情报学报　2018 年 3 月　第 37 卷　第 3 期
Journal of the China Society for Scientific and Technical Information, Mar. 2018, 37(3): 329-340

DOI: 10.3772/j.issn.1000-0135.2018.03.010

命名实体识别研究综述

刘　浏[1,2]，王东波[3,2]

（1. 南京大学信息管理学院，南京　210023；2. 江苏省数据工程与知识服务重点实验室（南京大学），南京　210023；3. 南京农业大学信息科学技术学院，南京　210095）

摘　要　命名实体识别一直以来都是信息抽取、自然语言处理等领域中重要的研究任务，随着机器学习技术的新发展，数字人文研究的兴起，事件知识和实体知识变得越发重要，命名实体识别焕发出新的发展动力。本文详细梳理了命名实体识别从提出至今的发展脉络，从实体的定义、重要的评测会议、主流的研究方法研究的应用价值等角度，全面考察了该领域的研究现状，并分析了命名实体识别未来的发展趋势。

关键词　命名实体识别；实体挖掘；信息抽取

A Review on Named Entity Recognition

Liu Liu[1,2] and Wang Dongbo[3,2]

(1. *School of Information Management, Nanjing University, Nanjing*　210023;
2. *Jiangsu Key Laboratory of Data Engineering and Knowledge Service (Nanjing University), Nanjing*　210023;
3. *College of Information Science and Technology, Nanjing Agricultural University, Nanjing*　210095)

Abstract: Named Entity Recognition has been an important research topic in information extraction and natural language processing. With the development of machine learning and an increasing interest in digital humanities, entity recognition has gained importance. More importantly, the Named Entity Recognition research has indicated the potential of development in the field. This study shows the arising and the development of Named Entity Recognition from the most important conferences, the main algorithms to the most popular implementations. The future possibilities in the research field are proposed at the end.

Key words: Named Entity Recognition; entity extraction; information extraction

1　引　言

命名实体识别（named entity recognition，NER）是信息抽取和信息检索中一项重要的任务，其目的是识别出文本中表示命名实体的成分，并对其进行分类，因此有时也称为命名实体识别和分类（named entity recognition and classification，NERC）。随着计算机技术的发展，自然语言理解和文本挖掘研究的不断深入，以及数字人文研究的兴起，文本语义层面知识显得愈发重要，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，而命名实体作为文本中重要的语义知识，其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、

收稿日期：2017-11-08；修回日期：2018-03-10

基金项目：国家社会科学基金重大项目“基于《汉学引得丛刊》的典籍知识库构建及人文计算研究”（15ZDB127）；国家自然科学基金面上项目“基于典籍引得的句法级汉英平行语料库构建及人文计算研究”（71673143）；江苏省普通高校学术学位研究生科研创新计划项目“引用内容分析——引文语义信息的自动挖掘”（KYZZ16_0033）。

作者简介：刘浏，男，1989 年生，博士研究生，主要研究领域为自然语言处理、信息计量，E-mail: liuliu.nju@outlook.com；王东波，男，1981 年生，博士，副教授，主要研究领域为自然语言处理、文本挖掘、信息计量。

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。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，从规则方法到统计机器学习，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，命名实体识别仍然是一个重要且具有挑战性的研究课题。

2 什么是命名实体

命名实体（named entity，NE）作为一个明确的概念和研究对象，是在 1995 年 11 月的第六届 MUC 会议（MUC-6，the Sixth Message Understanding Conferences）上被提出的。MUC-6 和后来的 MUC-7 并未对什么是命名实体进行深入的讨论和定义，只是说明了需要标注的实体是“实体的唯一标识符（unique identifiers of entities）”，规定了 NER 评测需要识别的三大类（命名实体、时间表达式、数量表达式）、七小类实体，其中命名实体分为：人名、机构名和地名[1-2]。MUC 之后的 ACE 将命名实体中的机构名和地名进行了细分，增加了地理-政治实体和设施两种实体[3]，之后又增加了交通工具和武器[4]。CoNLL-2002、CoNLL-2003 会议上将命名实体定义为包含名称的短语，包括人名、地名、机构名、时间和数量，基本沿用了 MUC 的定义和分类，但实际的任务主要是识别人名、地名、机构名和其他命名实体[5-6]。SIGHAN Bakeoff-2006、Bakeoff-2007 评测也大多采用了这种分类[7-8]。

除了主流的 NER 评测会议之外，也有学者专门就命名实体的含义和类型进行讨论，Petasis 等[9]认为命名实体就是专有名词（proper noun，PN），作为某人或某事的名称。Alfonseca 等[10]从构建本体的角度，提出命名实体就是能用来解决特定问题的我们感兴趣的对象（objects）。Sekine 等[11]认为通用的 7 小类命名实体并不能满足自动问答和信息检索应用的需求，提出了包含 150 种实体类别的扩展命名实体层级（extended named entity hierarchy），并在后来将类别种数增加到 200 个[12]。

Borrega 等[13]从语言学角度对命名实体进行了详细的定义，规定只有名词和名词短语可以作为命名实体，同时命名实体必须是唯一且没有歧义的。比较特别的是，该研究将命名实体分为强命名实体（strong named entities，SNE）和弱命名实体（weak named entities，WNE），其中 SNE 对应词汇，而 WNE 对应短语，SNE 和 WNE 又可以细分为若干个小类。虽然该研究将每种类别都进行了详细的定义和阐释，但可能由于过于复杂而不利于计算机自动识别，因此该研究并未得到太多关注。

Nadeau 等[14]指出，“命名实体”中的“命名（named）”表示：只关心那些表示所指对象（referent）的严格指示词（rigid designators）。严格指示词的概念源于 Kripke[15]的观点，“对于一个对象 *x*，如果在所有存在 *x* 的世界中，指示词 *d* 都表示 *x*，而不表示别的对象，那么 *x* 的指示词 *d* 是严格的（a designator d of an object *x* is rigid if it designates *x* with respect to all possible worlds where *x* exists, and never designates an object other than *x* with respect to any possible world）”。

Marrero 等[16]总结了前人对命名实体的定义，并将之归纳为语法类别、严格指示、唯一标识和应用目的四种类别。作者先假设每种类别都能作为定义命名实体的标准，再通过分析和举例等方式否定其作为标准的可行性。最后得出的结论是，应用方面的需求目的，是定义命名实体唯一可行的标准。

学者们对命名实体的外延和内涵的探讨和辩论还远未结束。关于命名实体，目前也未有一个较为官方的、普遍得到认可的定义。但是自然语言处理的研究特点是实用第一，与其探讨、争论并期待一个十全十美的定义，不如先找到一个能够识别出“我认为是命名实体”的方法。而 NER 确实也是这么发展的，有影响力的评测会议提出需要识别的实体对象，研究者们就针对这些对象设计方法进行识别，评测会议之外的研究也大体沿用了这些会议的评测标准。绝大多数的研究者并不关心什么是命名实体，什么不是命名实体，他们更关心的是如何设计模型和方法来有效地识别出文本中的人名、地名和机构名，而这些显然都是命名实体。因此，纵观整个 NER 研究的历史，所谓命名实体识别中的命名实体，其实就是识别出文本中的人名、地名和机构名。

3 NER 发展的推动者：评测会议

3.1 MUC 评测

早在 1991 年就有研究尝试从金融新闻中自动抽取公司名称[17]，该研究一般被认为是命名实体识别研究的前身。在此之后虽也有零星相关研究，但真正使得命名实体识别成为一项明确且重要的研究任务的，是在 1995 年 11 月举办的第六届 MUC 会议。MUC 会议是一系列面向信息抽取研究的会议。该会议的特点是以评测的形式定义会议主题，参会者必须是评测比赛的参赛者。评测的比分使用信息检索领域常用的正确率（Precision）、召回率（Recall）和

F1 值（F1 score）。该会议的这种评测的形式后来被广泛使用在各类信息检索和自然语言处理会议当中。

MUC-6 的任务规范[1]中定义的实体较为广泛，包括了命名实体（ENAMEX）、时间表达式（TIMEX）和数量表达式（NUMEX）。其中命名实体又分为人名、地名和机构名。由于语料规模较小（30 篇文档），类型单一（均是新闻），因此参会的方法大多都取得了较好的识别效果，F1 值最高达到 96.42%[18]，而且人名的识别效果要明显好于地名和机构名[19]。由于采用的是华尔街日报语料，任务只是对英语句中命名实体的识别。因此 1996 年春进一步举办了 MET（the multilingual entity task），将命名实体识别的任务扩展到汉语、西班牙语和日语。然而其中汉语的识别效果相对较低，F1 值最高只有 84.51%[20]。1998 年的 MUC-7 和 MET-2，继承并修订了 MUC-6 的标注规范，增加了训练语料的规模，然而此次评测的英文识别效果并不如 MUC-6[21]，F1 值最高达到 93.39%[22]，中文的识别效果却有了明显的提高，F1 值达到 86%[23]。

MUC-6 中所有的研究都采用了基于规则的方法，如包括词形、词性的词汇规则[24]，短语规则[25]等。而大部分方法是根据命名实体前后的提示词、上下文语境等制定字符序列匹配规则[26-27]。MUC-7 大部分的研究也都是基于规则的[22,28-29]，MET-2 中的中文命名实体识别同样如此[30]。这些方法通过制定有限的规则和模式，然后从文本中自动寻找匹配这些规则或模式的字符串，并标记为各类命名实体。基于规则的方法尤其适合识别时间表达式和数量表达式。但对于命名实体而言，由于其构造规则随意多变，试图通过有限的规则来识别近乎无限的命名实体，是不合适的。而基于规则的方法领域性非常强，尽管这些方法在会议评测上表现出色，但并不能适用在通用领域的文本上。

值得注意的是，MUC-7 出现了基于统计机器学习方法的初步尝试，如最大熵 ME[31]、隐马尔可夫 HMM[23]等。

3.2　ACE 项目

鉴于 MUC 在命名实体领域获得的成功，NIST 从 1999 年开始举办 ACE（automatic content extraction）项目，目的是发展从人类语言中提取信息的自动内容抽取技术，而语言的形式不限于文本，还包括了语音和图像。该项目由 LDC（the linguistic data consortium）提供语料和标注规范[32]。ACE 的研究面向的是目标对象（如实体、关系和事件）而不仅仅是 MUC 中文本中的词语，这种不同在于，MUC 旨在识别实体的名称，而 ACE 更关注的具有这些名称的实体，因此 ACE 的任务是更为抽象的，而且暗示了 ACE 必须将指代和指代消解作为研究的一个重要部分[33]。

ACE 分为多个阶段，第一阶段（1999—2001）的主要任务是实体识别和跟踪（entity detection and tracking，EDT），实际上包含四个子任务：实体识别，实体属性识别，实体指代识别和指代内容识别，其中后两个子任务就是实体跟踪①。实体识别方面，ACE 将实体分为五类：人名、机构名、地理-政治实体、地名、设施[3]。其中地理-政治实体和设施实际上是对地名和机构名的拓展，这部分实体的识别也是 MUC 所不包含的。

ACE 的第二阶段（2002—2003）在 EDT 中加入了转喻（metonymy）[34]，并将研究内容扩展增加了实体关系的识别和描述（relation detection and characterization，RDC）[35]；ACE-2003 进一步加入了汉语和阿拉伯语的研究[36]。

ACE 的第三阶段（2004—2008）将研究领域进一步扩大。ACE-2004 开始包含了主要的任务：实体识别（EDT）、关系识别（RDR）和事件识别（VDR）[37]。EDT 中增加了两种实体类别：交通工具（vehicle）和武器（weapon）。此外还在三种语言上加入了实体链接（entity link tracking，LNK）的研究[4]。ACE-2005 加入了值（values）和时间表达式（temporal expressions）②的识别[38-39]，并进一步定义了事件③（events）[40]。至此，ACE 的共包括 5 个主要任务[41]。

ACE-2007 加入了西班牙语[42]，并且加入了新的任务：实体翻译（entity translation）[43-44]，而其余的任务较之 ACE-2005 没有太大变化[45]。ACE-2008 将研究语言限定为英语和阿拉伯语，并在本地（local，within-document）实体识别之外又提出了全局（global，cross-document）的实体识别[46-47]，再一次将实体识别的研究推上新的角度。

① 所谓的实体跟踪，实际上就是对指代的识别。

② 时间表达式可以看作一种特殊的关系。

③ 事件实际上是对 ACE 之前所有相关研究的整合。在事件中，事件的类型（types）代表实体间的关系，事件的参与者（participants）代表了实体，事件的特征（Attributes）就代表了实体的值。

3.3 CoNLL 会议

CoNLL（Conference on Computational Natural Language Learning）是由 ACL 的自然语言理解专门的兴趣小组（special interest group on Natural Language Learning，SIGNLL）举办的一年一度的学术会议，会议从 1997 年开始举办，并从 1999 年开始有了类似于 MUC 的评测任务（shared task），每一年的评测任务主题都不一样，其中 CoNLL-2002 和 CoNLL-2003 的主题是独立于语言的命名实体识别（Language-Independent Named Entity Recognition），即寻找在不同语言中都能有效识别命名实体的方法，该任务难度明显要高于 MUC。CoNLL-2002 和 CoNLL-2003 定义的命名实体包括人名、地名、机构名、时间和数量，其中 CoNLL-2002 的语料为西班牙语和荷兰语，CoNLL-2003 的语料为英语和德语。

CoNLL 两届会议对于命名实体识别的研究明显不同于 MUC，绝大多数参赛方法都使用了统计机器学习的方法，如隐马尔可夫 HMM[48]、最大熵 ME[49-50]、支持向量机 SVM[51]、条件随机场 CRF[52]、AdaBoost[53]等。一些联结主义的模型如 Winnow 方法[54]，表决感知器 Voted Perceptrons[55]和长短记忆网络 LSTM[56]也得到了尝试。而基于规则的方法中，只有转换学习 TBL 方法仍得到使用，但效果并不理想[57]。将各种方法整合也是当时研究的一项重要尝试[58-60]。

其中 Carreras 等[53]使用的 AdaBoost.MH 方法在 CoNLL-2002 上取得了最好的识别效果，在西班牙语和荷兰语上的 F1 值分别为 81.39 和 77.05。而在 CoNLL-2003 上，最好的方法都采用或整合了最大熵模型[49,59-60]，评测第一名[59]在英语和德语上的 F1 值分别为 88.76 和 72.41。

纵观 CoNLL 的命名实体识别研究，统计机器学习已经成为主流，各种重要的机器学习方法均已得到了尝试，虽然当时的语料规模并不大，但机器学习仍然表现出强大的性能。尤其是在面向独立于语言的命名实体识别时，由于语言间的差异性，规则方法已经无法适用，统计机器学习成为唯一可行且有效的方法。而此时，选择并改进合适的机器学习方法，选择更加合适的文本表示特征便成为了命名实体识别研究的主要思路和潮流。另一方面，CoNLL 的评测都是基于双语的，NER 的研究在面向双语时困难程度明显增加了，尤其是机器学习方法十分依赖语料中的特征表示，在没有大规模的双语语料支撑的前提下，想得到效果较好的双语通用 NER 系统，必然十分困难。因而此后的 NER 研究很长一段时间多面向单一语言，很少涉及双语甚至多语，直到近年来大数据环境的推进，双语、多语语料库规模不断增长，跨语言的 NER 研究才又逐渐浮出水面。

3.4 汉语 NER 研究和 SIGHAN Bakeoff

汉语的命名实体识别较之英语要更为复杂困难，这主要表现在汉语文本中没有表示词语边界的分隔符号，命名实体的识别效果很大程度受自动分词结果影响[61]，而汉语自动分词的效果往往也受制于命名实体的识别。中文信息处理中类似于命名实体识别的研究，最早出现就是为了提高汉语自动分词①的效果[62-63]，早期的汉语命名实体识别主要关注某一类命名实体，如人名[64]、机构名[65]等，采用的也大多是基于规则的方法[66-68]。

最早将汉语命名实体识别作为评测任务提出的，是 2003 年举办的“863 评测”，该评测持续到 2005 年，汉语 NER 作为自动分词的子任务，只在 2003 年和 2004 年出现，评测结果中 F1 值最高的只有 82.38%[69]。真正将汉语 NER 研究作为重要的研究领域，并组织较大规模评测会议，是从 SIGHAN Bakeoff-2006 开始的。

SIGHAN（the special interest group for chinese language processing）也是 ACL 的一个专门的兴趣小组，主要研究汉语自动分词。在汉语分词中，未登录词（OOV）是影响分词效果非常重要的因素，而命名实体是未登录词中最为显著的一种，因此命名实体识别是汉语自动分词无法回避的问题。SIGHAN 于 2006 年正式将 NER 问题作为其评测比赛（bakeoff）的一项任务。Bakeoff-2006 提供了三组汉语语料（MSRA、LDC 和 CITYU）②，并借鉴 CoNLL-2002 的体系，定义了 4 类命名实体：人名、地名、机构名和地理-政治实体③（GPE）[7]。到了 Bakeoff-2007，减去了 LDC 语料，命名实体也减少为最常见的三类：人名、地名、机构名[8]。

① 当时自动分词也称作词语识别（word identification）

② 其中 MSRA 和 LDC 为简体中文语料，CITYU 为繁体中文语料。

③ Bakeoff-2006 提供了三组语料，只有 LDC 的语料包含了 GPE 实体。而大部分参赛的方法并未选用这个语料进行训练和测试，因此 Bakeoff-2006 大多研究还是对常见三类命名实体的识别（人名、地名、机构名）。

在 Bakeoff 评测中，统计机器学习方法依然占主导，且普遍取得了较好的效果，在 Bakeoff-2006、Bakeoff-2007 两届评测中，名列前茅的 NER 系统几乎都采用了 CRF 模型[70-75]和 ME 模型[71,76]。Bakeoff 的评测得到了几个重要的发现：语料含未登录词的比例直接影响了识别效果；三类命名实体中，机构名识别最难；在训练语料上增加外在数据，对识别效果影响很大[8]。

SIGHAN Bakeoff-2010 没有继续关注命名实体识别，而是转而关注了命名实体相关的另一项重要问题：命名实体消歧（NED）[77]。到了 SIGHAN Bakeoff-2012，命名实体识别和消歧（NERD）作为一个全新的问题，成为该届 Bakeoff 的评测任务[78]。参赛的 NERD 系统大多分为 NER 和 NED 两部分，而 CRF 模型仍然是 NER 部分的首选[79-80]。此后 SIGHAN Bakeoff 未再将 NER 作为评测任务。

4　NER 的主要方法

4.1　基于规则的方法

早期尤其是 MUC 会议前后的 NER 研究，人工构建有限规则，再从文本中寻找匹配这些规则的字符串，是一种主流的方法。但即便是基于规则，研究者们也试图借助机器自动地发现和生成规则，这其中最具代表性的便是 Collins 等[81]提出的 DL-CoTrain 方法，其先预定义种子规则集 Decision List，再根据语料对该集合进行无监督的训练迭代得到更多的规则，最终将规则集用于命名实体的分类，该方法对命名实体三种类别（人名、地名和机构名）的分类准确率均超过了 91%。类似的还有使用 Bootstrapping 进行规则自动生成的方法[82]。与此同时，也有研究者提出了规则和统计模型（ME）相结合的 NER 系统，并认为加入统计模型后，不使用地名词典仍然可以很好地识别出地名[83]。可见研究者们已经意识到，虽然基于规则的方法虽然能够在特定的语料上获得较高的识别效果，但是识别效果越好，越需要大量规则的制定，而人工制定这些规则可行性太低。而试图通过制定有限的规则来识别出变化无穷的命名实体，这样的方法愈发显得笨重。更不用说规则对领域知识的极度依赖，使得当领域差别很大时，制定的规则往往无法移植，不得不重新制定规则。这些固有的缺点使得研究者们转而采取新的研究思路，而此时正值机器学习在 NLP 领域兴起，NER 自然地转向了机器学习的阵营。

4.2　基于统计机器学习的方法

随着世纪交接时期机器学习在 NLP 领域的兴起，NER 研究也逐渐转向了火热的机器学习阵营。基于统计机器学习的 NER 研究大体可以总结为以下几个方向：选择合适的模型和方法，进行模型和方法的改进，选择合适的特征，多种方法的综合。

4.2.1　模型和方法的选择

基于机器学习的 NER 方法归根到底都是分类的方法，给定命名实体的多个类别，再使用模型对文本中的实体进行分类。但其中也可以分为两种思路，一种是先识别出文本中所有命名实体的边界，再对这些命名实体进行分类[84]，如 Collins 等[81]的 CoBoost 方法，其通过拼写和上下文分别训练得到两个分类器，再基于 AdaBoost 整合得到一个适用于无标签语料的分类器。该方法对于命名实体三种类别（人名、地名和机构名）的分类准确率均超过了 91%。

另一种是序列化标注方法。对于文本中每个词①，可以有若干个候选的类别标签，这些标签对应于其在各类命名实体中所处的位置②。此时 NER 的任务就是对文本中的每个词进行序列化的自动标注（其实也是分类），再将自动标注的标签进行整合，最终获得有若干个词构成的命名实体及其类别③。序列化标注是目前最为有效，也是最普遍的 NER 方法。经典机器学习分类模型如 HMM[85-86]、ME[87]、CRF[52]和 SVM[88]都被成功地用来进行命名实体的序列化标注，且获得了较好的效果。

4.2.2　模型和方法的改进

机器学习在 NER 上获得的迅速成功并没有阻碍研究者们的研究热情，怎么样设计出识别效果更好的 NER 方法，一直是研究者孜孜不倦努力的目标。改进模型并提高模型的计算效率，是机器学习中最常见的研究思路，而在 NER 中这种思路同样可行。如 Zhou 等[89]提出的基于 HMM 的块标注器，Leaman 等[90]提出的半马尔可夫模型等。值得注意的是，

① 对于汉语一般是每个字。

② 比如在上下文“XXX Christopher D. Manning XXX”中，“Christopher”的标签可以是“PERSON_START”，“D.”的标签是“PERSON_MID”，“Manning”的标签是“PERSON_END”。

③ 如注释 8 例，最终可以得到“Christopher Manning”是一个“PERSON”。

模型改进研究在汉语命名实体识别研究中较为多见，这是因为经典的 NER 模型大多面向英语文本提出，并不直接适用于汉语，根据汉语特点调整和优化经典模型，能够更有效地识别汉语文本中的命名实体，如层叠马尔可夫方法[91]、多层条件随机场方法[92-93]等。

4.2.3 特征的选择

另一种提高 NER 效果的思路是选择更好的特征表示，而这种思路在 NER 中更为普遍和有效。NER 任务中，最常见的特征包括形态、本地（local）词汇和句法信息，形态特征有如词形、大小写、前后缀等。本地词汇特征有如前后提示词、窗口词、连接词等。最近，通过未登录词和非常规词的识别来提高 NER 的效果，也得到了尝试[94]。句法特征有词性、浅层句法结构等。由于汉语的特殊性，除了词汇层面的特征外，汉字层面的特征也被充分地用来辅助提高 NER 的效果，如提示单字[95]、常用尾字[96]等。同时，由于汉语分词和 NER 的密切联系，有研究发现分词结果可以有效地提高汉语 NER 的效果[97]。

为了提高识别的效果，各种全局（global①）信息也作为特征被广泛地应用在 NER 中，尤其是远距离依存[98-99]和上下文同指[100]等。与此同时，各种外部知识如未标注文本[101]、人名词典[52]、地名词典[102]等也被普遍使用来提高 NER 模型的性能。有研究表明，在模型不变的情况下，全局信息和外部知识确实可以显著地提高识别的效果[102]。值得注意的是，维基百科知识是最常见且有效的外部知识[103-105]，而在汉语 NER 中，知网（HowNet）作为一个汉语特有的词汇语义知识库，也被充分地应用在 NER 研究中[106]。

4.2.4 综合的 NER 方法

在选择、改进机器学习模型，选择更好的、更多的特征表示之外，也有研究综合使用多种方法，并在 NER 上取得了较好的效果。比较常见的是模型的混合，如混合多个 SVM[107]、混合 HMM 和 ME[108]。统计和规则相结合的方法[109-110]也较为多见。有研究将 NER 与命名实体归一化（NEN）相结合，也取得了很好的效果[111]。

5 NER 的新阶段

5.1 跨语言 NER 研究

随着大数据时代的推进，越来越多大规模跨语言平行语料库的得到建立，跨语言 NER 逐渐得到了重新重视，有利用平行语料库和元数据的跨语言 NER 研究[112]，也有研究提出了不需要平行语料库[113]。有研究利用维基百科的多语言知识映射来提高跨语言 NER 效果[114-115]，还有研究将双语词语对齐和 NER 相结合[116]，发现结合后的方法对双语下的词语对齐和 NER 都带来了显著的提高。也有研究直接使用跨语言的协同训练算法来实现跨语言命名实体的识别[117]。

5.2 深度学习下的 NER

近年来，源于神经网络模型的深度学习技术成为机器学习领域新的热潮。尤其是使用词向量来表示词语的方法，一方面解决了高纬度向量空间带来的数据稀疏问题，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，对于 NER 这种典型的序列化标注问题，俨然能够带来强大的发展动力。因此，虽然 NER 已经不是命名实体众多相关研究领域中的热点，仍有不少学者将最新的深度学习技术使用在 NER 问题上，以求进一步提高 NER 的效果。这其中，使用词向量作为特征，是最为简单有效的方法[118]。而更多研究还力求借鉴借鉴和改进现有的模型和方法，如有研究借鉴 LSTM 在自动分词上得到的较好结果，提出一种 LSTM 与 CRF 相结合的模型，比之前的方法的 F 值提高了 5%[119]。Tomori 等[120]为了证明现实世界的数据可以用来提高 NER 的效果，利用日本将棋比赛的解说语料库和棋局数据，训练了一个 DNN+R 模型，发现该模型比单纯的 DNN 模型效果好很多。Lample 等[121]提出了 LSTM 和基于转换的两种神经网络模型，同时从标注语料和未标注语料中获取特征，在四种语言上均获得了目前最好的 NER 效果。Bharadwaj 等[122]在 LSTM 神经网络上，增加了一层音素特征，在土耳其语等形态变化较复杂的语言上取得了较好了 NER 效果。除此之外，卷积神经网络（CNN）[123]、混合神经网络（HNN）[124]等深度学习方法也被成功用来解决 NER 问题，并取得了较好的结果。

5.3 NER 的应用

随着 NER 效果不断提高，技术逐渐成熟，如今的 NER 研究重点逐渐从模型调整等转向了实际应用。这主要得益于机器学习方法下的 NER 效果，虽然还无法达到接近 100%的正确率和召回率，但绝大

① 也叫 Non-Local。

多数方法的效果已经能达到 80%~90%的 F1 值，这对于从大规模的文本中识别出命名实体来说，已经足以满足一定的应用需求。因此 NER 开始在各类学科、各领域文本中得到大量的尝试，这其中，生物医学领域中各种实体的识别的研究最为成熟和成功[125-128]。近年来社交媒体火热发展，对社交媒体的情感评价、网络分析研究成为 NLP 一个重要的研究领域，而从 Twitter[129-131]、微博[132-133]等社交媒体文本中识别出命名实体对于这类研究来说是一个重要的基础，因此社交媒体 NER 研究近几年一直保持着一定程度的研究热度。但值得注意的是，由于社交媒体文本的语句随意性较大，相对于评测会议的语料来说，语料质量要差很多，对于基于机器学习的 NER 效果会产生较大的影响，这种影响也直接反映在目前社交媒体 NER 的效果较之评测会议有着较大的差距。除了生物医学和社交媒体之外，NER 在其他诸多领域中，都发挥了积极的效果，如 E-Mail[134]、化学实体[135]、旅游实体[136]、商品商标名称[137-138]、古籍文本中的人名[139]、地名[140-141]等。除此之外，NER 还被用来帮助解决一些特殊的 NLP 问题，比如提高苏美尔语的词形还原效果[142]。

6　NER 的总结和未来发展

命名实体识别从提出以来，一直是信息检索、数据挖掘、自然语言处理等领域中一个重要的研究领域。从 MUC 到 ACE 再到 CoNLL，一系列重要的评测会议划定了 NER 的基本研究范围，也提出了大量经典的重要的研究方法。与大多数 NLP 问题类似，NER 的发展基本经历了一种从规则向统计的转向。早期的规则方法已经不再流行，但其研究思路仍然给人以宝贵的启示，且规则和统计相结合的方法仍不时得到有效的尝试。如今的 NLP 领域，统计机器学习方法日臻完善，NER 也在这辆高速列车上走向成熟，而深度学习带来的机器学习新热潮，将会使 NER 在统计机器学习的道路上继续高速地推进。

然而从目前已有的研究成果来看，NER 研究还远不是一个得到完善解决甚至将要完善解决的问题，各领域下对命名实体定义的模糊，实验结果在 80%~90%徘徊的 F1 值，使得 NER 仍然是一个有挑战性的研究领域。一方面，大数据环境下，机器学习乃至深度学习仍将是最有效的 NER 方法。而另一方面，虽然机器学习带来了 NER 的火热发展，但大量研究固化于调整经典模型、挑选更多特征、扩大语料规模这种三角模式，是值得研究者们反思的。NER 的研究不应局限于 F1 值的提高上，从更多的角度来思考 NER 这一问题，才能使这个研究领域获得更全面的发展。比如当语料规模不足时，不是考虑扩大语料的规模，而是使用迁移学习方法来解决[143]。

NER 在其他学科上的应用也是未来一个重要的研究方向。将已有的 NER 方法有效地应用在各种领域的文本上，帮助各种学科获取其所关注的命名实体，这本就是 NER 研究的意义和价值所在。

参　考　文　献

[1] Chinchor N. MUC-6 named entity task definition (version 2.1)[C]// Proceedings of the 6th Conference on Message Understanding, Columbia, Maryland, 1995.

[2] Chinchor N, Robinson P. MUC-7 named entity task definition[C]// Proceedings of the 7th Conference on Message Understanding, 1997.

[3] LDC. Entity detection and tracking-phase 1 ace pilot study task definition[EB/OL]. [2017-03-10]. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/edt-phase1-v2.2.pdf.

[4] LDC. Annotation guidelines for entity link tracking (LNK) Version 3.0 20040401[EB/OL]. [2017-03-10]. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-lnk-v3.0.PDF.

[5] Sang E F T K. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition[C]// Proceedings of the 6th Conference on Natural language Learning, 2002: 1-4.

[6] Sang E F T K, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, Association for Computational Linguistics, 2003: 142-147.

[7] Levow G. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, 2006: 108-117.

[8] Jin G, Chen X. The fourth international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and chinese pos tagging[C]// Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, 2008: 69-81.

[9] Petasis G, Cucchiarelli A, Velardi P, et al. Automatic adaptation of Proper Noun Dictionaries through cooperation of machine learning and probabilistic methods[C]// Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2000: 128-135.

[10] Alfonseca E, Manandhar S. An unsupervised method for general named entity recognition and automated concept discovery[C]// Proceedings of the 1st International Conference on General

WordNet, Mysore, India, 2002: 34-43.

[11] Sekine S, Sudo K, Nobata C. Extended named entity hierarchy[C]// Proceedings of the Third International Conference on Language Resources and Evaluation, 2002: 1818-1824.

[12] Sekine S, Nobata C. Definition, dictionaries and tagger for extended named entity hierarchy[C]// Proceedings of the International Conference on Language Resources and Evaluation, 2004: 1977-1980.

[13] Borrega O, Taulé M, Martí M A. What do we mean when we speak about Named Entities[C]// Proceedings of Corpus Linguistics, 2007.

[14] Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1): 3-26.

[15] Kripke S A. Naming and necessity[M]// Semantics of Natural Language. Springer, 1972: 253-355.

[16] Marrero M, Urbano J, Sánchez-Cuadrado S, et al. Named entity recognition: fallacies, challenges and opportunities[J]. Computer Standards & Interfaces, 2013, 35(5): 482-489.

[17] Rau L F. Extracting company names from text[C]// Proceedings of the Seventh IEEE Conference on Artificial Intelligence Applications, IEEE, 1991: 29-32.

[18] Krupka G R. SRA: Description of the SRA system as used for MUC-6[C]// Proceedings of the 6th Conference on Message Understanding. Stroudsburg: Association for Computational Linguistics, 1995: 221-235.

[19] Sundheim B M. Overview of results of the MUC-6 evaluation[C]// Proceedings of a Workshop on Held at Vienna, Virginia. Stroudsburg: Association for Computational Linguistics, 1996: 423-442.

[20] Merchant R, Okurowski M E, Chinchor N. The multilingual entity task (MET) overview[C]// Proceedings of a Workshop on Held at Vienna, Virginia. Stroudsburg: Association for Computational Linguistics, 1996: 445-447.

[21] Chinchor N. Overview of muc-7/met-2[R]. Science Applications International Corp San Diego, CA, 1998.

[22] Mikheev A, Grover C, Moens M. Description of the LTG system used for MUC-7[C]// Proceedings of 7th Message Understanding Conference, Fairfax, VA, 1998: 1-12.

[23] Yu S, Bai S, Wu P. Description of the Kent Ridge Digital Labs system used for MUC-7[C]// Proceedings of the Seventh Message Understanding Conference, 1998: 1-16.

[24] Weischedel R. BBN: description of the PLUM system as used for MUC-6[C]// Proceedings of the 6th Conference on Message Understanding. Stroudsburg: Association for Computational Linguistics, 1995: 55-69.

[25] Aberdeen J, Burger J, Day D, et al. MITRE: description of the alembic system used for MUC-6[C]// Proceedings of the 6th Conference on Message Understanding. Stroudsburg: Association for Computational Linguistics, 1995: 141-155.

[26] Appelt D E, Hobbs J R, Bear J, et al. SRI International FASTUS system: MUC-6 test results and analysis[C]// Proceedings of the 6th Conference on Message Understanding. Stroudsburg: Association for Computational Linguistics, 1995: 237-248.

[27] Grishman R. The NYU system for MUC-6 or where's the syntax?[C]// Proceedings of the 6th Conference on Message Understanding, Association for Computational Linguistics, 1995: 167-175.

[28] Black W J, Rinaldi F, Mowatt D. FACILE: Description of the NE system used for MUC-7[C]// Proceedings of the 7th Message Understanding Conference, 1998.

[29] Humphreys K, Gaizauskas R, Azzam S, et al. University of sheffield: Description of the LaSIE-II system as used for MUC-7[C]// Proceedings of the Seventh Message Understanding Conferences, 1998.

[30] Chen H, Ding Y, Tsai S, et al. Description of the NTU system used for MET2[C]// Proceedings of 7th Message Understanding Conference, 1998.

[31] 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 Seventh Message Understanding Conference, 1998.

[32] Strassel S, Mitchell A, Huang S. Multilingual resources for entity extraction[C]// Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition, Stroudsburg: Association for Computational Linguistics, 2003: 49-56.

[33] Doddington G R, Mitchell A, Przybocki M A, et al. The automatic content extraction program-tasks, data, and evaluation[C]// Proceedings of the International Conference on Language Resources and Evaluation, 2004: 1.

[34] LDC. Entity detection and tracking – Phase 1 EDT and metonymy annotation guidelines[EB/OL]. [2017-03-10]. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/edt-guidelines-v2-5.pdf.

[35] LDC. Annotation guidelines for relation detection and characterization (RDC) Version 3.6-4.22.2002[EB/OL]. [2017-03-10]. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/rdc-guidelines-v3.6.pdf.

[36] NIST. Automatic content extraction 2003 evaluation[EB/OL]. [2017-03-10]. http://itl.nist.gov/iad/mig/tests/ace/2003/.

[37] NIST. The ACE 2004 evaluation plan evaluation of the recognition of ACE entities, ACE relations and ACE events[EB/OL]. [2017-03-10]. http://itl.nist.gov/iad/mig/tests/ace/2004/doc/ace04-evalplan-v7.pdf.

[38] LDC. ACE (Automatic Content Extraction) English annotation guidelines for values[EB/OL]. [2017-03-10]. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-values-guidelines-v1.2. 4.pdf.

[39] LDC. Timestamping of ACE relations and events for 2005[EB/OL]. [2017-03-10]. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-timestamping-guidelines-v3.pdf.

[40] LDC. ACE (Automatic Content Extraction) English annotation

guidelines for events[EB/OL]. [2017-03-10]. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf.

[41] NIST. The ACE 2005 evaluation plan evaluation of the detection and recognition of ACE entities, values, temporal expressions, relations, and events[EB/OL]. [2017-03-10]. http://itl.nist.gov/iad/mig/tests/ace/2005/doc/ace05-evalplan.v3.pdf.

[42] LDC. ACE (Automatic Content Extraction) Spanish annotation guidelines for entities[EB/OL]. [2017-03-10]. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/spanish-entities-guidelines-v1.6.pdf.

[43] LDC. GALE Arabic translation guidelines V2.3[EB/OL]. [2017-03-10]. http://projects.ldc.upenn.edu/gale/Translation/specs/GALE_Arabic_translation_guidelines_v2.3.pdf.

[44] LDC. GALE Chinese translation guidelines V2.3[EB/OL]. [2017-03-10]. http://projects.ldc.upenn.edu/gale/Translation/specs/GALE_Chinese_translation_guidelines_v2.3.pdf.

[45] NIST. The ACE 2007 evaluation plan evaluation of the detection and recognition of ACE entities, values, temporal expressions, relations, and events[EB/OL]. [2017-03-10]. http://itl.nist.gov/iad/mig/tests/ace/2007/doc/ace07-evalplan.v1.3a.pdf.

[46] LDC. ACE 2008: Cross-document annotation guidelines (XDOC)[EB/OL]. [2017-03-10]. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/ace08-xdoc-1.6.pdf.

[47] NIST. Automatic content extraction 2008 evaluation plan assessment of detection and recognition of entities and relations within and across documents[EB/OL]. [2017-03-10]. http://itl.nist. gov/iad/mig/tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf.

[48] Burger J D, Henderson J C, Morgan W T. Statistical named entity recognizer adaptation[C]// Proceedings of the 6th Conference on Natural Language Learning, Stroudsburg: Association for Computational Linguistics, 2002, 20: 1-4.

[49] Chieu H L, Ng H T. Named entity recognition with a maximum entropy approach[C]// Conference on Natural Language Learning at HLT-NAACL, 2003: 160-163.

[50] Curran J R, Clark S. Language independent NER using a maximum entropy tagger[C]// Conference on Natural Language Learning at HLT-NAACL, 2003: 164-167.

[51] Mayfield J, McNamee P, Piatko C. Named entity recognition using hundreds of thousands of features[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, Stroudsburg: Association for Computational Linguistics, 2003: 184-187.

[52] McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, Stroudsburg: Association for Computational Linguistics, 2003,4: 188-191.

[53] Carreras X, Marquez L, Padró L. Named entity extraction using adaboost[C]// Proceedings of the 6th Conference on Natural Language Learning, Stroudsburg: Association for Computational Linguistics, 2002, 20: 1-4.

[54] Zhang T, Johnson D. A robust risk minimization based named entity recognition system[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, Stroudsburg: Association for Computational Linguistics, 2003, 4: 204-207.

[55] Carreras X, Màrquez L, Padró L. Learning a perceptron-based named entity chunker via online recognition feedback[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, Stroudsburg: Association for Computational Linguistics, 2003, 4: 156-159.

[56] Hammerton J. Named entity recognition with long short-term memory[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, Stroudsburg: Association for Computational Linguistics, 2003, 4: 172-175.

[57] Black W J, Vasilakopoulos A. Language independent named entity classification by modified transformation-based learning and by decision tree induction[C]// proceedings of the 6th Conference on Natural Language Learning, Stroudsburg: Association for Computational Linguistics, 2002, 20: 1-4.

[58] Florian R. Named entity recognition as a house of cards: Classifier stacking[C]// Proceedings of the 6th Conference on Natural Language Learning, Stroudsburg: Association for Computational Linguistics, 2002, 20: 1-4.

[59] Florian R, Ittycheriah A, Jing H, et al. Named entity recognition through classifier combination[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, Stroudsburg: Association for Computational Linguistics, 2003, 4: 168-171.

[60] Klein D, Smarr J, Nguyen H, et al. Named entity recognition with character-level models[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, Stroudsburg: Association for Computational Linguistics, 2003, 4: 180-183.

[61] 赵军. 命名实体识别、排歧和跨语言关联[J]. 中文信息学报, 2009, 23(2): 3-17.

[62] Chang J, Chen S, Chen Y, et al. A multiple-corpus approach to identification of Chinese surname-names[C]// Proceedings of Natural Language Processing Pacific Rim Symposium, 1991: 87-91.

[63] Wang L, Li W, Chang C. Recognizing unregistered names for mandarin word identification[C]// Proceedings of the 14th Conference on Computational Linguistics, Stroudsburg: Association for Computational Linguistics, 1992, 4: 1239-1243.

[64] 张俊盛, 陈舜德, 郑萦, 等. 多语料库作法之中文姓名辨识[J]. 中文信息学报, 1992, 6(3): 9-17.

[65] 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): 21-32.

[66] 宋柔, 朱宏. 基于语料库和规则库的人名识别法[C]// 全国第

二届计算机语言学联合学术会议. 北京: 北京语言学院出版社, 1993: 150-154.

[67] 郑家恒, 刘开瑛. 自动分词系统中姓氏人名处理策略探讨[C]// 全国第二届计算机语言学联合学术会议. 北京: 北京语言学院出版社, 1993: 139-143.

[68] 孙茂松, 黄昌宁, 高海燕, 等. 中文姓名的自动辨识[J]. 中文信息学报, 1995, 9(2): 16-27.

[69] 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, 26(6): 42-47.

[70] Zhou J S, He L, Dai X Y, et al. Chinese Named Entity Recognition with a Multi-Phase Model[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics, 2006: 213-216.

[71] Chen A, Peng F, Shan R, et al. Chinese named entity recognition with conditional probabilistic models[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics, 2006: 173-176.

[72] Chen W L, Zhang Y J, Isahara H. Chinese named entity recognition with conditional random fields[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics, 2006: 118-121.

[73] Mao X N, Dong Y, He S K, et al. Chinese word segmentation and named entity recognition based on conditional random fields[C]// Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, 2008: 90-93.

[74] Zhao H, Kit C. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition[C]// Proceedings of the Fourth International Chinese Language Processing Bakeoff & the First CIPS Chinese Language Processing Evaluation, 2008: 106-111.

[75] Yu X F, Lam W, Chan S K, et al. Chinese NER using crfs and logic for the fourth SIGHAN bakeoff[C]// Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, 2008: 102-105.

[76] Zhang S X, Qin Y, Wen J, et al. Word segmentation and named entity recognition for SIGHAN Bakeoff3[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics, 2006: 158-161.

[77] Chen Y, Jin P, Li W, et al. The Chinese persons name disambiguation evaluation: Exploration of personal name disambiguation in Chinese news[C]// Processings of the CIPS-SIGHAN Joint Conference on Chinese Language, 2010.

[78] He Z, Wang H, Li S. The Task 2 of CIPS-SIGHAN 2012 named entity recognition and disambiguation in Chinese Bakeoff[C]// Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing. 2012: 108-114.

[79] Zong H, Wong D F, Chao L S. A template based hybrid model for Chinese personal name disambiguation[C]// Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, Tianjin, China, 2012: 121-126.

[80] Tian W, Pan X, Yu Z T, et al. Chinese name disambiguation based on adaptive clustering with the attribute features[C]// Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, Tianjin, China, 2012: 132-137.

[81] Collins M, Singer Y. Unsupervised models for named entity classification[C]// Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999: 100-110.

[82] Cucerzan S, Yarowsky D. Language independent named entity recognition combining morphological and contextual evidence[C]// Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC, 1999: 90-99.

[83] Mikheev A, Moens M, Grover C. Named entity recognition without gazetteers[C]// Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1999: 1-8.

[84] 张晓艳, 王挺, 陈火旺. 命名实体识别研究[J]. 计算机科学, 2005, 32(4): 44-48.

[85] Bikel D M, Miller S, Schwartz R, et al. Nymble: a high-performance learning name-finder[C]// Proceedings of the Fifth Conference on Applied Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 1997: 194-201.

[86] Bikel D M, Schwartz R, Weischedel R M. An algorithm that learns what's in a name[J]. Machine Learning, 1999, 34(1-3): 211-231.

[87] Borthwick A E. A maximum entropy approach to named entity recognition[D]. New York: New York University, 1999.

[88] Isozaki H, Kazawa H. Efficient support vector classifiers for named entity recognition[C]// Proceedings of the 19th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2002, 1: 1-7.

[89] Zhou G, Su J. Named entity recognition using an HMM-based chunk tagger[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2002: 473-480.

[90] Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov Models[J]. Bioinformatics, 2016, 32(18): 2839-2846.

[91] 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.

[92] 胡文博, 都云程, 吕学强, 等. 基于多层条件随机场的中文命名实体识别[J]. 计算机工程与应用, 2009, 45(1): 163-165.

[93] 周俊生, 戴新宇, 尹存燕, 等. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报, 2006, 34(5): 804-809.

[94] Li C, Liu Y. Improving named entity recognition in Tweets via detecting non-standard words[C]// Proceedings of the 53rd An-

nual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 929-938.

[95] 冯元勇, 孙乐, 李文波, 等. 基于单字提示特征的中文命名实体识别快速算法[J]. 中文信息学报, 2008, 22(1): 104-110.

[96] 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.

[97] Luo W, Yang F. An empirical study of automatic Chinese word segmentation for spoken language understanding and named entity recognition[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies, San Diego, California, 2016: 238-248.

[98] 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 the 44th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2006: 1121-1128.

[99] 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.

[100] Chieu H L, Ng H T. Named entity recognition: a maximum entropy approach using global information[C]// Proceedings of the 19th International Conference on Computational linguistics. Stroudsburg: Association for Computational Linguistics, 2002, 1: 1-7.

[101] Miller S, Guinness J, Zamanian A. Name tagging with word clusters and discriminative training[C]// Proceedings of HLT-NAACL, 2004: 337-342.

[102] Ratinov L, Roth D. Design challenges and misconceptions in named entity recognition[C]// Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2009: 147-155.

[103] Kazama J I, Torisawa K. Exploiting Wikipedia as external knowledge for named entity recognition[C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007: 698-707.

[104] Richman A E, Schone P. Mining Wiki resources for multilingual named entity recognition[C]// Proceedings of ACL-08: HLT. Stroudsburg: Association for Computational Linguistics, 2008: 1-9.

[105] Nothman J, Ringland N, Radford W, et al. Learning multilingual named entity recognition from Wikipedia[J]. Artificial Intelligence, 2013, 194: 151-175.

[106] 郑逢强, 林磊, 刘秉权, 等. 《 知网》 在命名实体识别中的应用研究[J]. 中文信息学报, 2008, 22(5): 97-101.

[107] Li L, Mao T, Huang D, et al. Hybrid models for Chinese named entity recognition[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics, 2006: 72-78.

[108] 张晓艳, 王挺, 陈火旺. 基于混合统计模型的汉语命名实体识别方法[J]. 计算机工程与科学, 2006, 28(6): 135-139.

[109] 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, 25(10): 2404-2406.

[110] 潘正高. 基于规则和统计相结合的中文命名实体识别研究[J]. 情报科学, 2012, 30(5): 708-712.

[111] Liu X H, Zhou M, Wei F R, et al. Joint inference of named entity recognition and normalization for Tweets[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2012: 526-535.

[112] Kim S, Toutanova K, Yu H. Multilingual named entity recognition using parallel data and metadata from Wikipedia[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2012: 694-702.

[113] Zirikly A, Hagiwara M. Cross-lingual transfer of named entity recognizers without parallel corpora[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 390-396.

[114] Darwish K. Named entity recognition using cross-lingual resources: Arabic as an example[C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2013: 1558-1567.

[115] 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. Stroudsburg: Association for Computational Linguistics, 2016: 1275-1284.

[116] Wang M, Che W, Manning C D. Joint word alignment and bilingual named entity recognition using dual decomposition[C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2013: 1073-1082.

[117] Li Y, Huang H, Zhao X, et al. Named Entity recognition based on bilingual co-training[M]// Chinese Lexical Semantics. Berlin: Springer, 2013: 480-489.

[118] Cherry C, Guo H Y. The unreasonable effectiveness of word representations for Twitter named entity recognition[C]// The 2015 Annual Conference of the North American Chapter of the ACL. Stroudsburg: Association for Computational Linguistics, 2015: 735-745.

[119] 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. Stroudsburg: Association for Computational Linguistics, 2016: 149-155.

[120] Tomori S, Ninomiya T, Mori S. Domain specific named entity recognition referring to the real world by deep neural networks[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016: 236-242.

[121] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2016: 260-270.

[122] Bharadwaj A, Mortensen D, Dyer C, et al. Phonologically aware neural model for named entity recognition in low resource transfer settings[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 1462-1472.

[123] Dong X S, Qian L J, Guan Y, et al. A multiclass classification method based on deep learning for named entity recognition in electronic medical records[C]// Proceedings of the 2016 New York Scientific Data Summit. IEEE, 2016: 1-10.

[124] Shao Y, Hardmeier C, Nivre J. Multilingual named entity recognition using hybrid neural networks[C]// Proceedings of the Sixth Swedish Language Technology Conference, 2016.

[125] Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets[C]// Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. Stroudsburg: Association for Computational Linguistics, 2004: 104-107.

[126] Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition[C]// Pacific Symposium on Biocomputing, 2008: 652-663.

[127] Tang B Z, Cao H X, Wang X L, et al. Evaluating word representation features in biomedical named entity recognition tasks[J]. BioMed Research International, 2014, 2014: Article ID 240403.

[128] Kazama J, Makino T, Ohta Y, et al. Tuning support vector machines for biomedical named entity recognition[C]// Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain. Stroudsburg: Association for Computational Linguistics, 2002: 1-8.

[129] Ritter A, Clark S, Etzioni O. Named entity recognition in tweets: an experimental study[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011: 1524-1534.

[130] Liu X, Zhang S, Wei F, et al. Recognizing named entities in Tweets[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2011: 359-367.

[131] Li C, Weng J, He Q, et al. Twiner: named entity recognition in targeted twitter stream[C]// Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2012: 721-730.

[132] 邱泉清, 苗夺谦, 张志飞. 中文微博命名实体识别[J]. 计算机科学, 2013, 40(6): 196-198.

[133] Peng N, Dredze M. Named entity recognition for Chinese social media with jointly trained embeddings[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics 2015: 548-554.

[134] Minkov E, Wang R C, Cohen W W. Extracting personal names from email: Applying named entity recognition to informal text[C]// Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2005: 443-450.

[135] Leaman R, Wei C, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization[J]. Journal of Cheminformatics, 2015, 7: S3.

[136] 郭剑毅, 薛征山, 余正涛, 等. 基于层叠条件随机场的旅游领域命名实体识别[J]. 中文信息学报, 2009, 23(5): 47-52.

[137] 刘非凡, 赵军, 吕碧波, 等. 面向商务信息抽取的产品命名实体识别研究[J]. 中文信息学报, 2006, 20(1): 7-13.

[138] Putthividhya D, Hu J L. Bootstrapped named entity recognition for product attribute extraction[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011: 1557-1567.

[139] 汤亚芬. 先秦古汉语典籍中的人名自动识别研究[J]. 现代图书情报技术, 2013, 29(7-8): 63-68.

[140] 朱锁玲, 包平. 方志类古籍地名识别及系统构建[J]. 中国图书馆学报, 2011, 37(3): 118-124.

[141] 黄水清, 王东波, 何琳. 基于先秦语料库的古汉语地名自动识别模型构建研究[J]. 图书情报工作, 2015, 59(12): 135-140.

[142] Liu Y D, Burkhart C, Hearne J, et al. Enhancing sumerian lemmatization by unsupervised named-entity recognition[C]// Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL. Stroudsburg: Association for Computational Linguistics, 2015: 1446-1451.

[143] Qu L, Ferraro G, Zhou L, et al. Named Entity Recognition for Novel Types by Transfer Learning[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 899-905.

（责任编辑 魏瑞斌）