DOI:10.16644/j.cnki.cn33-1094/tp.2021.04.005

中文命名实体识别研究方法综述

李嘉欣，王　平

(陆军工程大学国防工程学院，江苏 南京 210001)

摘　要：命名实体是存在于现实世界里的事物，它们与现实世界有着相互作用、相互影响的关系，因此命名实体在一些场景里是很重要的。文章从命名实体识别的定义着手，逐步阐述它从始初到如今的发展状况和识别方法及手段，分析命名实体识别的主要难点，最后通过命名实体识别的三个评价指标来判断实体的边界是否正确，以及实体的类型是否标注正确。

关键词：自然语言处理；命名实体识别；条件随机场；评价指标；信息抽取

中图分类号：TP391　　**文献标识码：**A　　**文章编号：**1006-8228(2021)04-18-04

A review of research methods of Chinese named entity recognition

Li Jiaxin, Wang Ping

(*College of National Defense Engineering, Army University of Engineering, Nanjing, Jiangsu 210001, China*)

Abstract: Named entities are things that exist in the real world. They interact and influence each other with the real world. Therefore, named entities are very important in some scenarios. Starting with the definition of named entity recognition, this paper gradually elaborates its development from the beginning to the present and its recognition methods and means, and analyzes the main difficulties of named entity recognition, and finally judges whether the entity boundary is correct and whether the entity type is marked correctly through three evaluation indexes of named entity recognition.

Key words: natural language processing; named entity recognition; conditional random fields; evaluation index; information extraction

0 引言

随着大数据时代的出现和机器学习的发展，自然语言处理NLP变得越来越重要，而自然语言处理中的一个热门的研究方向--命名实体识别NER也发展了起来。命名实体识别(Namedentityrecognition，NER)是机器翻译、问答系统、信息抽取和自然语言处理中一项重要的任务[1]，它的目的是从给出的一段文本中找出其中所有的实体，并将实体的属类标注出来。如今，命名实体识别在生物医学、警情军事及农业渔业等方面均有广泛应用。

1 命名实体识别的定义

在1991年LisaF.Rau[2]研究如何从文本中抽取公司的名称，论文主要是将人工编写规则的方法与启发式算法结合以此来实现公司名称识别的问题。在1993年宋柔等[3]人出了一本基于规则库的识别方法的书用于识别人名。命名实体识别被正式的提出是在1995年的第六届MUC会议上，在之后的第七届MUC会议上给出了命名实体识别需要识别的三大类和七小类的实体。后来在CONLL会议上将其定义为包含名称的短语[5]。SIGHANBakeoff-2006，Bakeoff-2007等也大多采用了这种分类[1]。除了主流的NER评测会议之外，Petasis、Alfonseca、Sekine、Borrega、Marrero等[1]学者也对命名实体的含义和类型进行热烈的讨论。

2 命名实体识别的研究方法

早期用来命名实体识别的方法主要是基于规则的方法和基于字典的方法，但随着语料的增加制定的

收稿日期：2020-11-16

作者简介：李嘉欣(1997-)，女，安徽阜阳人，硕士研究生，主要研究方向：语音交互与自然语言理解。

规则也将越来越繁琐，使用基于规则的方法和基于字典的方法就会过于费时费力。随着大数据时代的到来，如HMM、SVM和CRF等传统的机器学习方法也被学者用于命名实体识别的任务上，到后期的深度学习方法，将神经网络模型引入进行命名实体识别和近期开始将注意力机制和迁移学习应用其中，命名实体识别的发展正受到了各方面的持续关注。

2.1 基于词典和规则的方法

在正式提出命名实体概念后，早期的研究主要是采用规则和字典的方法[6]。基于规则的方法和基于字典的方法都是要构建大量的规则集或字典，然后按照需求将需要识别的汉字串放入制定的规则集中或与所构建的字典进行匹配，经过多次修正直到匹配成功。但这种方法只能在小数据集上得到很好的准确率，而随着数据集中数据的增加这种方式变得不再适用。

由于使用基于词典和规则的方法进行命名实体识别存在限制，它只能在特定的语料上识别能够得到很高的准确度，这样随着需要识别的实体多样化就需要制定更多的规则和更大的词典，这样的工作也会越来越重和越来越复杂。而随着机器学习的发展，在进行命名实体识别的任务时也开始考虑基于统计机器学习的方法。

2.2 基于统计机器学习的方法

基于统计机器学习的方法有基于隐马尔可夫模型（HiddenMarkovMode，HMM）的方法、基于最大熵（MaxmiumEntropy）的方法、基于支持向量机（SupportVectorMachine，SVM）的方法和基于条件随机场（ConditionalRandomFields，CRF）的方法。

2.2.1 统计机器学习方法在NER的第一步

第一个基于支持向量机（SVM）的NE系统是由Yamada等人在2001年提出来的。

2002年HidekiIsozak和HidetoKazawa将支持向量机应用在命名实体识别任务上。同年，McCallum等人将条件随机场也应用到命名实体识别的任务上。

在2004年庄明，老松杨，吴玲达[7]等人提出了基于统计的命名实体识别的方法。该方法首先将一段文本切分为几段较短的句子，再使用自增长统计算法从切分后的句子中生成最初的数据集，并利用得到的信息筛选出实体。这个方法不需要建立专业领域的大规模语料库，而是基于文本自身的用词特点进行统计分析，在实验中取得了良好的效果。但它的缺点是对带有前缀的模式的生成过程难以进行有效分析而且自增长生成算法作为一种统计方法对低频词条无法有效识别。

2.2.2 统计机器学习方法在NER的第二步

继统计机器学习的方法提出之后，人们尝试使用统计与规则结合的方法用于命名实体识别。如2005年向晓雯等[8]采用了统计与规则结合的方法进行命名实体识别，其首先使用HMM模型对文本做词性标注，然后使用制定好的规则对已经标注的文本再有一个修改的过程，他们采用的这种方法来做命名实体识别时，其三项评价标准均得到较高的数值。

2006年张剑[9]主要是对英文文本进行命名实体识别的研究，文中分别利用了改进的隐马尔可夫模型和条件随机域模型两种方法进行英文文本的命名实体识别，这个方法既兼顾了每个句子内部的局部特征又兼顾到每个词在同一文档中全局特征。同年，Okanohara[10]在生物领域上使用改进的半监督条件随机场模型进行命名实体识别，主要用于识别蛋白质，DNA和RNA等实体。

2009年高国洋，戚银城，潘德锋[11]等对中文地名识别进行了研究，提出了一种结合多知识的识别方法，该方法首先以CRF模型为框架，将专家知识与局部特征以及复合特征相融合进行命名实体识别；并利用构建的规则库对识别结果进行修正。

2010年鞠久朋等[12]也提出将条件随机场与规则相结合的方法用于地理空间中的命名实体识别。

2.3 深度学习下的命名实体识别

随着基于神经网络模型的深度学习技术在机器学习领域的发展越来越深入[4]。尤其是使用词向量来表示词语的方法，一方面解决了由于高纬度向量空间的原因导致的数据稀疏问题，另一方面词向量本身也比人工选择的特征包含更多的语义信息，而且该方法可以从异构的文本中获取统一向量空间下的特征表示，给NER中的命名实体识别问题带来强大的发展动力。

在2011年Collobert[13]等学者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。

2015年Yonghuiwu[14]等学者在文中先后使用了两次深度神经网络，首先是对未标注的文本使用一次深度神经网络进行训练，然后再使用一次深度神经网络对训练后的词向量进行实体识别，该方法在生物学上

F值超过了使用统计的方法进行命名实体识别。同时ZhihengHuang[15]等人使用双向长短期记忆模型用于命名实体，文中采用双向长短期记忆模型和条件随机场即BILSTM-CRF模型进行命名实体识别。

2016年XuezheMal和EdurdHovy[16]提出的BILSTM-CNNs-CRF模型，相比于ZhihengHuang提出的BILSTM-CRF模型来说，文中多一步使用卷积神经网络训练词向量，再将训练后的向量作为双向长短期记忆模型的输入，再使用CRF对输出建立标签关系，以此来更有效的实现实体识别。

MouradGridach，HatemHaddad在2017年先前用于现代标准阿拉伯语（MSA）的命名实体识别（NER）模型在很大程度上依赖于功能和地名词典的使用，这非常耗时。在文中介绍了一种基于双向选通递归单元（GRU）结合条件随机场（CRF）的新型神经网络架构，这种新颖的体系结构可以消除了对大多数手工工程功能的需求。

2.4 近期的命名实体识别研究方法

近期的命名实体识别方法除了在循环神经网络模型进行模型改进提出门控GRU模型外，还尝试使用CNN、SVM、Lattice、BERT等模型进行语言预训练，并在模型中引入注意力机制、迁移学习等来提高命名实体识别的准确度。

周晓磊[17]等人提出了使用SVM-BILSTM-CRF的神经网络模型对财产纠纷案件进行命名实体识别。文中提出构建一个关于财产纠纷裁判文书的语料库，首先使用支持向量机将文本中包含实体的句子筛选出来，然后输入给BILSTM-CRF模型进行训练，最后使用训练好的模型对财产纠纷裁判文书进行命名实体识别王博冉[18]等人方法利用LatticeLSTM模型来提取句子中的词汇词，从而将文本中词信息输入到基于字符的循环神经网络-条件随机场模型中。与基于字符和基于词的命名实体识别方法相比，本文提出模型的优势是其利用显性词汇信息而不是字符序列进行标注，且得到的分词误差率也得到了大幅度下降。

王子牛[19]等人提出了基于BERT的神经网络方法进行命名实体识别。BERT预训练语言模型具有双向Transformer结构不仅可以增强字的语义表示，还可以根据其上下文生成语义向量。文中提出的方法结合BERT和BILSTM-CRF模型对中文实体进行识别，以无需添加任何特征的方式提升了实体识别的准确率、召回率及F1值，验证了该方法的有效性。

武惠[20]等人提出了一种基于迁移学习的命名实体识别模型即TrBILSTM-CRF模型，该模型通过迁移学习算法将源域的信息迁移到所需的目标域中以此可以从其他领域获取到目前工作任务中所需要的信息。实验结果表明，TrBILSTM-CRF模型在小规模数据集上进行中文机构名命名实体识别时，其准确率、召回率和F值相比于其他方法，取得了较好的效果。

3 命名实体识别的研究难点

由于中文的多样性且并没有可以将其划分开的明显标注，这导致我们在对中文进行命名实体识时会更加的困难。实体能否被准确的从文本中识别出来主要在这两个方面：是否可以准确的划分出实体的边界；是否可以准确的判断出实体属类。

中文命名实体识别的难点有以下几个方面。

⑴ 命名实体的定义：在对中文进行实体识别时，可以发现中文实体的数量庞大且属类也很模糊，这就导致我们在对中文命名实体进行定义时会有很大的分歧。首先是在对其进行命名实体识别时需要大量的标注数据而这恰恰也是我们缺少的，而且在对数据进行标注时还需要有一个命名实体的标准，这也是目前的难点之一。

⑵ 歧义的消解：传统的词典规则方法可以很容易召回文本序列中在词表匹配到的词，但它的局限在无法解决歧义问题。一种典型的歧义是多种可能划分的问题，比如下面这个例子。输入序列：看到良方正在浇花。可以分为：看到/良方/正在/浇花；也能分为：看到/良方正/在/浇花。

⑶ 边界的界定：虽然深度学习对歧义的消解有显著优势，但它通常会遇到的问题是对新词的边界把握模糊。而词典中包含了大量词的边界信息。因此如何把词典信息融入到深度学习模型中是近几年研究的主流。现如今的方法是先对文本进行中文分词，再对得到的词进行标注。

⑷ 缺少标注数据：我们在进行命名实体识别的过程中不可缺少的就是已经标注好的数据，这也是在进行机器训练中不能缺少的一环。但这些标注好的数据都是需要人工去标注的，这样就需要更多的人力投入到这个过程中，而这个已有的标注数据并不完全适用于各个领域，这样，在我们对特殊领域进行命名

实体识别时就必须先人工构建这个领域的标注数据库，一旦这个标注数据库不够准确或者数据不够多，就会使得计算机的学习能力大幅度下降且难以训练出好的结果，这也是命名实体识别的难点之一。

4 展望

命名实体识别作为机器翻译、问答系统、信息抽取和自然语言处理的研究热点之一。通过阅读，近年来学者们在神经网络模型的基础上通过引入注意力机制和迁移学习等方法以此取得了大量新的研究成果，而命名实体识别未来的发展也将围绕这些方面。如今我们通过构建法律、生物、医学、军事等领域的数据库并在该数据库上进行实验，都取得了不错的结果。

但这里不可避免会遇到未登录词的问题，未登录词指的是那些没有被收录在词典中但必须切分出来的词。而如何识别并处理未登录词将是命名实体识别未来的一个重要研究方向。因此，后面的工作我们不仅要将已有的命名实体识别方法应用在各个领域上还要通过不断的改进模型来提高命名实体识别的准确度，并在命名实体识别的基础上对未登录词展开研究。

参考文献(References):

[1] 刘浏，王东波．命名实体识别研究综述[J]．情报学报，2018.37(3):329-340

[2] RauLF. Extracting Company Names from Text[C]. In: Proceeding softhe 7th IEEE Conference on Artificial Intelligence Applications.1991:29-32

[3] 宋柔．基于语料库和规则库的人名识别方法[M]．计算语言学研究与应用，北京语言学院出版社，1993.

[4] 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

[5] Grishman R, Sundheim B. Message Understanding Conference-6: ABriefHistory[C]. In: Proceeding softhe 16th International Conferenceon Computational Linguistics,1996.

[6] 隋臣．基于深度学习的中文命名实体识别研究[D]．浙江大学硕士学位论文，2017.

[7] 左明，老松扬，吴玲达．一种统计和词性相结合的命名实体发现方法[J]．计算机应用，2004.1:22-24

[8] 向晓雯，史晓东，曾华琳．一个统计与规则相结合的中文命名实体识别系统[J]．计算机应用，2005.10:2404-2406

[9] 张剑．基于CRF的英文命名实体识别研究[D]．哈尔滨工业大学硕士学位论文，2006.

[10] Daisuke Okanohara, Yusuke Misyao. Yoshimasa Tsuruka. Improvingthe Scalability of Semi-Markov Conditional Random Fiekds for Named Entity Recognition[C]. Proceeding softhe21 "International Conferenceon Computational Linguisticsand 44th Annual Meetingofthe ACL,2006:465-472

[11] 高国洋，戚银城，潘德锋．基于条件随机场与规则相结合的中文地名识别[J]．电脑开发与应用，2009.22(8):26-28

[12] 鞠久朋，张伟伟，宇建军，周国栋．CRF与规则相结合的地理空间命名实体识别[J]．计算机工程，2011.37(7):210-212,215

[13] COLLOBERTR, WESTONJ, BOTTOUL, etal. Natural Language Processing(almost) from Scratch[J]. Journal of Machine Learning Research,2011.12(Aug):2493

[14] YonghuiW, MinJiang, JianboLei, HuaXu. Named Entity Recognitionin Chinese Clinical Text Using Deep Neural Network. Studiesin Health Technology and Informatics,2015:624-628

[15] Zhiheng Huang, Wei Xu and Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. arXiv, 2015.1508.01991

[16] MAX, HOVYE. End-to-end Sequence Labeling ViaBi-directional LSTM-CNNs-CRF[J]. arXivpreprintarXiv: 1603.01354,2016.

[17] 周晓磊，赵薛蛟，刘堂亮，宋子潇，王其乐，里创桥．基于SVM-BILSTM-CRF模型的财产纠纷命名实体识别方法[J]．计算机系统应用，2019.28(1):245-250

[18] 王博冉，林夏，朱晓东，朱万琳，马学华．LatticeLSTM神经网络法中文医学文本命名实体识别模型研究[J]．中国卫生信息管理杂志，2019.16(1):84-88

[19] 王子牛，姜猛，高建瓴，陈姚先．基于BERT的中文命名实体识别方法[J]．计算机科学，2019.46(S2):138-142

[20] 武惠，吕立，于碧辉．基于迁移学习和BILSTM-CRF的中文命名实体识别[J]．小型微型计算机系统，2019.40(6):1142-1147