文章编号：1003-0077(2021)09-0030-16

基于深度学习的命名实体识别综述

邓依依，邬昌兴，魏永丰，万仲保，黄兆华

（华东交通大学 软件学院，江西 南昌 330013）

摘　要：命名实体识别是自然语言处理的基础任务之一，目的是从非结构化的文本中识别出所需的实体及类型，其识别的结果可用于实体关系抽取、知识图谱构建等众多实际应用。近些年，随着深度学习在自然语言处理领域的广泛应用，各种基于深度学习的命名实体识别方法均取得了较好的效果，其性能全面超越传统的基于人工特征的方法。该文从三个方面介绍近期基于深度学习的命名实体识别方法：第一，从输入层、编码层和解码层出发，介绍命名实体识别的一般框架；第二，分析汉语命名实体识别的特点，着重介绍各种融合字词信息的模型；第三，介绍低资源的命名实体识别，主要包括跨语言迁移方法、跨领域迁移方法、跨任务迁移方法和集成自动标注语料的方法等。最后，总结相关工作，并提出未来可能的研究方向。

关键词：命名实体识别；汉语命名实体识别；低资源命名实体识别；深度学习

中图分类号：TP391　　**文献标识码**：A

A Survey on Named Entity Recognition Based on Deep Learning

DENG Yiyi, WU Changxing, WEI Yongfeng, WAN Zhongbao, HUANG Zhaohua

(School of Software, East China Jiaotong University, Nanchang, Jiangxi 330013, China)

Abstract: Named entity recognition (NER), as one of the basic tasks in natural language processing, aims to identify the required entities and their types in unstructured text. In recent years, various named entity recognition methods based on deep learning have achieved much better performance than that of traditional methods based on manual features. This paper summarizes recent named entity recognition methods from the following three aspects: 1) A general framework is introduced, which consists of an input layer, an encoding layer and a decoding layer. 2) After analyzing the characteristics of Chinese named entity recognition, this paper introduces Chinese NER models which incorporate both character-level and word-level information. 3) The methods for low-resource named entity recognition are described, including cross-lingual transfer methods, cross-domain transfer methods, cross-task transfer methods, and methods incorporating automatically labeled data. Finally, the conclusions and possible research directions are given.

Keywords: named entity recognition(NER); Chinese NER; low-resource NER; deep learning

0　引言

随着数据的爆炸式增长，人工从海量的文本中寻找有用的信息无疑是一项费时费力的任务，因此信息抽取研究应运而生。作为其关键技术之一的命名实体识别(Named Entity Recognition，NER)多年来受到学术领域和工业界的广泛关注。命名实体识别同时也是众多自然语言处理(Natural Language Processing，NLP)应用的基础，如实体关系抽取、知识图谱构建和智能问答等。

命名实体识别任务旨在从非结构化的文本中自动识别出所需的实体，并将其标记为预定义的类别，例如人名、地名和组织机构名等。该任务于 1995 年在第六届 MUC(the Sixth Message Understanding Conference，MUC-6)会议上首次被提出[1]。随后，

收稿日期：2020-08-24　　**定稿日期**：2020-11-16

基金项目：国家重点研发计划(2018YFC0831106)；国家自然科学基金(61866012)；江西省自然科学基金(20181BAB202012)；江西省教育厅科学技术研究项目(GJJ180329)

命名实体识别的研究在不同语言和不同领域中得到广泛开展，关注的实体从人名等通用实体扩展到包含疾病名等领域特定实体，实体类别的数量从几种到上百种不等。表 1 列出了英语和汉语中常用于 NER 模型训练和性能评估的数据集。这些数据集来自新闻、财经和生物医学等领域，涉及的文体包括规范的新闻文本、维基百科文本和用户生成网络文本等。从当前的研究情况来看，虽然大量的研究工作针对语料资源丰富的语言（例如，英语）和领域（例如，生物医学）展开，但近年来低资源语言和领域的命名实体识别受到越来越多的关注，而汉语命名实体识别则一直受到国内研究人员的高度重视。从识别性能来看，基于深度学习的命名实体识别方法在规范的文本上（如 CoNLL2003 数据集[2]）识别人名、地名和机构名的 F_1 值达到了 93.3%[3]，可以满足基本应用的要求。但是，在用户生成网络文本数据集 W-NUT17 上，Lin 等人的实验结果显示，F_1 值不到 50%[4]，远未达到实用的要求。因此，命名实体识别依然是一个重要且值得深入研究的课题。

表 1　常用的命名实体识别数据集

数据集	语种	类别数	来源或文体
MUC-6/7(1995,1997)	英语	7	华尔街新闻等
CoNLL2003(2003)	英语	4	路透社新闻
OntoNotes5.0(2013)	英语	18	电话语音、新闻等
JNLPBA2004(2004)	英语	5	生物医学
GENIA(2004)	英语	36	生物医学
WiNER(2012)	英语	4	维基百科
WikiFiger(2012)	英语	112	维基百科
W-NUT(15-18)(2015-2018)	英语	6/10	用户生成网络文本
OntoNotes4.0(2011)	汉语	18	新闻专线、广播等
MSRA(2006)	汉语	3	新闻
Weibo(2015)	汉语	4	新浪微博
Resume(2018)	汉语	8	新浪财经

非重叠（非嵌套）命名实体识别通常被建模成一个序列标注任务，即给序列（句子）中的每个字或词指定一个标签。如图 1 所示，根据常用的 BIO 标注模式，通过预测输入句子中每个词的标签，并连接相应的标签就可得出该句子中实体的边界及类型（<1-1，地名，武汉市>、<2-3，机构名，长江医院>、<5-5，人名，王林>）。其中，B 代表命名实体的起始词，I 代表实体的非起始词，O 为其他字符；B-Org代表机构名的起始词，I-Org 代表机构名的非起始词，B-Per 代表人名的起始词。近年来，重叠（嵌套）命名实体识别受到越来越多研究者的关注。重叠命名实体是一种特殊的命名实体，即在一个实体的内部还存在着一个或多个其他的实体，例如组织机构名“武汉大学”中的“武汉”也是地名。重叠命名实体的识别难以直接使用上述基于序列标注的模型，而通常采用层叠式模型[5]或基于区域的模型[6]。本文主要介绍非重叠命名实体识别的相关研究，在不造成理解歧义的情况下，后文把非重叠命名实体（识别）简称为命名实体（识别）。

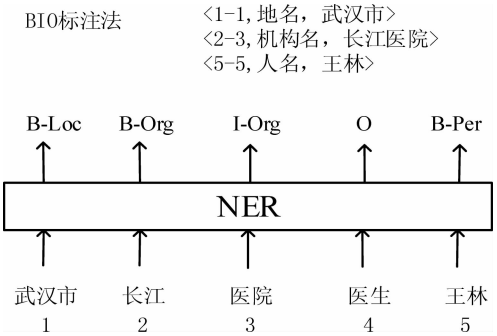

图 1　命名实体识别任务实例

早期的命名实体识别方法主要包括基于规则的方法和基于人工特征的方法。当制定的规则能较准确地反映出文本的特性时，基于规则的方法往往能取得较高的准确率，但这极大地依赖于语言学家的专业知识，且有限的规则难以将变化无穷的实体较全面地识别出来[7]。而后，随着统计机器学习算法在 NLP 领域的广泛使用，基于人工特征的方法取得了比基于规则的方法更好的性能。这类方法通常基于大量人工定义的特征，使用隐马尔可夫模型（Hidden Markov Model，HMM）[8]或条件随机场（Conditional Random Fields，CRF）[9]在大量人工标注的语料上训练命名实体识别模型。基于人工特征的方法通过统计机器学习算法从大量标注语料中学习知识，而不再需要人工定义的规则。这类方法的不足之处主要包括：①需要人工定义能反映实体特性的特征集合，方法的性能主要依赖于所采用的特征是否具有识别度；②对标注语料的依赖性也较强，需要在大量人工标注的语料上训练模型，而构建大规模标注语料库是一项费时费力的事情。

近年来，基于深度学习的方法广泛应用于自然语言处理领域中，在多数任务上都取得了较好的效果[10-13]。与早期的统计机器学习方法相比，基于深度学习的方法在自动学习特征、运用深层次语义知

识和缓解数据稀疏问题等方面具有明显的优势。具体表现在：①可以自动学习特定于任务的分布式特征，从而避免了需要人工定义特征的问题；②可以自动学习词、短语和句子等不同粒度语言单位的语义向量表示，从而有利于深层次语义的理解和计算；③从数据稀疏的角度看，自动学习的分布式特征的低维连续向量表示也优于人工定义特征的高维离散向量表示；④能够方便地整合并迁移来自各种异构数据源的信息，从而有效地缓解低资源语言和领域人工标注语料短缺的问题。就命名实体识别而言，研究人员探索了大量基于深度学习的方法，并取得了实质性的进展。这类方法借助神经网络自动学习特征并训练序列标注模型，性能超过了传统的基于人工特征的方法，是当前的研究热点之一。为了方便这个领域的研究人员和应用人员，本文对近年来基于深度学习的命名实体识别（NER）的研究工作进行梳理，将其大致分为以下三大类进行介绍：

（1）NER 的一般框架：从输入层、编码层和解码层出发，介绍主流命名实体识别模型的一般框架，详细介绍现有工作中各层的典型实现，并分析它们的优缺点。

（2）汉语 NER：在阐述汉语命名实体识别的特点后，重点介绍各种融合字词信息[①]的模型，其既能够利用词的相关信息又能避免汉语分词可能带来的错误。

（3）低资源 NER：介绍低资源语言和领域的命名实体识别，主要包括跨语言迁移的方法、跨领域迁移的方法、跨任务迁移的方法以及集成自动标注语料的方法等，其能够有效地缓解人工标注语料短缺的问题。

基于深度学习的命名实体识别研究已经持续多年，现在已进入一个相对成熟的阶段，但该方向的综述还比较少，而有关汉语命名实体识别和低资源命名实体识别研究进展的介绍更是少之又少。Yadav 和 Bethard[14] 依据当前基于深度学习的命名实体识别模型中输入层表示的不同进行分类介绍，分为基于字符表示的模型、基于词表示的模型和基于字词混合表示的模型。Li 等[15] 指出上述综述更多地关注命名实体识别模型的输入，进而从提取字词信息的输入层、融合上下文信息的编码层和标记解码层出发介绍当前主流的模型。不同于现有综述的分类方式，本文首先介绍基于深度学习的命名实体识别模型的一般框架，然后重点介绍汉语命名实体识别及低资源语言和领域的命名实体识别的研究现状。目的是让读者对基于深度学习的命名实体识别研究进展有一个较为全面的了解，便于日后研究和应用工作的开展。

1 NER 的一般框架

基于深度学习的 NER 模型通常以词作为基本的标记单元，即为文本中的每个词预测一个标签，连接相应的标签就可得出该文本中实体的边界及类型。其一般框架如图 2(a)所示：输入层用于把词相关的信息表示为向量；编码层学习融合上下文信息的词的向量表示，该表示可以认为是特定于任务的

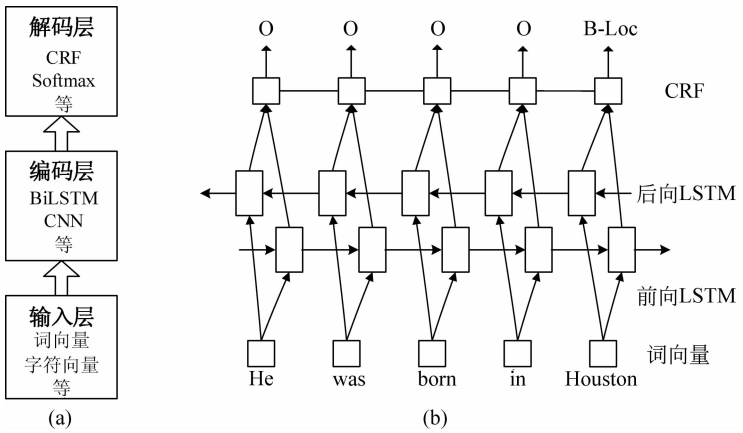

图 2 NER 的一般框架(a)及经典的 BiLSTM-CRF 模型(b)

① 在汉语中，词指分词处理后的文本单元，例如“医院”，而字即指单个汉字，例如，“医”和“院”；在英语等语言中，对应的是词(word)和字符(character)。

特征；解码层用于预测文本中每个词对应的标签。Huang 等[16]提出的基于 BiLSTM-CRF 的命名实体识别模型是极具代表性的工作之一，如图 2(b)所示。该模型的输入层仅使用预训练的词向量，不使用任何人工特征。编码层使用 BiLSTM(Bidirectional Long Short Term Memory)从两个方向建模词的上下文信息，前向 LSTM 从左至右学习词在上文中的表示，后向 LSTM 从右至左学习词在下文中的表示。解码时使用一个 CRF 层(conditional random field，CRF)利用标签之间的依赖关系，搜索最优的标签序列。例如，表示人名开头的标签 B-Per 后面不可能出现表示地名中间的标签 I-Loc。实验显示，上述基于 BiLSTM-CRF 的模型取得了与当时最好模型可比的性能。随后，基于深度学习的命名实体识别方法成为研究的热点，众多研究人员分别探索了不同的输入层、编码层和解码层对识别性能的影响。下面分别对相关的研究工作进行介绍。

1.1　输入层

除了最主要的词向量特征外，输入层还经常使用字符向量特征、形态学特征和基于实体词典(Gazetteer)的特征等作为补充信息。

字符向量特征被证实是非常通用且有效的信息，主要表现在以下两点：①可以显式地利用前缀和后缀等子词级(Sub-word Level)的特征；②可以很自然地缓解低频词的词向量质量不可靠、未登录词没有词向量的问题。形态学特征对词形丰富语言的 NER 非常有用，例如，屈折语系语言拉丁语和德语等，少数民族语言维吾尔语和哈萨克语等。

基于实体词典(Gazetteer)的特征主要是为了利用已有的地名词典、机构名词典以及药品名词典等，对特定领域的 NER 非常有效。例如，Lample 等[17]在输入层除了使用预训练的词向量外，还使用一个字符级的 BiLSTM 网络学习基于字符的词表示(Character-based Word Representations)作为补充。实验发现，与 Luo 等[18]提出的添加大量人工特征并且将实体识别与实体链接进行联合训练的复杂模型相比，Lample 等[17]提出的模型取得了可比的性能。更进一步，Chiu 和 Nichols[19]联合使用预训练的词向量、基于字符级 CNN (Conventional Neural Network)学习的词表示、字符的大小写特征和基于实体词典的特征作为模型的输入。Yadav 等[20]首次在基于深度学习的 NER 模型中融入词缀信息等形态学特征，在多种语言上获得了更好的性能。Lin 等[21]针对实体中经常包含低频词和未登录词的情况，设计了一种基于词频的可靠性机制，以灵活地选择和组合词向量特征和字符向量特征，在复杂数据集 OntoNotes 5.0[22]上其性能远超基线模型。

总的来说，在输入层中通常以词作为基本单位，即以预训练的词向量为主要特征，同时把字符特征和形态学等特征作为补充信息。

1.2　编码层

在编码层，常用的 BiLSTM 网络已表现出良好的序列建模能力，能较好地学习文本中词之间的依赖关系。然而，BiLSTM 也存在以下几个方面的缺陷：①序列中当前词的计算依赖于前一个词的计算结果，导致其不能并行计算，计算效率不如卷积神经网络(Convolution Neural Network，CNN)和基于注意力机制的 Transformer 网络[23]；②建模局部上下文(也称短距离的词之间的依赖)的能力不如 CNN；③理论上，BiLSTM 可以建模任意长距离的词之间的依赖，但实际中由于梯度消失问题，其建模长距离依赖的能力不如 Transformer 网络；④BiLSTM 没有考虑句子的结构信息。

基于上述原因，研究人员探索了不同网络结构的编码层对 NER 性能的影响。例如，Strubell 等[24]提出一种改进的 CNN 用作编码器，不但充分利用 CNN 捕获局部上下文的能力及其运算的可并行性，还通过层叠和允许不连续的输入等方式扩展 CNN 捕获长距离词之间依赖的能力。与经典的 BiLSTM -CRF 模型相比，其运算速度大幅提升，且达到了可比的性能。Chen 等[25]首先使用 CNN 建模词的局部上下文，然后层叠一种门控关系网络(Gated Relation Network)建模句子中词之间的长距离依赖关系，在 CoNLL2003 数据集上获得了高达 91.44%的 F_1 值。Li 等[26]提出基于双向递归神经网络(Bi-directional Recursive Neural Networks，BRNN)的模型以引入短语句法树信息；Jie 和 Lu[27]改进 BiLSTM-CRF 模型，进而利用依存句法树信息。虽然在编码层中利用句法信息可以提高 NER 的性能，但是在大多数语言中如何自动获取句子的句法树是一个问题。

近年来，Transformer 模型逐渐进入了大家的视野，其开创性地将自注意力机制(Self-Attention)作为编码器的核心，直接建模句子中任意距离的词之间的依赖。Transformer 虽然很快在机器翻译和语言模型等任务上展示了很好的性能，但 Guo 等[28]

的实验证实原始的 Transformer 在 NER 上的性能并不理想。随后，Yan 等[29]指出原始 Transformer 中的位置编码方式虽然能捕获词之间的距离信息，但不能得到词之间的前后关系信息，而这些信息对 NER 是非常重要的。基于这一问题，他们提出一种能同时感知距离和前后关系的注意力机制，用于改进 Transformer，在 NER 上取得了较好的效果。

从近几年的研究可以发现，BiLSTM 依然是 NER 模型中用得较多的编码层，可能的原因之一是其可以同时较好地建模词之间的短距离依赖（虽然不如 CNN）和长距离依赖（虽然不如 Transformer）。基于 CNN 或 Transformer 的编码层则具有可以并行计算、速度相对更快的特点。

1.3 解码层

在解码阶段，常用的 CRF 层不但考虑对应于每个词的分类标签的概率，还建模了相邻标签之间的依赖关系。CRF 解码层输出一个最优的标签序列，而不是单独为序列中的每个词预测一个标签，在多数序列标注任务（不仅是 NER）上都能取得较好的效果。然而，CRF 层也有以下方面的不足：①在输入序列较长或需要标记的实体类型较多时，速度较慢；②只建模了相邻的分类标签之间的依赖关系。

对于上述第一个不足，如果对性能要求不是很高，可以直接使用一个 Softmax 分类层为句子中的每个词单独解码。对于上述第二个不足，Shen 等[30]把命名实体识别看成是一个序列生成问题，他们基于 RNN 网络（Recurrent Neural Network）逐个生成句子中词的分类标签，并把前一个词的预测标签用作当前词的标签预测的输入。Shen 等[30]的模型不仅在性能上优于使用 CRF 解码的模型，且当标记的实体类型较多时其训练速度也更快。另外，Zhai 等[31]则首次应用指针网络（Pointer Networks）[32]来生成序列标签，其目的是利用已识别出的实体信息辅助预测，同样取得了较好的效果。虽然基于 RNN 等网络进行解码可以利用标签之间的长距离依赖关系，但在 NER 任务上的性能与 CRF 层相比并没有实质性的提升，可能的原因是命名实体标签之间更多的是局部的依赖关系。

目前，以 BERT 等基于超大规模语料预训练的语言模型为基础的 NER 模型在多个常用的数据集上取得了最佳的性能[3,11]，得益于超大规模语料中的知识，这类模型的性能远超上述以普通词向量等作为输入的 BiLSTM-CRF 模型。例如，在 BERT 上简单叠加一个 Softmax 分类层在 CoNLL2003 数据集上获得了高达 92.8% 的 F_1 值[11]。然而，这类 NER 模型的不足之处在于其规模太大，需要很强的计算能力，难以运行在大多数便携式设备上。最近，研究人员基于知识蒸馏相关方法[33]把 BERT 等大模型学到的知识迁移到小模型中[34-35]，成倍地减少了运行所需的空间和时间。这类小模型可以运行在便携式设备上，并在文本分类等多个任务上取得了可比的性能。

2 汉语 NER

与英语等语言相比，汉语的一个显著特点是词之间没有明确的边界。基于深度学习的汉语 NER 模型可大致分为以下三类：①基于词的模型，其首先对文本分词，然后再基于词进行命名实体识别，如图 1 所示[36]。这类模型可以利用词的相关信息，但主要缺点是分词不可避免地存在错误[37]，从而引起实体识别的错误。②基于字的模型，其不对文本分词，直接以汉字作为实体识别模型的输入[38-39]。这类模型可以避免分词错误带来的问题，通常能取得比基于词的模型更好的效果，但其没有利用词的相关信息，例如词的边界和语义等[40]。③融合字词信息的模型，其主要研究如何在基于字的模型中利用词的相关信息[41-42]。

基于词的汉语 NER 模型和基于字的汉语 NER 模型大都遵循第 1 节介绍的一般框架，这里不再赘述。目前，汉语 NER 的研究热点在于融合字词信息的模型，其性能优于前两类模型[43]。前期，研究者在多任务学习框架下同时训练汉语 NER 模型和分词模型，通过信息共享利用分词模型学到的词边界特征[44-46]。这类方法仅仅利用了词的边界信息，没有利用词的语义信息，而且还需要大量人工标注的汉语分词语料，因此其适用性并不强。下面详细介绍近期融合字词信息的汉语 NER 的相关工作，大致可分为基于字词图的方法和基于字词编码的方法。

2.1 基于字词图的方法

基于字词图的方法首先基于句子中的字和所有潜在词构建一个字词图作为模型的输入，然后设计专门的编码层来融合字和词的信息。与早期基于多任务学习的方法相比，基于字词图的方法需要的外部资源较少，通常仅需一个已有的或自动构建的词

典，而且可以同时利用词的边界信息和语义信息。

Zhang 和 Yang [47] 首次提出了基于 Lattice-LSTM 的汉语 NER 模型，如图 3 所示。该模型的输入层是一个由当前句子中的所有字以及所有潜在词构成的 Lattice，其中潜在词可以通过匹配已有的或自动构建的词典得到。Lattice 可以看作是一个字词图，其中相邻的字之间有边相连，潜在词的首字和尾字之间同样通过边相连接。在编码层，扩展常用的基于字的 LSTM 网络，在从左至右编码句子中字信息的同时，通过潜在词所在的边融合词的信息。该模型在获取字信息的基础上显式地融合了词的边界信息和语义信息，还避免了因分词错误而可能导致的错误传播问题，提高了汉语命名实体识别的性能。

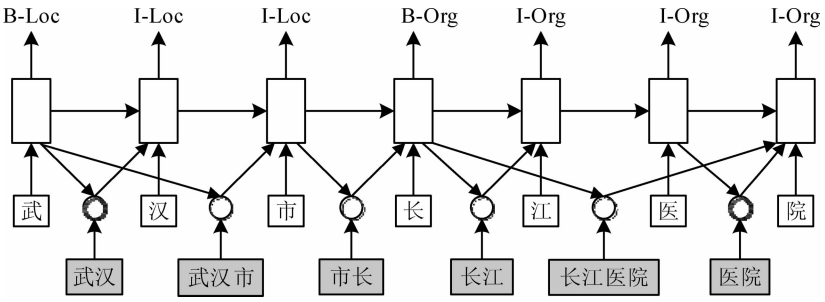

图 3　基于 Lattice-LSTM 的汉语 NER 模型

针对基于字的汉语命名实体识别中如何利用词信息的问题，上述基于 Lattice-LSTM 的模型的提出是解决该问题的重要一步。但是，其也存在以下几个方面的缺点：①Lattice 保留了所有潜在词的信息，这也带来了潜在词冲突的问题，从而可能引入噪声。例如，图 3 的 Lattice 中“市长”和“长江”之间存在的冲突往往需要全局语义才能区分。②沿着字序列从左至右处理信息，导致无法很好地融合与字自匹配的词（即词中包含该字）的信息，而这对命名实体识别是很重要的。例如，在标注“长”时，其自匹配的词“长江”的信息还未输入到模型中。③Lattice-LSTM 本质仍为 LSTM 网络，无法实现并行化，通过引入额外的边来处理词的信息，模型的复杂性大幅度增加，导致运行速度进一步降低。另外，由于与句子中每个字关联的潜在词的个数可能不同，使模型无法进行批处理训练，导致训练速度较慢。

为了缓解基于 Lattice-LSTM 的模型中潜在词冲突的问题，同时提升运行速度，Gui 等[48] 在 CNN 模型的基础上结合 Rethinking 机制[49] 对汉语命名实体识别进行研究，其编码层如图 4 所示。该模型把句子中的字和潜在词组织成了一个层次类型的结构，也可以认为是一个字词图。具体地，给定输入的句子和所有潜在词，模型层叠多个窗口为 2 的基于字的 CNN 层[50] 编码字特征和潜在词特征；使用注意力机制融合字和词的信息（简洁起见，图 4 中未画出）；引入 Rethinking 机制把 CNN 顶层所得的全局语义信息反馈到 CNN 底部的各层，调整潜在词对应的权重，从而缓解潜在词之间的冲突问题。实验表明，其在多个汉语数据集上的结果均优于基于 Lattice-LSTM 的模型，且运行速度更快。

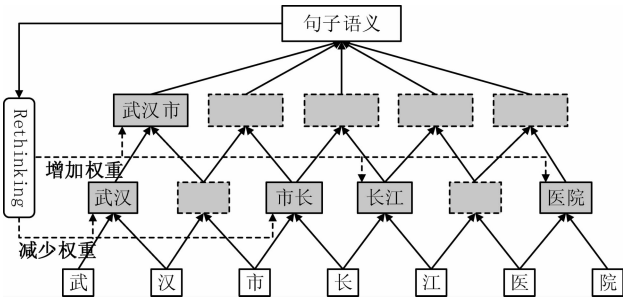

图 4　基于 CNN＋Rethinking 机制的编码层

Gui 等[51] 提出一种基于图神经网络的汉语 NER 模型，并把汉语 NER 问题看作是一个图节点的分类问题。如图 5 所示，他们把输入句子转换成一个基于字和词的有向图。句子中的每个字对应一个节点，潜在词看成是连接其首尾字的边节点，整个句子对应一个全局节点，全局节点与图中其他节点都有边相连（简洁起见，图 5 中省略了这些边）。基于“融合→更新→融合→……”的图信息处理方法[52]，模型能够很好地建模字、词和整个句子之间信息的交互。实验表明，Gui 等[51] 的模型在四个常用的汉语数据集 OntoNotes 4.0[53]、MSRA[54]、Weibo[55] 和 Resume[47] 上都取得了很好的结果。该模型在利用字词信息的同时，有效地缓解了基于 Lattice-LSTM 的模型存在的以下问题：① 通过引入全局节点建模句子的语义，有利于缓解潜在词冲突的问题；②基于图的模型易于并行化，能提高运

行的速度。

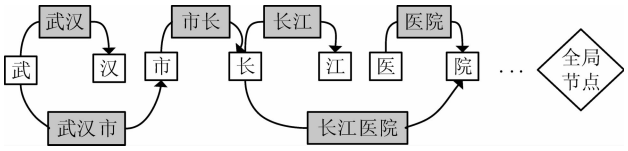

图 5 基于字和词的句子有向图

同一时期，Sui 等[56]提出一种基于协同图神经网络的汉语 NER 模型，其在编码层使用三个不同结构的图网络建模句子中字与潜在词之间不同类型的信息交互。其中，C-图（Contain-graph）用于融合字及其自匹配的词信息，T-图（Transition-graph）用于融合字与其最邻近的词信息，L-图（Lattice-graph）用于匹配 Lattice-LSTM 模型中所建模的词信息。模型通过叠加一个聚合层（Fusion Layer）整合三个图网络建模的信息，达到信息互补的目的。实验表明，Sui 等[56]的模型在多个数据集上获得了较好的性能，且运行速度成倍地快于基于 Lattice-LSTM 的模型。

2.2 基于字词编码的方法

上述基于字词图的方法虽然可以较好地融合字和词的信息以提高汉语 NER 的性能，但不足之处是其编码层通常依赖于字词图的结构，导致方法的可移植性不高。另外，引入的编码层通常相对复杂，不能满足需要实时响应的相关工业领域的应用需求。基于字词编码的方法通常只需改变 NER 模型的输入层，即把字和词的信息统一编码成联合表示作为模型的输入。这类方法比较简单，可以适用于多种类型的编码层，易于移植，而且能取得与基于字词图的方法可比的性能。

Ma 等[40]提出了一种基于 Soft-Lexicon 编码字词信息的汉语 NER 方法，其主要思想是在模型的输入层把字和词的信息编码成联合表示。如图 6 所示，对于句子中的字“长”，其对应的 Soft-Lexicon 表示为 B、M、E 和 S 四个集合，其中 B 表示以当前字开头的潜在词的集合，M 表示中间包含当前字的潜在词的集合，E 表示以当前字结尾的词的集合，S 集合中包含的潜在词是当前字本身，不存在相应的词则用 None 代替。为了进一步利用预训练的词向量信息，每个集合分别表示为其包含词的向量的融合。如果集合中只有一个词，则直接用这个词的词向量作为该集合的向量表示；如果有多个词，则使用一种基于词频计算的权重加权求和多个词向量。最后，拼接当前字的向量表示及其对应的 B、M、E 和 S 集合的向量表示作为字词信息的联合表示，用作模型的输入。上述基于 Soft-Lexicon 的输入层不仅利用了潜在词的边界信息，还利用了词的语义信息。由于该方法只调整了输入层，能适用于常用的如 BiLSTM、CNN 或 Transformer 等编码层，方法的可移植性强。实验表明，Ma 等[40]的方法能获得与当前最好模型可比的性能，同时成倍地加快实体识别的速度。

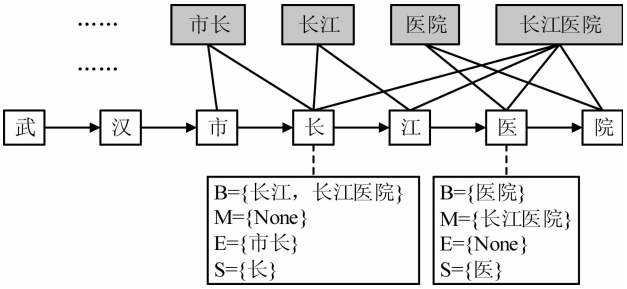

图 6 基于 Soft-Lexicon 的输入层

Liu 等[57]提出一种简单的字词编码输入层，并以 BiLSTM 作为模型编码层的方法。具体地，把句子中当前字的向量和以该字结尾的潜在词的向量拼接起来作为前向 LSTM 的输入，把句子中当前字的向量和以该字开头的潜在词向量拼接起来作为后向 LSTM 的输入。当一个位置有多个潜在词时，他们尝试了最短词优先、最长词优先、平均词向量和基于注意力机制融合词向量的方法，把多个潜在词表示为一个向量。虽然文章中以 BiLSTM 作为编码层，但提出的字词编码输入层经过简单调整后可适用于其他的编码层。Li 等[58]提出一种基于 Flat-Lattice 的输入层，其把所有潜在词直接拼接在输入句子后面形成一个扩展的字词序列，然后设计了一种位置编码的方法，用于编码字和词的相对位置。在模型的编码层，直接使用原始的 Transformer 来融合字和词的信息。与 Ma 等[40]提出的方法相比，Li 等[58]提出的字词编码方法只适用于基于 Transformer 的 NER 模型。

与英语等语言的命名实体相比，汉语命名实体有其自身的特点，导致其更难识别。主要包括：①汉语文本的词之间没有明确的边界；②汉语命名实体缺少明显的词形变换特征，比如英语等语言中的前后缀、大小写等；③汉语命名实体中存在大量缩写、中英文混用和实体相互重叠等现象[59]。近年来，汉语命名实体识别的研究主要针对汉语的词之间没有明确的边界这一特点展开。在重叠命名实体

识别方面，虽然已有大量针对英语重叠命名实体识别的研究[60-62]，但聚焦于汉语重叠命名实体识别的研究还较少，主要原因之一可能是没有被广泛认可的相关语料库[63]。

3　低资源的 NER

训练基于深度学习的命名实体识别模型通常需要大量人工标注的语料，然而，在大多数语言和领域中并没有或者只有少量标注的语料。低资源的 NER 是当前研究热点之一，其性能的提高是 NER 走向广泛实际应用的前提。该方面相关的研究工作可大致分为以下几类：跨语言迁移的方法、跨领域迁移的方法、跨任务迁移的方法和集成自动标注语料的方法。

3.1　跨语言迁移的方法

跨语言迁移方法的基本思路是利用资源丰富语言的标注数据帮助低资源语言进行命名实体识别。通常把资源丰富的语言称为源语言，把低资源的语言称为目标语言。当目标语言没有任何标注语料时(zero resource)，跨语言迁移的方法可大致分为基于数据迁移的方法和基于模型迁移的方法两大类。

3.1.1　基于数据迁移的方法

基于数据迁移的方法通常借助文本翻译和标签映射等手段把源语言中的标注数据转换成目标语言的标注数据，然后基于这些数据训练 NER 模型用于目标语言。例如，Ni 等[64]提出了一种在可比的语料库上进行标签映射的方法，用于创建自动标记的目标语言数据，并设计了一种启发式的方案筛选出高质量的标注数据。Mayhew 等[65]基于容易获得的双语词典，使用一种类似短语机器翻译[66]的方法自动翻译源语言的标注文本。Xie 等[67]首先基于双语词向量自动构建双语词典，然后把源语言的标注文本翻译成目标语言，在训练目标语言的 NER 模型时使用自注意力机制代替 BiLSTM 作为编码器，以缓解不同语言词序不同的问题。基于数据迁移方法的优点是可以利用目标语言的相关信息，缺点是自动生成的目标语言的标注数据不可避免地存在错误。

3.1.2　基于模型迁移的方法

基于模型迁移的方法通常先学习语言无关的特征，然后在源语言的标注语料上训练 NER 模型直接用于目标语言。例如，Wu 和 Dredze[68]直接使用基于 104 种语言训练的多语言版本 BERT[11]学习语言无关的词和句子表示，用于分本分类、词性标注和 NER 等多个跨语言任务，取得了很好的效果。Keung 等[69]在多语言版本 BERT 的基础上进一步使用对抗学习[70]的方法，以学习更好的与语言无关的特征。Chen 等[71]同样基于对抗学习的方法提取语言无关的特征，并动态地计算源语言和目标语言之间的相似度，从而更有效地实现从多个源语言到目标语言的知识迁移。Bari 等[72]首先训练一个源语言 NER 模型，然后基于无标注的目标语言语料进行调优(Fine-tuning)。为了在两种语言之间建立联系，他们基于对抗学习的方法把两种语言的词向量映射到同一语义空间；为了在两种语言的 NER 任务之间建立联系，他们提出了一种融合参数共享和特征增强的调优方法。Wu 等[73]提出了一种加强的元学习(meta-learning)方法，基于少量的目标语言测试数据对训练好的源语言模型进行调优，取得了很好的效果。基于模型迁移方法的优点是不需要生成目标语言的标注数据，缺点是没有充分利用目标语言的相关信息。

最近，Wu 等[74]指出基于数据迁移的方法和基于模型迁移的方法是可以互补的，但这两种方法都没有充分利用大量容易获得的目标语言中的未标注数据。虽然在基于对抗学习的方法中[69,71]通常会用到目标语言的文本，但其目的是学习语言无关的特征，往往丢失了特定于目标语言的相关信息。因此，他们首先分别基于数据迁移和模型迁移的方法训练两个目标语言的 NER 模型；然后，基于这两个模型标注大量目标语言的文本用于进一步调优，得到第三个目标语言的 NER 模型；最后，利用知识蒸馏[33]的方法集成这三个模型中的知识，取得了当前最好的性能。类似地，Wu 等[75]提出了一种基于知识蒸馏的方法，用于在只有训练好的源语言 NER 模型而源语言训练数据不可得到的情况下实现知识跨语言的迁移。

当目标语言有少量的标注语料时(Few resource)，一种可行的方法是先在源语言语料上基于模型迁移或数据迁移的方法训练一个目标语言 NER 模型，然后使用少量目标语言的标注语料对学到的模型进一步调优。另一种常见的方法是在不同语言的 NER 模型之间通过共享参数的策略迁移知识。例如，Yang 等[76]共享不同语言的 NER 模型中字符级的编码层；Lin 等[77]同时共享字符级和单词级的编码层；Zhou 等[78]在共享参数的基础上，提出

一种双重对抗训练的方法用于学习更好的语言无关特征,同时处理了源语言和目标语言数据极度不平衡的问题。

3.2 跨领域迁移的方法

跨领域迁移方法的基本思路是利用资源丰富领域的标注数据帮助低资源领域进行命名实体识别。跨领域和跨语言本质上是一致的,不同的语言也可以认为是不同的领域,因此跨语言迁移的方法与跨领域迁移的方法是基本类似的。

当目标领域没有标注语料时,通常使用模型迁移的方法。例如,Jia 等[79]提出了一种基于跨领域语言模型的方法,用于目标领域无监督的命名实体识别。该方法基于大量无标注的源领域文本和目标领域文本分别训练语言模型,基于源领域标注语言训练一个 NER 模型,并设计了一种参数生成网络,以实现跨领域的知识迁移和跨任务(语言模型和NER)的知识迁移。在多个目标领域上的实验显示,该方法取得了较好的效果。Liu 等[80]指出在有些资源极度缺乏的目标领域,无标注的文本也不容易获得,因此提出了一种仅需源领域标注语料的跨领域 NER 模型。具体地,他们首先引入一个辅助任务用来识别句子中的词是否为实体,以学习实体的一般表示,从而减轻不同领域之间的差异性;其次,引入了一个混合的实体专家框架(Mixture of Entity Experts)来避免模型过拟合源领域训练数据。以英文 CoNLL-2003 为源领域语料,该方法在多个零资源的目标领域取得了与 Jia 等[79]的模型可比的性能。

当目标领域有少量的标注语料时,通常基于共享参数或特征映射等策略在领域之间迁移知识。例如,He 和 Sun[81]联合使用源领域和目标领域的标注语料训练多个共享参数的 NER 模型,并在训练时基于源领域句子和目标领域句子的相似度调整该句子的学习率。Yang 等[76]基于层次循环神经网络和参数共享策略,提出了多个分别用于跨领域、跨语言和跨任务的序列标注模型。Lee 等[82]首先使用源领域的标注语料训练 NER 模型,然后使用少量目标领域的标注语料对模型进行调优。Wang 等[83]在共享源领域和目标领域 NER 模型的词向量层和编码层的基础上,设计了两种标签感知约束代价用于特征迁移和参数迁移。Lin 和 Lu[84]提出了一种轻量级的跨领域自适应性方法。具体地,他们先在一个已经训练好的源领域 NER 模型中添加自适应的神经网络层,然后基于少量的目标领域语料进行调优。这种方法的好处是仅需要训练好的源领域 NER 模型,而不再需要源领域的标注语料。Wang 等[85]提出了一种多任务学习框架,同时利用多个源领域的标注数据帮助目标领域,取得了较好的效果。

3.3 跨任务迁移的方法

跨任务迁移方法的基本思路是利用相关任务中的信息帮助命名实体识别。例如,词性信息和汉语中词的边界信息显然对命名实体是有用的。早在 2008 年,Collobert 和 Weston[86]在多任务学习框架下基于 CNN 网络联合训练词性标注、语义角色标注和命名实体识别等任务,通过共享参数的方式迁移知识。Lin 等[77]在多任务学习框架下联合训练多种语言下的多个相关任务。Sanh 等[87]则在层次多任务学习框架下联合训练命名实体识别、实体提及和关系抽取等任务。他们认为这些任务需要的语义具有层次性,应该关联到不同的神经网络层。Aguilar 等[88]针对社交媒体数据中存在不正确的语法结构、拼写错误和非正式缩写等问题,提出使用更具一般性的命名实体分割任务(预测一个词是否为实体)作为辅助任务帮助命名实体识别任务,在 WNUT-17 数据集上获得了很好的效果。Kruengkrai 等[89]则提出联合训练句子级的分类任务和命名实体识别任务,好处是可以利用大量较容易获得的句子级的标注语料。

上述方法大都需要利用相关任务中有标注的训练数据,与之不同,Rei[90]和 Liu 等[91]联合训练语言模型和命名实体识别模型,好处是可以在大量无标注的语料上基于语言模型学习语法和语义知识,取得了较好的效果。最近,以 BERT 等基于超大规模语料预训练的语言模型为基础的 NER 模型在多个数据集上取得了最佳的性能[3,11],这也可以看作是一种跨任务迁移的方法。该类模型通常以预训练好的 BERT 作为编码层,然后在命名实体标注任务上进一步调优。总之,跨任务迁移的方法大都通过共享参数的方式迁移知识,以提高命名实体识别的性能。

3.4 集成自动标注语料的方法

基于语言/领域迁移的方法虽然能够有效地缓解标注语料短缺的问题,但是具有丰富标注资源的语言或领域是非常少的,且存在标注类别不同的问题,因此在很多实际应用中难以找到可以迁移的资

源。例如，在电子商务领域中通常需要识别商品的名称，而资源丰富的领域往往标注的是人名、地名和机构名等通用的实体。

为此，一些研究者提出集成自动标注语料的方法，首先通过某种方法自动标注大量语料，然后集成它们用于提高低资源 NER 的性能。自动标注语料一般使用基于 Wikipedia 等 Web 资源自动抽取的方法[92]或基于知识库/领域实体词典匹配的方法[93]。虽然可以较容易地生成大量自动标注的语料，但这些语料中往往存在较多的噪声。如图 7 所示，基于商品名称词典匹配的方法自动生成的标注语料可能出现以下几种情况：①正确标注，句子中所有的商品都正确标注了；②部分标注，句子中的商品“普通冰箱”正确标注了，而“智能冰箱”漏标了；③噪声标注，句子中的商品“机器人”标注错误，正确的应该是“扫地机器人”。因此，集成自动标注语料的方法重点研究如何有效地利用正确标注和部分标注的数据，同时减少噪声标注数据的影响。

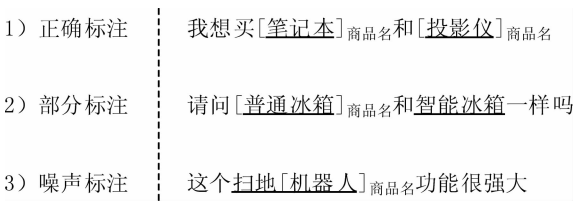

图 7　自动标注语料的示例

Yang 等[93]首先基于词典匹配的方法自动标注语料，然后使用 Partial-CRF[94]在少量人工标注的语料和大量自动标注的语料上训练 NER 模型。此外，他们还基于强化学习[95]训练一个选择器，用于筛选掉具有噪声的标注数据。Shang 等[96]提出一种“连接-断开”的标注方法代替常用的基于 CRF 或者 Partial-CRF 的方法。他们训练一个二分类器用于预测相邻的两个字/词是否在同一个实体内，是则为“连接”，不是则为“断开”。采用这种标注方法的主要原因是自动标注语料中的某些实体边界可能有误，但其中大部分字/词之间的“连接”关系是对的。例如，图 7 噪声标注的商品名“机器人”中，“机”与“器”、“器”与“人”之间的“连接”关系是对的。实验结果表示，他们提出的方法在多个数据集上的结果优于常用的 Partial-CRF 方法。Mayhew 等[97]提出一种自动调整权重的方法，其训练一个迭代的二分类器为自动标注语料中的每个实体计算权重，主要目的是降低漏标实体的权重。例如，图 7 部分标注句子中的“智能冰箱”被漏标记为其他字符，降低其权重有利于处理这类噪声问题。Peng 等[98]则把命名实体识别任务建模成一个 PU（Positive-unlabeled）学习问题，其把基于词典匹配方法自动标注的实体作为正例，把剩余的部分作为未标注文本训练模型。这种方法的优点是可以较好地解决漏标实体的问题，从而降低对词典规模和质量的要求。Liang 等[99]提出了一种两阶段的 NER 模型，其首先使用大量自动标注的语料训练以 BERT 为编码层的 NER 模型，然后使用自训练（Self-training）的方法进一步调优模型。

Cao 等[92]提出了一种不需要任何人工标注语料训练 NER 模型的方法。他们首先基于 Wikipedia 自动构建大量标注语料，然后通过计算标注的置信度和覆盖度两个指标把语料分成高质量和低质量两部分。例如，图 7 中正确标注句子是高质量的标注语料，而部分标注句子和噪声标注句子则是低质量的。为了充分利用低质量的语料，他们设计了一个基于字的分类任务，即针对其中标注的实体中的每个字，预测它们的实体类别。例如，对图 7 噪声标注句子仅分别预测“机”“器”和“人”的类别。训练这个分类模型利用了大量正确的标注信息，同时降低了噪声的影响。因此，上述分类模型中的编码器建模了大量的上下文语义信息，可用于初始化 NER 模型中的编码器。最后，他们在高质量的自动标注语料上继续对该 NER 模型进行调优。Lison 等[100]则融合基于多种方法自动标注的语料用于训练 NER 模型。他们首先基于训练好的领域外 NER 模型、实体词典和启发式规则等方式标注语料；然后，训练一个隐马尔可夫模型把经过多种方式自动标注的语料（一份领域内的语料，多份标签）融合在一起（一份标签）；最后，基于融合后的语料训练 NER 模型。同期，Safranchik 等[101]则引入一类连接规则（Linking rules）用于推断句子中可能是实体的文本片段，例如，同一文档中多次出现的 n-grams。然后，他们提出了一种基于扩展的 HMM 的 NER 模型，并联合使用基于词典匹配等方法自动标注的语料和基于连接规则生成的语料训练模型。

实际应用中，绝大部分的语言和领域都是低资源（甚至零资源）的，如何提高这些语言和领域命名实体识别的性能是当前的研究热点之一。基于知识迁移的思路，大量跨语言迁移或跨领域迁移的相关研究工作不同程度地提高了低资源 NER 的性能。然而，由于标注语料丰富的语言和领域非常少，这些方法的适用范围受到一定的限制。相比较而言，自

动标注语料的方法可以快速地、低成本地获取大量含噪声的标注语料。现有研究工作已证实,集成这些自动标注的语料可以实质性地提高低资源NER的性能。鉴于BERT等预训练语言模型的成功(超大规模无标注文本的利用),有理由相信如何更好地利用大量自动标注的命名实体语料将是未来重要的研究方向之一。

4 性能对比

4.1 不同模型 F_1 值对比

为了让读者对基于深度学习的NER模型的性能有一个直观的了解,本节列举了一些具有代表性的模型和方法在常用数据集上的 F_1 值。F_1 值常用于评估命名实体识别模型的性能,具体计算如式(1)所示。

$$\begin{aligned} F_1 &= 2*P*R/(P+R) \\ P &= T_1/(T_1+T_2) \\ R &= T_1/(T_1+T_3) \end{aligned} \tag{1}$$

其中,T_1 表示正确识别出的实体数,T_2 表示错误识别出的实体数,T_3 表示未被识别出的实体数,P(Precision)表示查准率,R(Recall)表示查全率。需要说明的是,本节所列结果均来源于所引用的文献。

表2列出最近的基于深度学习的NER模型在CoNLL2003和OntoNotes 5.0这两个常用的英语数据集上的 F_1 值。其中,CoNLL2003数据集中标注了人名、地名、组织机构名及其他类别4种实体,来自于路透社新闻;而OntoNotes 5.0所标注的实体类别有18种之多,且由多个不同领域的文本组成,实体识别的难度更大。

表2 基于深度学习的NER模型在英文数据集CoNLL2003和OntoNotes 5.0上的性能对比

工作	输入层	编码层	解码层	CoNLL2003/%	OntoNotes5.0/%
Lample 等[17]	词	BiLSTM	CRF	90.20	—
Lample 等[17]	词+字符	BiLSTM	CRF	90.94	—
Chiu 和 Nichols[19]	词+字符	BiLSTM	Softmax	90.91	86.17
Shen 等[30]	词+字符	CNN	RNN	90.69	86.52
Strubell 等[24]	词	CNN	CRF	90.65	86.84
Chen 等[25]	词+字符	CNN	CRF	91.44	87.67
Li 等[26]	词+字符	句法树引导的BRNN	Softmax	—	87.21
Jie 和 Lu[27]	词+字符	依存树引导的BiLSTM	CRF	—	88.52
Yan 等[29]	词+字符	Transformer	CRF	91.43	88.43
Devlin 等[11]	BERT	—	Softmax	92.80	—
Li 等[3]	BERT	—	Softmax	**93.33**	**92.07**

从表2可以看出:①在输入层中,字符信息是词信息的有益补充,“词+字符”的输入模式是事实上的标准;②采用不同编码层(BiLSTM、CNN、树结构引导的神经网络或Transformer)的模型之间性能上并没有明显的差别;③解码层使用简单的Softmax直接分类能取得与CRF可比的性能,这说明编码层可能已经捕获到了标签之间的依赖关系;④基于BERT的NER模型的性能显著地高于以前的模型,在两个数据集上均取得了当前最好的性能,这主要是因为基于大规模文本预训练的BERT中包含有大量的语义信息。基于深度学习的NER模型虽然在CoNLL2003和OntoNotes 5.0等比较正规的文本上取得了超过90%的 F_1 值,但在网络文本数据集W-NUT17上 F_1 值还不到50%[4],远未达到实用的要求。这充分说明命名实体识别依然是一个极具挑战性的任务。

4.2 不同数据集上 F_1 值对比

表3列出了最近的汉语NER模型在OntoNotes 4.0(ON)、MSRA、Weibo(WB)以及Resume(RS)四个常用的汉语数据集上的 F_1 值。数据集OntoNotes 4.0来自新闻、广播等多个领域,标注了人名、产品名、日期等18种实体;数据集MSRA来自新闻领域,只标注了人名、组织机构名和地名3种实体。数据集Weibo和Resume分别来自社交媒体新浪微博和新浪财经,Weibo中标注了

人名、地名、组织机构名和地缘政治 4 种实体，Resume标注了教育机构、职业、职称等 8 种实体。

表 3　汉语 NER 模型的性能对比

工作	输入层	ON /%	MSRA /%	WB /%	RS /%
Zhang 和 Yang[47]	词	65.63	86.65	47.33	93.58
Zhang 和 Yang[47]	字	64.30	88.81	52.77	93.48
Zhang 和 Yang[47]	字词图	73.88	93.18	58.79	94.46
Gui 等[48]	字词图	74.45	93.71	59.92	95.11
Gui 等[51]	字词图	74.89	93.46	60.21	95.37
Sui 等[56]	字词图	74.79	93.47	63.09	—
Ma 等[40]	字词编码	75.54	93.50	61.24	**95.59**
Li 等[58]	字词编码	**75.70**	**94.35**	**63.42**	94.93

从表 3 中可以看出：①在基于字的汉语 NER 模型中融入词的信息实质性地提高了识别的性能，原因是其在利用词信息的同时避免了汉语分词可能带来的错误。这也表明针对汉语的特点设计相应的 NER 模型是非常重要的；②简单的基于字词编码的方法取得了与基于字词图的方法可比的、甚至更好的性能；③在识别难度较小的 MSRA 和 RS 数据集上取得了超过 90% 的 F_1 值，而在识别难度较大的 ON 和 WB 数据集上的 F_1 值则低得多。对比表 3 中数据集 OntoNotes 4.0(汉语)与表 2 中数据集 OntoNotes 5.0(英语)的 F_1 值，可以发现模型在这两个相似数据集上的性能差异非常明显，这从某种程度上说明汉语命名实体识别的难度要高于英文命名实体识别，如何针对汉语的特点设计 NER 模型是一个值得深入研究的问题。

4.3　跨语言迁移的 NER 方法的 F_1 值对比

在低资源 NER 的相关研究中，研究者使用的数据集不太统一，难以进行直接对比。表 4 中仅列出了最近跨语言迁移的 NER 方法的 F_1 值，这些方法假定目标语言是零资源的，并采用单一源语言到单一目标语言的迁移方式(一对一迁移)。常使用 CoNLL2003(English-en)作为资源丰富的源语言，使用 CoNLL2002(Spanish-es 和 Dutch-nl)和 CoNLL2003(German-de)中的一部分语料作为目标语言的测试集。以上数据集标注的实体包括人名、地名、机构名和其他类别 4 种。从表 4 中可以看出：①近年来，跨语言迁移的 NER 方法的性能取得了显著的提高；②基于模型迁移的方法的性能明显好于基于数据迁移的方法；③Wu 等[74]的方法综合数据迁移方法和模型迁移方法的优点，取得了当前最好的性能。这些结果充分说明了跨语言迁移方法的有效性，是一个值得深入研究的方向。

表 4　跨语言迁移的 NER 方法的性能对比

工作	方法类别	es/%	nl/%	de/%
Ni 等[64]	零资源＋数据迁移	65.10	65.40	58.50
Mayhew 等[65]	零资源＋数据迁移	65.95	66.50	59.11
Xie 等[67]	零资源＋数据迁移	72.37	71.25	57.76
Wu 和 Dredze[68]	零资源＋模型迁移	74.50	79.50	71.10
Wu 等[75]	零资源＋模型迁移	76.94	80.89	73.22
Wu 等[74]	零资源＋数据迁移＋模型迁移	**79.31**	**82.90**	**74.82**

5　结语

基于深度学习的命名实体识别方法在性能上已经超过了早期基于人工特征的方法，是目前的研究热点之一。本文从命名实体识别的一般框架、汉语命名实体识别和低资源的命名实体识别三个方面着手介绍近年来的相关研究工作，并分析了它们的优缺点。基于这些分析，我们建议未来命名实体识别的研究工作可以从以下几个方面展开：

(1) 非正式文本的命名实体识别研究。目前基于深度学习的命名实体识别模型的一般框架趋于成熟，在新闻等正式文本上也取得了比较满意的性能。但是，这些模型在非正式文本上的性能还比较低，离实际应用仍有较大的差距。在自媒体快速发展的今天，如何提高非正式文本的命名实体识别的性能是一个迫切需要解决的问题。现有框架是否适用于非正式文本也是一个值得考虑的问题。

(2) 领域特定命名实体识别研究。目前针对人名、地名和机构名等通用命名实体识别的研究较多，而对于领域特定命名实体识别的研究则相对较少，例如，电子商务文本中商品名等实体的识别。领域特定命名实体的准确识别是信息抽取的前提和基础，是这些领域走向智能信息处理的关键之一。

(3) 篇章级的命名实体识别研究。目前的模型主要利用实体本身及其所在句子中的上下文信息进行识别，而没有考虑篇章级信息的应用。例如，Gui 等[102]指出一个篇章中多次出现的文本片段的类型(某种实体或非实体)大部分情况下是一致的，并设计出一个两阶段的模型以利用这些信息，取得了较

好的性能。篇章级的命名实体识别研究才刚刚起步,一方面篇章级的哪些信息对命名实体识别有用尚不清楚,另一方面如何改进现有模型以高效地利用篇章信息也是一个问题。

(4) 经济高效的命名实体识别研究。现有的模型虽然取得了较好的性能,但往往在训练和应用时都需要耗费大量的计算资源。例如,性能领先的基于BERT的命名实体识别模型[11]对计算资源的要求非常高。如何在识别性能和计算效率之间达到平衡是一个很实际的问题,也是命名实体识别技术广泛运用的关键之一。

(5) 汉语命名实体识别的进一步研究。与英语等语言相比,汉语具有词之间没有明确的边界、缺少词形变换等特点,使得其命名实体识别的难度更大。当前汉语命名实体识别的研究主要针对词之间没有明确的边界这一特点展开,对其他特点的探索和利用非常不充分。深入研究汉语与英语等语言中实体的差异,并设计相应的模型是提高汉语命名实体识别性能的有效手段之一。

(6) 集成自动标注语料的命名实体识别研究。虽然跨语言迁移和跨领域迁移的方法都能在一定程度上缓解标注资源短缺的问题,但标注资源丰富的语言和领域毕竟非常少,限制了这些方法的适用范围。集成自动标注语料的方法仅需要实体词典等相对容易获得的资源,可以快速地应用于一种新的语言或一个新的领域,适用范围更广。但是,如何有效地克服自动标注语料中噪声的影响,依然是一个极具挑战性的问题。

参考文献

[1] Sundheim B. Named entity task definition (v2.1)[C]//Proceedings of Message Understanding Conference, 1995: 319-332.

[2] Sang E F T K, Meulder F D. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition[C]//Proceedings of NAACL, 2003: 142-147.

[3] Li X, Sun X, Meng Y, et al. Dice loss for data-imbalanced NLP tasks[C]//Proceedings of ACL, 2020: 465-476.

[4] Lin B Y, Xu 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.

[5] Wang J, Shou L, Chen K, et al. Pyramid: A layered model for nested named entity recognition[C]//Proceedings of ACL, 2020: 5918-5928.

[6] Sohrab M G, Miwa M. Deep exhaustive model for nested named entity recognition[C]//Proceedings of EMNLP, 2018: 2843-2849.

[7] Li Y, Chiticariu L, Reiss F, et al. Domain adaptation of rule-based annotators for named-entity recognition tasks[C]//Proceedings of EMNLP, 2010: 1002-1012.

[8] Gayen V, Sarkar K. An HMM based named entity recognition system for Indian languages: the JU System at ICON 2013[J]. CoRR, 2014, abs/1405.7397.

[9] 朱颢东, 杨立志, 丁温雪, 等. 基于主题标签和CRF的中文微博命名实体识别[J]. 华中师范大学学报(自然科学版), 2018, 052(3): 316-321.

[10] Su J, Tan Z, Xiong D, et al. Lattice-based recurrent neural network encoders for neural machine translation[C]//Proceedings of AAAI, 2017: 3302-3308.

[11] Devlin J, Chang M-W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL-HLT, 2019: 4171-4186.

[12] Cui L, Zhang Y. Hierarchically-refined label attention network for sequence labeling[C]//Proceedings of EMNLP-IJCNLP, 2019: 4115-4128.

[13] Wu C, Hu C, Li R, et al. Hierarchical multi-task learning with CRF for implicit discourse relation recognition[J]. Knowledge-Based Systems, 2020, 195: 105637.

[14] Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models[C]//Proceedings of COLING, 2018: 2145-2158.

[15] Li J, Sun A, Han J, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020.

[16] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[17] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[C]//Proceedings of NAACL-HLT, 2016: 260-270.

[18] Luo G, Huang X, Lin C Y, et al. Joint entity recognition and disambiguation[C]//Proceedings of EMNLP, 2015: 879-888.

[19] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNS[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[20] Yadav V, Sharp R, Bethard S. Deep affix features improve neural named entity recognizers[C]//Proceedings of Lexical and Computational Semantics, 2018: 167-172.

[21] Lin Y, Liu L, Ji H, et al. Reliability-aware dynamic feature composition for name tagging[C]//Proceedings of ACL, 2019: 165-174.

[22] Pradhan S, Moschitti A, Xue N, et al. Towards robust linguistic analysis using ontonotes[C]//Proceedings of CoNLL, 2013: 143-152.

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of NIPS, 2017: 6000-6010.

[24] Strubell E, Verga P, Belanger D, et al. Fast and accurate entity recognition with iterated dilated convolutions[C]//Proceedings of EMNLP, 2017: 2670-2680.

[25] Chen H, Lin Z, Ding G, et al. GRN: gated relation network to enhance convolutional neural network for named entity recognition[C]//Proceedings of AAAI, 2019:6236-6243.

[26] Li P-H, Dong R-P, Wang Y-S, et al. Leveraging linguistic structures for named entity recognition with bidirectional recursive neural networks[C]//Proceedings of EMNLP, 2017: 2664-2669.

[27] Jie Z, Lu W. Dependency-guided LSTM-CRF for named entity recognition[C]//Proceedings of EMNLP-IJCNLP, 2019: 3862-3872.

[28] Guo Q, Qiu X, Liu P, et al. Star-transformer[C]//Proceedings of NAACL-HLT, 2019: 1315-1325.

[29] Yan H, Deng B, Li X, et al. TENER: Adapting transformer encoder for named entity recognition[J]. CoRR, 2019, abs/1911.04474.

[30] Shen Y, Yun H, Lipton Z C, et al. Deep active learning for named entity recognition[C]//Proceedings of the 2nd Workshop on Representation Learning for NLP, 2018: 252-256.

[31] Zhai F, Potdar S, Xiang B, et al. Neural models for sequence chunking[C]//Proceedings of AAAI, 2017: 3365-3371.

[32] Vinyals O, Fortunato M, Jaitly N. Pointer networks[C]//Proceedings of Advances in Neural Information Processing Systems 28, 2015: 2692-2700.

[33] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[C]//Proceedings of NIPS Workshop. 2015.

[34] Liu W, Zhou P, Wang Z, et al. FastBERT: A self-distilling BERT with adaptive inference time[C]//Proceedings of ACL, 2020: 6035-6044.

[35] Sun Z, Yu H, Song X, et al. MobileBERT: A compact task-agnostic BERT for resource-limited devices[C]//Proceedings of ACL, 2020: 2158-2170.

[36] 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.

[37] Tian Y, Song Y, Xia F, et al. Improving Chinese word segmentation with wordhood memory networks[C]//Proceedings of ACL, 2020: 8274-8285.

[38] Lu Y, Zhang Y, Ji D. Multi-prototype Chinese character embedding[C]//Proceedings of LREC, 2016: 855-859.

[39] Dong C, Zhang J, Zong C, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition[G]//Natural Language Understanding and Intelligent Applications, 2016, 10102: 239-250.

[40] Ma R, Peng M, Zhang Q, et al. Simplify the usage of lexicon in Chinese NER[C]//Proceedings of ACL, 2020: 5951-5960.

[41] 张海楠，伍大勇，刘悦，等. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报，2017(4)：33-40.

[42] 殷章志，李玖一. 融合字词模型的中文命名实体识别研究[J]. 中文信息学报，2019，33(11)：95-100.

[43] 石春丹，秦岭. 基于 BGRU-CRF 的中文命名实体识别方法[J]. 计算机科学，2019，046(009)：237-242.

[44] Peng N, Dredze M. Improving named entity recognition for Chinese social media with word segmentation representation learning[C]//Proceedings of ACL, 2016: 149-155.

[45] Cao P, Chen Y, Liu K, et al. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism[C]//Proceedings of EMNLP, 2018: 182-192.

[46] Wu F, Liu J, Wu C, et al. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation[C]//Proceedings of the World Wide Web Conference on -WWW, 2019: 3342-3348.

[47] Zhang Y, Yang J. Chinese NER using lattice LSTM[C]//Proceedings of ACL, 2018: 1554-1564.

[48] Gui T, Ma R, Zhang Q, et al. CNN-based Chinese NER with lexicon rethinking[C]//Proceedings of IJCAI, 2019: 4982-4988.

[49] Li X, Jie Z, Feng J, et al. Learning with rethinking: recurrently improving convolutional neural networks through feedback[J]. Pattern Recognition, 2018, 79: 183-194.

[50] Kim Y. Convolutional neural networks for sentence classification[C]//Proceedings of EMNLP, 2014: 1746-1751.

[51] Gui T, Zou Y, Zhang Q, et al. A lexicon-based graph neural network for Chinese NER[C]//Proceedings of EMNLP-IJCNLP, 2019: 1040-1050.

[52] Gilmer J, Schoenholz S S, Riley P F, et al. Neural

message passing for quantum chemistry[C]//Proceedings of the 34th International Conference on Machine Learning, 2017: 2053-2070.

[53] Weischedel R, Pradhan S, Ramshaw L, et al. OntoNotes release 4.0. LDC2011T03, Philadelphia, Penn: Linguistic Data Consortium, 2011.

[54] Levow G-A. The third international Chinese language processing bakeoff: word segmentation and named entity recognition[C]//Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing, 2006: 108-117.

[55] Peng N, Dredze M. Named entity recognition for Chinese social media with jointly trained embeddings [C]//Proceedings of EMNLP, 2015: 548-554.

[56] Sui D, Chen Y, Liu K, et al. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network[C]//Proceedings of EMNLP-IJCNLP, 2019: 3828-3838.

[57] Liu W, Xu T, Xu Q, et al. An encoding strategy based word-character LSTM for Chinese NER[C]// Proceedings of NAACL-HLT, 2019: 2379-2389.

[58] Li X, Yan H, Qiu X, et al. FLAT: Chinese NER using flat-lattice transformer [C]//Proceedings of ACL, 2020: 6836-6842.

[59] Duan H, Zheng Y. A study on features of the CRFs-based Chinese named entity recognition[J]. International Journal of Advanced Intelligence Paradigms, 2011.

[60] Ju M, Miwa M, Ananiadou S. A neural layered model for nested named entity recognition[C]//Proceedings of NAACL-HLT, 2018: 1446-1459.

[61] Fisher J, Vlachos A. Merge and label: A novel neural network architecture for nested NER[C]//Proceedings of ACL, 2019: 5840-5850.

[62] Luo Y, Zhao H. Bipartite flat-graph network for nested named entity recognition [C]//Proceedings of ACL, 2020: 6408-6418.

[63] 李雁群，何云琪，钱龙华，等. 中文嵌套命名实体识别语料库的构建[J]. 中文信息学报，2018，32(8)：19-26.

[64] Ni J, Dinu G, Florian R. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection[C]//Proceedings of ACL, 2017: 1470-1480.

[65] Mayhew S, Tsai C-T, Roth D. Cheap translation for cross-lingual named entity recognition[C]//Proceedings of EMNLP, 2017: 2536-2545.

[66] Koehn P, Zens R, Dyer C, et al. Moses: Open source toolkit for statistical machine translation[C]// Proceedings of ACL, 2007: 177-180.

[67] Xie J, Yang Z, Neubig G, et al. Neural cross-lingual named entity recognition with minimal resources [C]//Proceedings of EMNLP, 2018: 369-379.

[68] Wu S, Dredze M. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT[C]//Proceedings of EMNLP-IJCNLP, 2019: 833-844.

[69] Keung P, Lu Y, Bhardwaj V. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER[C]//Proceedings of EMNLP-IJCNLP, 2019: 1355-1360.

[70] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[C]//Proceedings of NIPS, 2014: 2672-2680.

[71] Chen X, Awadallah A H, Hassan H, et al. Multi-source cross-lingual model transfer: Learning what to share[C]//Proceedings of ACL, 2019: 3098-3112.

[72] Bari M S, Joty S, Jwalapuram P. Zero-resource cross-lingual named entity recognition[C]//Proceedings of AAAI, 2020, 34: 7415-7423.

[73] Wu Q, Lin Z, Wang G, et al. Enhanced meta-Learning for cross-lingual named entity recognition with minimal resources[C]//Proceedings of AAAI. 2020, 34: 9274-9281.

[74] Wu Q, Lin Z, Karlsson B F, et al. Unitrans: Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data[C]// Proceedings of IJCAI, 2020: 3926-3932.

[75] Wu Q, Lin Z, Karlsson B F, et al. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language[C]//Proceedings of ACL, 2020: 6505-6514.

[76] Yang Z, Salakhutdinov R, Cohen W W. Transfer learning for sequence tagging with hierarchical recurrent Networks[C]//Proceedings of ICLR, 2017.

[77] Lin Y, Yang S, Stoyanov V, et al. A multi-lingual multi-task architecture for low-resource sequence labeling[C]//Proceedings of ACL, 2018: 799-809.

[78] Zhou J T, Zhang H, Jin D, et al. Dual adversarial neural transfer for low-resource named entity recognition[C]//Proceedings of ACL, 2019: 3461-3471.

[79] Jia C, Liang X, Zhang Y. Cross-domain NER using cross-domain language modeling[C]//Proceedings of ACL, 2019: 2464-2474.

[80] Liu Z, Winata G I, Fung P. Zero-resource cross-domain named entity recognition[C]//Proceedings of the 5th Workshop on Representation Learning for NLP, 2020: 1-6.

[81] He H, Sun X. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media[C]//Proceedings of AAAI, 2017: 3216-3222.

[82] Lee J Y, Dernoncourt F, Szolovits P. Transfer learn-

ing for named-entity recognition with neural networks [C]//Proceedings of LREC, 2018: 4470-4473.

[83] Wang Z, Qu Y, Chen L, et al. Label-aware double transfer learning for cross-specialty medical named entity recognition [C]//Proceedings of NAACL, 2018: 1-15.

[84] Lin B Y, Lu W. Neural adaptation layers for cross-domain named entity recognition[C]//Proceedings of EMNLP, 2018:2012-2022.

[85] Wang J, Kulkarni M, Preotiuc-Pietro D. Multi-domain named entity recognition with genre-aware and agnostic inference[C]//Proceedings of ACL, 2020: 8476-8488.

[86] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multi-task learning [C]//Proceedings of Machine Learning, 2008: 160.

[87] Sanh V, Wolf T, Ruder S. A hierarchical multi-task approach for learning embeddings from semantic tasks[C]//Proceedings of AAAI, 2019, 33: 6949-6956.

[88] Aguilar G, Maharjan S, LÓPEZ Monroy A P, et al. A multi-task approach for named entity recognition in social media data[C]//Proceedings of the 3rd Workshop on Noisy User-generated Text, 2017: 148-153.

[89] 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 ACL, 2020: 5898-5905.

[90] Rei M. Semi-supervised multitask learning for sequence labeling [C]//Proceedings of ACL, 2017: 2121-2130.

[91] Liu L, Shang J, Xu F F, et al. Empower sequence labeling with task-aware neural language model[C]//Proceedings of AAAI, 2018: 5253-5260.

[92] Cao Y, Hu Z, Chua T, et al. Low-resource name tagging learned with weakly labeled data[C]//Proceedings of EMNLP-IJCNLP, 2019: 261-270.

[93] Yang Y, Chen W, Li Z, et al. Distantly supervised ner with partial annotation learning and reinforcement learning[C]//Proceedings of Computational Linguistics, 2018: 2159-2169.

[94] Tsuboi Y, Kashima H, Oda H, et al. Training conditional random fields using incomplete annotations [C]//Proceedings of COLING, 2008, 1: 897-904.

[95] Feng J, Huang M, Zhao L, et al. Reinforcement learning for relation classification from noisy data [C]//Proceedings of AAAI, 2018: 5779-5786.

[96] Shang J, Liu L, Gu X, et al. Learning named entity tagger using domain-specific dictionary[C]//Proceedings of EMNLP, 2018: 2054-2064.

[97] Mayhew S, Chaturvedi S, Tsai C-T, et al. Named entity recognition with partially annotated training data[C]//Proceedings of CoNLL, 2019: 645-655.

[98] Peng M, Xing X, Zhang Q, et al. Distantly supervised named entity recognition using positive-unlabeled learning [C]//Proceedings of ACL, 2019: 2409-2419.

[99] Liang C, Yu Y, Jiang H, et al. BOND: BERT-assisted open-domain named entity recognition with distant supervision[C]//Proceedings of SIGKDD. 2020: 1054-1064.

[100] Lison P, Barnes J, Hubin A, et al. Named entity recognition without labelled data: A weak supervision approach [C]//Proceedings of ACL, 2020: 1518-1533.

[101] Safranchik E, Luo S, Bach S. Weakly supervised sequence tagging from noisy rules[C]//Proceedings of AAAI, 2020, 34: 5570-5578.

[102] Gui T, Ye J, Zhang Q, et al. Leveraging document-level label consistency for named entity recognition [C]//Proceedings of IJCAI, 2020: 3976-3982.

邓依依(1996—)，硕士研究生，主要研究领域为自然语言处理、深度学习
E-mail: dyiii_20@163.com

魏永丰(1975—)，硕士，讲师，主要研究领域为数据库性能优化。
E-mail: 1052@ecjtu.edu.cn

邬昌兴(1981—)，通信作者，博士，讲师，主要研究领域为基于机器学习的自然语言处理。
E-mail: wuchangxing@ecjtu.edu.cn