Computer Science and Application 计算机科学与应用, 2021, 11(5), 1588-1597
Published Online May 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.115164

Hans汉斯

复杂数据上的实体识别综述

王诗怡，贺　萍*

河北经贸大学信息技术学院，河北　石家庄
Email: wangshiyi_1001@163.com, *heping@heuet.edu.cn

收稿日期：2021年4月26日；录用日期：2021年5月21日；发布日期：2021年5月28日

摘　要

复杂数据的数据量大和数据源不同的特征导致在挖掘复杂数据中的潜在价值时，需要利用实体识别技术。实体识别技术能实现对传统数据进行完整刻画、对数据质量进行管理的重要操作。而在复杂数据进行实体识别具有识别效果差、识别精度不高等问题。本文首先从应用领域的角度探讨复杂数据上的实体识别技术，包括社交网络领域的敏感实体识别、军事领域的目标实体识别、商业领域的商情实体识别。其次，对不同领域中的各个实体识别常用方法进行对比，分析了各个方法的问题与不足。最后，对在不同领域中进行实体识别的难点进行总结。

关键词

复杂数据，实体识别，敏感实体，目标实体，商情实体

A Survey of Entity Recognition on Complex Data

Shiyi Wang, Ping He*

College of Information Technology, Hebei University of Economics and Business, Shijiazhuang Hebei
Email: wangshiyi_1001@163.com, *heping@heuet.edu.cn

Received: Apr. 26th, 2021; accepted: May 21st, 2021; published: May 28th, 2021

Abstract

Complex data is characterized by a large amount of data and different data sources, which lead to the use of entity recognition technology in mining the potential value of complex data. Entity recognition technology can realize some important operations, such as complete description of tradi-

*通讯作者。

文章引用: 王诗怡, 贺萍. 复杂数据上的实体识别综述[J]. 计算机科学与应用, 2021, 11(5): 1588-1597.
DOI: 10.12677/csa.2021.115164

tional data and data quality management. However, entity recognition technology applied in complex data has the problems of poor recognition effect and low recognition accuracy. This paper first discusses entity recognition technology on complex data from the perspective of application field, including sensitive entity recognition in social network field, target entity recognition in military field and business entity recognition in commercial field. Secondly, the usual methods of entity recognition in different fields are compared, and the problems and shortcomings of each method are analyzed. Finally, the difficulties of entity recognition in different fields are summarized.

Keywords

Complex Data, Entity Recognition, Sensitive Entities, Target Entity, Business Entity

1. 引言

随着各个行业内复杂数据的数据量在信息化时代均呈爆发式增长，这使得挖掘数据的潜在价值愈发重要，如何更好地利用这些复杂数据逐渐被人们重视。而复杂数据通常具有数据量大和数据源不同的特征。数据量大使得处理数据时需要使用大量的计算资源，数据源不同导致在数据使用之前数据内部逻辑或结构需要统一整合。这两个特征是判断数据是否属于复杂数据的基础，另外，复杂数据的复杂性需要通过数据结构、数据量大小、数据的颗粒度、数据查询语言、数据类型、数据分散性以及数据增长率 7 个指标进行深入判断[1] [2]。这些指标导致在复杂数据上开展相关研究将比传统数据更加困难，并存在一些难以解决的问题。

实体识别技术是自然语言处理(Natural Language Processing, NLP)技术[3] [4]中的一项重要的基本任务。这项基本任务可以简单总结为在目标文本中识别出命名实体，并标注其位置与类型的过程。根据 MUC-6 将命名实体划分为时间类(TIMEX)、实体类(EMAMEX)与数字类(NUMEX)，可以将实体识别问题看作一种特殊的序列标注问题(Sequence Labeling Problem) [5] [6] [7] [8]，这类问题不但与常规的输入特征相关，而且与目标文本的上下文之间有一定的联系。如表 1，早期实体识别技术通常使用基于规则的方法[9] [10] [11]建立字典、规则库、语料库以及规则模板，从而使目标文本中的特征词能够根据人工规则被计算机识别并提取，但人工规则的构建需要耗费大量人力与时间进行人工标注，不能适应当前复杂数据拥有的数据量大的特征，进而发展形成基于统计的方法。基于统计的方法[12]利用机器学习算法弥补了基于规则的方法需要人工工作带来的损失，同时通过构建模型实现特征词语的自动提取，实体识别的效率得到增强。但基于统计的方法仅能够通过大量数据完成对模型的训练，面对数据量较小的数据集往往难以保证实体识别的正确率，因此能够处理小规模数据的基于深度学习的方法应运而生。基于深度学习的方法[13] [14] [15]根据深度学习模型完成小样本条件下实体的自动识别[16]-[21]，同时更好地联系上下文关系。但使用基于深度学习的方法仅能针对某一领域进行实体识别，而非提取实体在各个领域的全部特征，无法做到目标实体全部特征的联合，导致目标实体识别的准确性受到影响。传统实体识别使用的基于规则的方法与基于统计的方法一般将目标文本的识别分为特征工程与文本分类两个部分，在特征工程部分完成人工标注，在文本分类部分完成特征提取，但这两个阶段往往消耗大量人力成本，且最后得到的识别结果也并不准确，因此运用深度学习方法进行实体识别是较好的选择。实体识别技术又被称为

实体消歧[22]、数据消重[23]、重复检测等[24] [25]，常用于社交媒体领域[26] [27]的敏感实体识别、军事领域[28]的目标实体识别或是商业领域[29]的商情实体识别。在自然语言处理(Natural Language Processing, NLP)中，作为识别单位的文本包括人名、地名、机构名、时间和数量[30] [31]，这些具有重要意义的主体均可以作为实体识别的对象进行研究。在对现实世界的复杂数据进行实体识别操作时，现实世界存在的多条记录可能指向同一个实体对象，而这些杂乱的、待清洗的数据记录未必以相同的表现形式呈现。另外，这些作为实体识别对象的复杂数据还存在重复、缺失、陈旧、虚假问题。因此，理清这些记录与实体对象之间的关系，消除多余的、无用的信息对于后续的相关应用研究非常重要。

Table 1. The development of named entity recognition method (time sequence from left to right)
表 1. 根据时间顺序发展(由左至右)的实体识别方法对比分析

方法	基于规则的方法	基于统计的方法	基于深度学习的方法
优点	能够从目标文本中识别出特定实体	自动提取实体特征模板，提高识别效率	能够处理小规模数据集；更好地联系上下文
缺点	耗费人力	难以保证小规模样本识别的准确率	仅能完成单一领域的实体识别

Table 2. Traditional method of named entity recognition
表 2. 传统实体识别的方法

基于规则的方法		沈达洋[73]等通过统计地名用字概率建立了地名规则与地名规则库 郑家恒[74]等建立地名语料库 张晓恒[75]等通过总结高校组织名的文本特征设计了大量规则模板
基于统计的方法	有监督方法	王红斌[76]等采用最大熵模型的方法针对泰语句子进行了实体关系抽取 张玥杰[77]等提出了一种融合多特征的最大熵汉语命名实体识别模型 高冰涛[78]等将生物医学文本中的命名实体识别问题转换为基于迁移学习的隐马尔可夫模型问题，降低了生物医学文本中命名实体识别对目标领域标注数据的需求 何炎祥[79]等根据自设定规则结合条件随机场模型提高了地理命名实体识别效果 李博[80]等利用条件随机场模型对收集到的特征进行分类识别，提高了中文电子病历命名实体识别效果
	无监督方法	朱祥[81]等提出了用于降低语音特征参数维数的分段均值算法，聚类交叉分组算法和 HMM 分组算法的组合形式，提高了英语语音的识别率与系统识别速度 王浩畅[82]等使用了基于支持向量机的方法对生物医学文本中的命名实体进行了识别 孙琛琛[83]等提出一种面向实体识别的聚类算法弥补了以往聚类算法没有涉及到的匹配决定问题
基于深度学习的方法		张娜娜[84]等提出一种基于 BiLSTM-CRF 模型与词典规则相结合的识别方法识别工艺操作说明文本中的命名实体 张帆[85]等设计了一种通过神经网络语言模型获得词和词性特征分布式表达的医疗文本实体识别方法

根据识别对象的不同，在识别过程中运用的方法也不尽相同，如表 2 所示。传统的复杂数据实体识别技术按时间发展顺序可以分为三类：基于规则(Rule-Based)的方法、基于统计(Statistic-Based)的方法和基于深度学习的方法[32]。基于规则的方法是早期研究中的常见的一类方法，该类方法通过人工构造规则，以模式和字符串相匹配的方式识别实体。随后，基于统计的方法大多采用基于机器学习的序列标记方法与自动构造规则的监督学习方法进行复杂数据实体识别。以上两种方法通常以语言模型与机器学习算法为基础，包括：1) 解决序列标记问题的有监督方法，如隐马尔科夫模型(Hidden Markov Model, HMM)、最大熵(Maximum Entropy, ME)、最大熵马尔科夫模型(Maximum Entropy Markov Model, MEMM)、条件随机场(Conditional Random Fields, CRF)等[33] [34] [35] [36]；2) 让上下文中的同义文本自行聚类的无监督

方法，如支持向量机(Support Vector Machine, SVM)等。最近，基于深度学习的方法在实体识别领域的研究热点，如卷积神经网络(Convolutional Neural Networks, CNN) [37] [38]与循环神经网络(Rerrent Neural Network, RNN) [39]是应用最广泛的方法。基于深度学习的复杂数据实体识别方法通过对深层神经网络的研究，在识别词性、抽取字段等方面均在不断发展[40]。

目前，对于实体识别技术的研究大部分集中于文本边界识别技术、文本重复检测技术、歧义识别技术与关系数据[2]等方面。而对复杂数据的实体识别技术从应用领域角度进行分析的研究却很少。因此，本文从应用领域方面对现有命名实体识别技术进行对比与分析，分别从社交网络、军事、商业为主的特殊领域对实体识别方法进行综述，并对各领域主要的实体识别技术的难点进行总结。

2. 面向社交网络领域的敏感实体识别

面向社交网络领域的敏感实体识别是特定领域的命名实体识别(Name-entity recognition, NER)，主要是对互联网领域中的专业词汇进行识别和分类。所识别的实体相较于普通命名实体更具专业性与复杂性，识别范围更小，要求识别精度也更高。社交网络领域的敏感实体识别主要任务是通过识别敏感信息文本数据的方式将社交网络中潜在的危险识别出来。建立强有力的监管措施已经成为了网络治理的有效手段，对互联网中极端思想和联络潜在人员传播的敏感实体进行识别与处理已经成为了当下维护社交网络安全的首要问题[41]。互联网敏感实体的识别能够帮助网警对于涉恐涉黑等危险信息提前做出预警，维护网络环境安全与社会稳定；使研究人员在收集互联网信息时更有针对性地对网络安全治理进行研究；提高互联网从业人员对互联网生态的了解，从而设计出更贴合人们日常需求的新产品。

2.1. 敏感实体识别常用方法

2.1.1. 基于规则的方法与基于统计的方法

Webb S [42]、Qiang C [43]、Liang Z [44]、Ding W [45]等人提出基于规则与基于统计的实体识别方法，分别对社交网络中的恶意实体、虚假实体、关键实体以及不良话题进行了识别。其中，Weeb S [42]及其研究团队提出基于机器学习的分类器模型分析恶意实体的行为特征的方法，但该方法仅能通过被动接收好友申请的方式获取垃圾邮件配置文件，导致识别恶意实体存在限制与风险。Qiang C [43]等人提出一种基于图的方法的实体识别工具，在一定程度上解决了识别具有被动性的问题，但其提出的 SybilRank 在识别准确性与效率方面却大打折扣。Liang Z [44]等人提出了基于动态查询扩展的方法，该方法通过动态提取网络中的关键字构建图模型，在一定程度上提高了识别虚假实体的准确性，并考虑到了社交网络的异构性，但其针对范围小，且效率仍有待提升。Ding W [45]等人开发并使用了半监督的 Dirichlet 处理过程进行社交网络短文本的实体识别，但联系上下文方面仍存在问题。

2.1.2. 基于深度学习的方法

与基于规则的方法和基于统计的方法相比，深度学习方法[46]具有更强的泛化能力和对人工特征的低依赖性，在实体识别的各个领域均取得了良好的效果。其中，卷积神经网络(CNN)与循环神经网络(RNN)是命名实体识别领域运用最广泛的深度学习方法，使用过程中通过与条件随机场(CRF)结合联系上下文特征，从而提高传统实体识别方法的识别效果。基于深度学习的方法中，CNN 与 RNN 在提取文本字符特征与序列特征方面有所提升。

CNN 通过“端到端”学习，能够很好地实现对数据样本特征的学习与表示。对于通常是由中英文混合的互联网敏感实体，CNN 主要用于处理文本中的英文、拼音等字母实体。相较于基于规则和基于统计的方法，CNN 对声母、韵母、字符大小写的处理更加细腻，所以通常利用 CNN 进行敏感实体字符特征

的提取。为了提高识别社交网络领域中中英文混合文本的准确率，魏笑[47]等人提出了一种基于部件 CNN 的网络安全命名实体识别方法(CC-NS-NER)。

为了处理更长的数据序列，RNN 逐渐由长短时记忆网络(LSTM)演变为双向长短时记忆网络(BiLSTM)。Chen T [48]等人提出一种使用 BiLSTM 分类文本语句并搭建序列模型过程的方法，相较于 Tai K S [49]等人基于 LSTM 技术进行语义分割的方法，该方法得到的准确率与效率都更高。原因是 BiLSTM 比起 LSTM 能够更加有效地联系文本的上下文信息。而融合双向门控循环单元(BiGRU) [50]是 BiLSTM 的一种优化结构，在保持了 BILSTM 原有的效果同时使得结构更加简单，但在社交网络领域的敏感实体识别应用中仍待进一步研究。

2.2. 敏感实体识别的难点

互联网敏感实体的识别难点可归结如下：一是互联网敏感实体具有数量大、种类多以及更新迅速的特性，由于互联网用户具有创造性，许多新的命名实体被创造，这为敏感实体识别的工作进行带来困难；二是互联网敏感实体具有“一词多义”的特性，同一名称的敏感实体之间界限不明确将造成旧词新用和不同语境下实体含义不同的问题；三是互联网敏感实体具有字符混杂的特性，识别过程当中常常难以区分混杂的中英文字符以及标点符号；四是互联网敏感实体存在着大量的代称与缩写，敏感实体识别技术难以联系上下文语境自行识别。以上特性都使得敏感实体难以识别，需要在识别效果方面进行优化[51]。

3. 面向军事领域的目标实体识别

面向军事领域的目标实体识别是特定领域的命名实体识别。部队作战中累积的文本情报数据蕴含着大量活动、动向、意图、趋势等揭示战场内幕的重要信息，具有重要的军事价值。为了能让情报效能得到充分发挥，为情报分析研判提供支撑[52]，从文本情报中自动抽取军事目标及其相关活动以完成目标实体识别成为了军事领域的重要任务。在军事领域，可以将目标实体分为军事人员、军事保障机构、军事保障设备、军事保障设施、军事装备名称 5 种类型。军事领域目标实体识别的主要任务通常包括从作战情报中识别出作战指令、作战文书与军事武器，这对识别精度以及实时性有着更高的要求。

3.1. 目标实体识别常用方法

3.1.1. 基于规则的方法与基于统计的方法

面向军事领域的目标实体识别前期多使用基于规则的方法和基于统计的方法结合而成的基于混合的方法[53]，但此类方法对于目标实体的识别效果很大程度上依赖于人工设计的规则与设计的特征模板。基于规则的目标实体识别方法主要使用 BIO (begin：开始，internal：中间，other：其他)方法表示每个输入单元，从而对输入序列进行标注。而条件随机场(CRF)作为基于统计的目标实体识别方法，能够综合利用包括字、词、词性在内的上下文信息与外部特征，在一定程度上提高了军事目标实体识别的效果。姜文志[54]等人提出一种 CRF 与规则相结合的方法，通过结合基本特征与外部词典特征进行军事目标实体识别，但其识别效果仅能达到通用领域水平，仍有待改进。

3.1.2. 基于深度学习的方法

基于深度学习的双向长短时记忆(BiLSTM)方法能够有效地利用上下文信息，识别文本序列间的顺序关系，降低目标实体识别中对于规则与特征模板的依赖。齐玉东[55]、李健龙[56]、高学攀[57]等人利用 BiLSTM 具有的自动学习任务特征与 CRF 模型结合的方法有效联系了上下文信息，提升了目标实体的识别效果，但 BiLSTM 与 RNN 均难以避免分词时原始字向量的长短对实体识别结果的影响，所以双向循环神经网络(Bidirectional Recurrent Neural Network, BiRNN) [58]逐渐替代神经网络按照军事目标实体实

例关系构建分词模型。丁海强[59]等人提出一种基于双向循环神经网络(Bidirectional Recurrent Neural Network, BiRNN 模型的海军军械不均衡文本数据集处理方法，利用 BiRNN 模型自动学习文本序列的特征均衡目标实体数据，提高目标实体文本分类的性能[60]。

3.2. 目标实体识别的难点

相较于传统领域的命名实体识别，军事领域目标实体具有独特的命名规则[57]，目标实体之间的辨析难度更大。增加军事领域知识更够提高军事实体的识别效果，但相应领域知识具有一定保密性，在学习上具有局限性与滞后性的问题。在目标实体识别过程中，面对军用文书以及标绘图等不同的信息载体，以及专有名词数量较多且嵌套使用的情况，切词分词也是一项重要难点。目标实体在识别过程中常常难以做到根据上下文联系进行准确识别，而是依赖于构造的特征模板与规则进行词义联系，这将造成识别结果的不准确。

4. 面向商业领域的商情实体识别

面向商业领域的商情实体识别是特定领域的命名实体识别。社会中存在的大量商业信息可供人们挖掘其潜在价值以发现商机，进而带来经济效益。商业领域的商情实体通常出现在工程单位发标时编制的标书，网络销售产品网站以及包装产业生产的包装电子文档中，其产品名具有构成复杂、长度较长等特点，如“防静电透明 PVC 板棒”、“双通道连卷背心袋机”等，这种结构使得商情实体识别比传统的命名实体识别更为复杂和困难。商情实体识别是充分挖掘商业信息价值必不可少的步骤。

4.1. 商情实体识别常用方法

4.1.1. 基于规则的方法与基于统计的方法

面向商业领域的商情实体识别早期主要使用基于规则的方法与基于统计的方法。基于规则的方法需要使用 BIO 技术对商业领域出现的商情实体进行标注。条件随机场(CRF)作为基于统计的方法，能够考虑标签之间的关系来获取全局最优标签序列，利用句子级别的标签信息集成任意知识源，使实体抽取结果更加准确[61] [62] [63] [64]。方莹[65]等人在爬取到的英文农产品语料库中使用 CRF 对商情实体名称包含的特有关键词进行了特征模板抽取与特征参数训练，使商情识别的准确度有所提升，但 CRF 模型训练的时间依赖于特征空间的规则，导致商情识别的效率仍需提高。

4.1.2. 基于深度学习的方法

基于深度学习的方法常使用循环神经网络(RNN)及其衍生模型与注意力(Attention)等机制结合的方法[66] [67]。RNN 类型的神经网络模型能对序列数据中非限定长度的上下文信息进行表示，不同于基于统计的方法对人工提取特征的依赖性，RNN 适用于构建词的分布式特征，提升了商情实体识别效率。Attention 机制能实现不同层级间的信息传递，加快商情实体语言特征识别的速度与准确度。贾全烨[68]等人提出基于 RNN 与 CRF 结合的方法识别商情实体，发现在识别准确性与模型训练成本等方面均未达到理想效果。黄晓[69]等人将 RNN 与 Attention 机制结合的方法应用于商情实体识别过程中，虽然准确率相较于 RNN-CRF 模型有所提升，但复杂度较高，且执行效率不如其他 RNN 衍生模型。

BiLSTM、GRU 等作为 RNN 的衍生模型，在解决序列标注问题、识别字符较长的商情实体以及利用上下文信息提取句子特征方面均取得了良好的效果。张应成[70]等人提出一种 BiLSTM-CRF 模型，以获取商情实体序列化文本的整体特征。虽然在双向语义识别以及识别准确率方面均获得了提升，但模型训练时间消耗仍然较大。李一斌[71]等人提出一种 BiGRU-CRF 模型，该模型能够自动学习商情实体的分布式特征，在识别准确率方面相较于传统 CRF 模型有所提高，但仅能保证小数量的商情实体识别效果。

4.2. 商情实体识别的难点

商业领域的商情实体具有行业种类繁多的特性，这常对商情实体识别的上下文联系产生阻碍。商情实体还具有领域特殊性与复杂性，且实体名称用词随意，难以在识别过程中快速确定商情实体，导致识别效率低下。其呈现的文本类资源结构化低，可利用和传播性较差，造成商情实体识别难以从大量不同来源的非结构化文本资源中自动提取相关实体。商情实体名称的构成复杂、关键词混乱、长度较长等特性也使得商情实体识别比一般的实体识别更为复杂和困难[72]。

5. 总结与展望

网络智能化与数据化使得数据呈现复杂的特点，而在复杂数据中挖掘有用的潜在信息，需要利用有效的实体识别方法。目前，对复杂数据上的实体识别技术通常从技术角度进行总结研究，而实体识别技术从应用领域进行汇总的研究却很少体现。本文从社交网络、军事、商业为主的特殊领域对实体识别技术进行具体对比与分析，发现各领域实体识别过程中存在的一些难点，如实体种类繁杂、实体名称不规范、实体数量庞大等，以及各领域实体识别使用的方法虽然相似，但各模型对于不同领域的实体识别起到的作用却不尽相同的规律。

在复杂数据不断涌现的背景下，早期基于规则的方法的研究思路仍然给人以宝贵的启示，基于规则和统计相结合的混合方法仍不时得到有效的尝试，基于深度学习的方法在复杂数据上的实体识别应用中也具有巨大的发展空间。基于规则和基于统计的实体识别方法中出现的问题可以由不断发展的基于深度学习的方法进行解决[86] [87]，针对各领域复杂数据实体的特性完善实体识别的方法，以及对于基于深度学习的实体识别方法存在的仅能针对某一领域进行实体识别，而非提取实体在各个领域的全部特征，无法做到目标实体全部特征的联合的问题，仍可以作为复杂数据上实体识别的研究重点。实体识别在特定领域上的应用已经成为了各领域处理复杂数据的关键技术，未来面对复杂数据潜在价值的挖掘工作仍需要继续研究在复杂数据上的具备高效特性的实体识别技术。考虑到基于广度学习的方法和基于联邦学习的方法能够融合多个领域的实体特征的特点，复杂数据上实体的特性导致难以识别和难以联合的问题与广度学习以及联邦学习相结合有着广阔的发展空间。

基金项目

河北省教育厅科学研究计划项目(重点项目，ZD2019017)。

参考文献

[1] Elmagarmid, A.K., Ipeirotis, P.G. and Verykios, V.S. (2007) Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge & Data Engineering*, **19**, 1-16. https://doi.org/10.1109/TKDE.2007.250581

[2] 梁吉业, 冯晨娇, 宋鹏. 大数据相关分析综述[J]. 计算机学报, 2016(1): 1-18.

[3] 王丁. 关于自然语言处理技术的分析与研究[J]. 科技创新导报, 2020, 17(7): 141-142.

[4] 林莉. 人工智能时代背景下自然语言处理技术的发展[J]. 电子世界, 2020(22): 24-25.

[5] 黄睿, 李辰, 王涛, 等. 语言序列标注方法, 装置存储介质及计算机设备[P]. 中国专利, CN201811481219.2. 2020-06-12.

[6] 阳萍, 谢志鹏. 基于 BiLSTM 模型的定义抽取方法[J]. 计算机工程, 2020(3): 40-45.

[7] 黄胜, 王博博, 朱菁. 基于文档结构与深度学习的金融公告信息抽取[J]. 计算机工程与设计, 2020, 41(1): 115-121.

[8] 王宗极. 基于深度学习的复杂场景车牌识别研究[D]: [硕士学位论文]. 北京: 中国地质大学(北京), 2020.

[9] 孔玲玲. 面向少量标注数据的中文命名实体识别技术研究[D]: [硕士学位论文]. 杭州: 浙江大学, 2019.

[10] 纪文璐, 王海龙, 苏贵斌, 柳林. 基于关联规则算法的推荐方法研究综述[J]. 计算机工程与应用, 2020, 56(22): 39-47.

[11] 谢德鹏, 常青. 关系抽取综述[J]. 计算机应用研究, 2020, 37(7): 1921-1924, 1930.

[12] 黄超. 基于统计方法从文本中抽取分词词典[J]. 电脑知识与技术, 2020, 16(4): 213-214.

[13] 陈娟, 王卓薇, 程良伦. 基于深度学习的命名实体识别算法[J]. 计算机科学与应用, 2021, 11(3): 628-634. https://doi.org/10.12677/CSA.2021.113064

[14] 仇增辉, 赫明杰, 林正奎. 基于深度学习的网购评论命名实体识别方法[J]. 计算机工程与科学, 2020, 42(12): 189-196.

[15] 高亮亮. 基于深度神经网络的中文医疗文本实体识别[D]: [硕士学位论文]. 成都: 电子科技大学, 2020.

[16] 陈茹, 卢先领. 融合空洞卷积神经网络与层次注意力机制的中文命名实体识别[J]. 中文信息学报, 2020, 34(8): 70-77.

[17] 唐国强, 高大启, 阮彤, 等. 融入语言模型和注意力机制的临床电子病历命名实体识别[J]. 计算机科学, 2020, 47(3): 211-216.

[18] 徐凯, 王崎, 李振彰, 等. 基于结合多头注意力机制 BiGRU 网络的生物医学命名实体识别[J]. 计算机应用与软件, 2020, 37(5): 151-155+232.

[19] 马千程, 王崑声, 周晓纪. 基于深度学习的竞争情报命名实体识别研究[J]. 情报探索, 2020(9): 1-7.

[20] 丁晟春, 方振, 王楠. 基于 Bi-LSTM-CRF 的商业领域命名实体识别[J]. 现代情报, 2020, 40(3): 103-110.

[21] 刘小安, 彭涛. 基于卷积神经网络的中文景点识别研究[J]. 计算机工程与应用, 2020, 56(4): 140-145.

[22] 周国民, 宣鑫乐, 沈佳琪, 等. 基于实体关联的消歧算法研究[J]. 中国电子科学研究院学报, 2020, 15(3): 271-277.

[23] Ahmed, A.M., Patel, A. and Khan, M. (2021) Super-MAC: Data Duplication and Combining for Reliability Enhancements in Next-Generation Networks. *IEEE Access*, **9**, 54671-54689. https://doi.org/10.1109/ACCESS.2021.3070993

[24] Che, S., Yang, W. and Wang, W. (2020) Improved Streaming Quotient Filter: A Duplicate Detection Approach for Data Streams. *The International Arab Journal of Information Technology*, **17**, 769-777. https://doi.org/10.34028/iajit/17/5/10

[25] 王宏志, 樊文飞. 复杂数据上的实体识别技术研究[J]. 计算机学报, 2011, 34(10): 1843-1852.

[26] 徐啸, 朱艳辉, 冀相冰. 基于自注意力深度学习的微博实体识别研究[J]. 湖南工业大学学报, 2019, 33(2): 48-52.

[27] 王超, 王峥. 基于改进分词标注集的中文微博命名实体识别方法[J]. 计算机与数字工程, 2019, 47(1): 211-215.

[28] 刘卫平, 张豹, 陈伟荣, 等. 基于迁移表示学习的军事命名实体识别[J]. 指挥信息系统与技术, 2020, 11(2): 64-69.

[29] 刘程波. 基于实体识别和情感分析的商品评论主体观点挖掘[D]: [硕士学位论文]. 上海: 东华大学, 2020.

[30] Grishman, R. and Sundheim, B. (1996) Message Understanding Conference 6: A Brief History. *Proceedings of the* 16*th Conference on Computational Linguistics*, **1**, 466-471. https://doi.org/10.3115/992628.992709

[31] United States Defense Advanced Research Projects Agency (DARPA), Information Technology Office (1995) Named Entity Task Definition, Version 2.1. *Message Understanding Conference*-6 (*MUC*-6), Morgan Kaufmann, Columbia, Maryland, November 1995, 319 -332.

[32] 莎仁, 梁琼芳, 李长明, 张家鑫. 大数据实体识别相关技术研究[J]. 软件导刊, 2020, 19(3): 125-127.

[33] Ratnaparkhi, A. (1996) A Maximum Entropy Model for Part-of-Speech Tagging. *Proceedings of the Empirical Method for Natural Language Processing*, Stroudsburg, 1996, 133-142.

[34] McCallum, A., Freitag, D. and Pereira, F.C.N. (2000) Maximum Entropy Markov Models for Information Extraction and Segmentation. *Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford, CA, June 2000, 591-598.

[35] Lafferty, J. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the* 18*th International Conference on Machine Learning*, Williams College, Williamstown, MA, 28 June-1 July 2001, 282-289.

[36] 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.

[37] 徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述[J]. 计算机学报, 2020, 43(5): 755-780.

[38] 谢博, 申国伟, 郭春, 等. 基于残差空洞卷积神经网络的网络安全实体识别方法[J]. 网络与信息安全学报, 2020,

6(5): 126-138.

[39] 王栋，李业刚，张晓，等. 基于准循环神经网络的中文命名实体识别[J]. 计算机工程与设计，2020，41(7): 2038-2043.

[40] 陈基. 命名实体识别综述[J]. 现代计算机, 2016(2): 24-26.

[41] 黄炜，童青云，李岳峰. 基于广度学习的异构社交网络敏感实体识别模型研究[J]. 情报学报，2020，39(6): 579-588.

[42] Webb, S., Caverlee, J. and Pu, C. (2008) Social Honeypots: Making Friends with a Spammer near You. *The Fifth Conference on Email and Anti-Spam*, Mountain View, CA, 21-22 August 2008.

[43] Cao, Q., Sirivianos, M., Yang, X., *et al.* (2012) Aiding the Detection of Fake Accounts in Large Scale Social Online Services. *USENIX Conference on Networked Systems Design & Implementation*, San Jose, CA, 25-27 April 2012, 1-14.

[44] Liang, Z., Feng, C., Jing, D., *et al.* (2014) Unsupervised Spatial Event Detection in Targeted Domains with Applications to Civil Unrest Modeling. *PLoS ONE*, **9**, e110206. https://doi.org/10.1371/journal.pone.0110206

[45] Ding, W., Yue, Z., Chen, C., *et al.* (2017) Semi-Supervised Dirichlet-Hawkes Process with Applications of Topic Detection and Tracking in Twitter. 2016 *IEEE International Conference on Big Data* (*Big Data*). Washington DC, 5-8 December 2016, 869-874. https://doi.org/10.1109/BigData.2016.7840680

[46] 张政馗，庞为光，谢文静，吕鸣松，王义. 面向实时应用的深度学习研究综述[J]. 软件学报，2020，31(9): 2654-2677.

[47] 魏笑，秦永彬，陈艳平. 一种基于部件 CNN 的网络安全命名实体识别方法[J]. 计算机与数字工程，2020，48(1): 106-111.

[48] Chen, T., Xu, R., He, Y., *et al.* (2016) Improving Sentiment Analysis via Sentence Type Classification Using BiLSTM-CRF and CNN. *Expert Systems with Applications*, **72**, 221-230. https://doi.org/10.1016/j.eswa.2016.10.065

[49] Tai, K.S., Socher, R. and Manning, C.D. (2015) Improved Semantic Representations from Tree-Structured Long Short-Term Memory Networks. *Proceedings of the* 53*rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, 26-31 July 2015, 1556-1566. https://doi.org/10.3115/v1/P15-1150

[50] 饶竹一，张云翔. 基于 BiGRU 和注意力机制的多标签文本分类模型[J]. 现代计算机, 2020(1): 31-35.

[51] 张建权. 基于 CNN 和 BiGRU-attention 的互联网敏感实体识别方法[J]. 网络安全技术与应用, 2020(4): 61-65.

[52] 徐树奎，曹劲然. 基于层级式 Bi-LSTM-CRF 模型的军事目标实体识别方法[J]. 信息化研究，2019，45(6): 18-22, 46.

[53] 王传栋，徐娇，张永. 实体关系抽取综述[J]. 计算机工程与应用, 2020, 56(12): 25-36.

[54] 姜文志，顾佼佼，丛林虎. CRF 与规则相结合的军事命名实体识别研究[J]. 指挥控制与仿真, 2011, 33(4): 13-15.

[55] 齐玉东，丁海强，吴晋豫，等. 融合本体特征的 BiLSTM-CRF 军事实体识别模型[J]. 兵器装备工程学报，2020, 41(5): 118-123.

[56] 李健龙，王盼卿，韩琪羽. 基于双向 LSTM 的军事命名实体识别[J]. 计算机工程与科学, 2019, 41(4): 713-718.

[57] 高学攀，杜楚，吴金亮. 基于 BiLSTM-CRF 的军事命名实体识别方法[J]. 无线电工程, 2020, 50(12): 1050-1054.

[58] 刘明明，李震霄，郑丽丽. 基于双向循环神经网络的字符级文本分类[J]. 江苏建筑职业技术学院学报，2019, 19(4): 29-34.

[59] 齐玉东，丁海强，赵锦超，等. 基于 biRNN 的海军军械不均衡文本数据集处理方法[J]. 计算机与现代化，2019(12): 21-26.

[60] Xu, B., Yan, S. and Yang, D. (2019) BiRNN-DKT: Transfer Bi-Directional LSTM RNN for Knowledge Tracing. In: Ni, W., Wang, X., Song, W. and Li, Y., Eds., *Web Information Systems and Applications. WISA* 2019. *Lecture Notes in Computer Science*, Vol. 11817, Springer, Cham, 22-27. https://doi.org/10.1007/978-3-030-30952-7_3

[61] 李培英，杨鉴. 基于 BERT-CRF 模型的缅甸语韵律单元边界预测[J]. 计算机科学与应用, 2021, 11(3): 505-514. https://doi.org/10.12677/CSA.2021.113051

[62] 付瑶，万静，邢立栋. 基于条件随机场与信息熵的特定领域概念发现[J]. 计算机应用研究，2020，37(3): 708-711, 730.

[63] 王莉，陈兆熙，余丽. 基于条件随机场的多标签图像分类识别方法[J]. 计算机仿真, 2020, 37(8): 394-397.

[64] 黄定琦，史晟辉. 基于条件随机场的汉语词汇特征研究[J]. 计算机应用研究, 2020, 37(6): 1724-1728, 1754.

[65] 方莹. 基于条件随机场的英文农产品名识别[J]. 河南科学, 2011, 29(3): 350-353.

[66] 石磊, 王毅, 成颖, 等. 自然语言处理中的注意力机制研究综述[J]. 数据分析与知识发现, 2020, 4(5): 1-14.

[67] 单义栋, 王衡军, 黄河, 等. 基于注意力机制的命名实体识别模型研究——以军事文本为例[J]. 计算机科学, 2019, 46(z1): 111-114, 119.

[68] 贾全烨, 张强, 宋博川. 一种基于循环神经网络的电网客服语音文本实体识别算法[J]. 供用电, 2020, 37(6): 13-20. https://doi.org/10.19421/j.cnki.1006-6357.2020.06.003

[69] 黄晓, 林嘉良, 滕蔚, 等. 基于多卷积窗尺寸注意力卷积神经网络实体关系抽取方法[P]. 中国专利, CN201911143069.9. 2020-03-17.

[70] 张应成, 杨洋, 蒋瑞, 等. 基于 BiLSTM-CRF 的商情实体识别模型[J]. 计算机工程, 2019, 45(5): 308-314.

[71] 李一斌. 基于双向 GRU-CRF 的中文包装产品实体识别[J]. 华东理工大学学报(自然科学版), 2019, 45(3): 486-490.

[72] 王海宁, 周菊香, 徐天伟. 融合深度学习与规则的民族工艺品领域命名实体识别[J]. 云南师范大学学报(自然科学版), 2020, 40(2): 48-54.

[73] 沈达阳, 孙茂松. 中国地名的自动辨识[C]//全国第三届计算语言学联合学术会议论文集. 北京: 清华大学出版社, 1995: 68-74.

[74] 郑家恒, 李鑫, 谭红叶. 基于语料库的中文姓名识别方法研究[J]. 中文信息学报, 2000, 14(1): 7-12.

[75] 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, 11(4): 22-33.

[76] 王红斌, 李金绘, 沈强, 等. 基于最大熵的泰语句子级实体从属关系抽取[J]. 南京大学学报(自然科学), 2017, 53(4): 738-746.

[77] 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.

[78] 高冰涛, 张阳, 刘斌. BioTrHMM: 基于迁移学习的生物医学命名实体识别算法[J]. 计算机应用研究, 2019, 36(1): 45-48.

[79] 何炎祥, 罗楚威, 胡彬尧. 基于 CRF 和规则相结合的地理命名实体识别方法[J]. 计算机应用与软件, 2015, 32(1): 179-185, 202.

[80] 李博, 康晓东, 张华丽, 等. 采用 Transformer-CRF 的中文电子病历命名实体识别[J]. 计算机工程与应用, 2020, 56(5): 153-159.

[81] 朱祥. 基于隐马尔可夫模型和聚类的英语语音识别混合算法[J]. 计算机测量与控制, 2020, 28(5): 175-179.

[82] 王浩畅, 赵铁军. 基于 SVM 的生物医学命名实体的识别[J]. 哈尔滨工程大学学报, 2006, 27(z1): 570-574.

[83] 孙琛琛, 申德荣, 寇月, 等. 面向实体识别的聚类算法[J]. 软件学报, 2016, 27(9): 2303-2319.

[84] 张娜娜, 王裴岩, 张桂平. 面向工艺操作说明文本的命名实体深度学习识别方法[J]. 计算机应用与软件, 2019, 36(11): 188-195, 261.

[85] 张帆, 王敏. 基于深度学习的医疗命名实体识别[J]. 计算技术与自动化, 2017, 36(1): 123-127.

[86] Ma, C. and Zhang, C. (2021) Joint Pre-Trained Chinese Named Entity Recognition Based on Bi-Directional Language Model. *International Journal of Pattern Recognition and Artificial Intelligence*. https://www.worldscientific.com/doi/10.1142/S0218001421530037

[87] Yang, J., Wang, H., Tang, Y., *et al.* (2021) Incorporating Lexicon and Character Glyph and Morphological Features into BiLSTM-CRF for Chinese Medical NER. 2021 *IEEE International Conference on Consumer Electronics and Computer Engineering* (*ICCECE*), Guangzhou, 15-17 January 2021, 12-17. https://doi.org/10.1109/ICCECE51280.2021.9342121