MOSCHITTI A, XUE N, et al. Towards robust linguistic analysis using ontonotes[C]// Proceedings of the Seventeenth Conference on Computational Natural Language Learning. 2013: 143-152.

[59] ZHANG Y, YANG J. Chinese NER Using Lattice LSTM[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers). 2018: 1554-1564.

[60] LEVOWG A. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition [C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006: 108-117.

GAO Xiang, born in 1996, postgraduate. His main research interests is natural language processing named entity recognition.

WANG Shi, born in 1981, Ph. D, associate researcher, is a member of China Computer Federation. His main research interests include natural language processing semantic analysis and knowledge graph.