命名实体识别研究综述

摘要

命名实体识别作为自让语言处理中的基本问题之一，其性能的好坏将直接影响下游的众多任务。近年来，随着“连接主义”的复苏，命名实体识别任务也从传统的基于统计的方法逐渐过渡至深度神经网络模型。本文详细介绍了命名实体识别任务的定义与起源，主流方法的发展脉络，以及现有的问题困难及未来的发展方向。

关键词 命名实体识别；信息抽取；自然语言处理；深度学习

1. 引 言

命名实体识别（Named entity recognition）是指识别文本中指定类别的实体，如人名、地名、组织机构名、时间等专有名词[1]。例如，识别“2018年4月3日，中华人民共和国商务部宣布对原产于美国的大豆农产品、汽车、水产品等进口商品增加25%的关税”这句中的地名（美国）、时间（2018年4 月3日）、商品（大豆、汽车、水产品）和机构（中华人民共和国商务部）。命名实体识别任务为自然语言处理领域的基本问题之一，其可以作为许多下游问题的前驱，如情感分析[2]、信息检索[3]、推荐系统[4]、知识图谱[5]等。命名实体识别通常可以被分解为两个子任务，即实体边界的识别和实体的分类，其中实体边界识别是指逐字符判断其是否属于一个实体，而实体分类则是将识别出的实体对应至已有的识别类别中。

命名实体识别任务一般可以转化为序列标记问题（Sequence labelling or named as tagging）[6]。其实，自然语言处理中的众多任务其本质上均可以看作为序列标记，如分词，词性标注等。因此序列标记的基本方法也均可以迁移至命名实体识别的任务中。早期对于命名实体任务的解决方法主要包括基于规则的抽取方法（在实际操作中则主要体现在规则，即正则表达式模板的书写和匹配）以及基于统计的方法如，隐马尔可夫模型（Hidden Markov Model，HMM）[7]、最大熵分类模型（Maximum Entropy）[8]、条件随机场（Conditional random field，CRF）[9]等。近年来随着深度学习的强势崛起，得益于深度神经网络强大的特征表示能力，命名实体识别任务的精度不断被提高，其中最具代表性的网络结构如Bi-LSTM+CRF[10]模型。

命名实体识别是一项极具实用价值的技术，目前中英文上通用命名实体识别（人名、地名、机构名）的F1值均能达到90%以上[11]。而命名实体识别的主要难点在于表达不规律、且缺乏训练语料的开放域命名实体类别（如电影、歌曲名），即未登录词的识别，以及歧义的问题。其中未登录词的识别是影响命名实体识别任务效果的主要原因。根据SIGHAN Bakeoff数据评测结果显示，未登录词造成的精度损失至少是歧义的5倍以上[12]。此外现有的评测结果均是在特定的数据集上使用标注语料训练的结果，而当面对真实复杂环境时，现有的模型和方案其效果均会大打折扣。因此如何在海量、异构的数据中进行命名实体识别又或是信息抽取是目前的研究难点和主要的研究方向。

1. 命名实体识别研究背景及历史

现在普遍认为命名实体识别任务于1996年举办的第六届信息抽取研究会议（Message Understanding Conference, MUC-6）上正式被提出[13]。MUC会议主要由美国国防部主导，其主要关注于从非结构化的文本，如报纸、文章中抽取出结构化的信息（与公司和国防的相关的活动）。随着研究的进行，学者们发现对于人名、组织机构名、地名等专有信息，以及数字，如时间、日期、货币、百分比这类信息十分关键，故将上述7种类别的信息抽取，即命名实体识别任务作为MUC的子任务单独提出。而对于信息抽取（Information Extraction）的研究，学术界和工业界早在上世纪60年代中期就已开始，如上世纪60年代中期纽约大学Linguistic String项目，旨在从医疗邻域的相关报告（X光报告、出院报告）中抽取和格式化信息[14]；上世纪70年代耶鲁大学Gerald De Jong根据故事脚本理论建立信息抽取系统，同时从新闻报道中抽取信息，其内容涉及地震、工人罢工等多个领域和场景[15]；上世纪80年代的ATRANS系统实现银行转账信息的自动处理[16]；卡内基集团为路透社开发的JASPER系统实现从新闻那种分析公司的收入与盈利情况[17]，以及美国通用公司开发的SCISOR系统，分析公司合并与盈利情况[18]。

MUC-6中命名实体识别任务使用30篇新闻报道作为数据，其语料库较小且类型单一，因此参会方法的识别效果均就好，其F1值最高达到96.42%[19]。1996年春进一步举办了MET（the multilingual entity task）命名实体识别任务，其将语种扩展至汉语、日语和西班牙语，其中汉语任务的F1值最高仅为84.51%[20]。1998年又相继举办了MUC-7和MET-2，其在MUC-6的基础上进一步扩展了训练语料的规模，其共包含100篇SGML标注格式的新闻领域文档，共涉及1880个组织机构名，1324个地名，887个人名。在MUC-7中英文最高的F1值为93.39%[21]。MUC会议于1998年结束，此后由美国国家标准技术局主导的ACE（Automatic Content Extraction，1999-2008）评测任务将实体类型扩展至人名、地名、机构名、地理-政治实体、设施这五类。同时其语言形式也从早期的文本扩展至语言和图像部分。2004年ACE在原有基础上进一步扩展其研究领域，包括：实体识别（EDT）、关系识别（RDR）和事件识别（VDR）[22]。其中实体识别新增加了两种实体类别，分别为交通工具（vehicle）和武器（weapon）。对于英文命名实体识别任务其常见的数据集如表1所示：

表1 英文命名实体识别主要数据集[23]

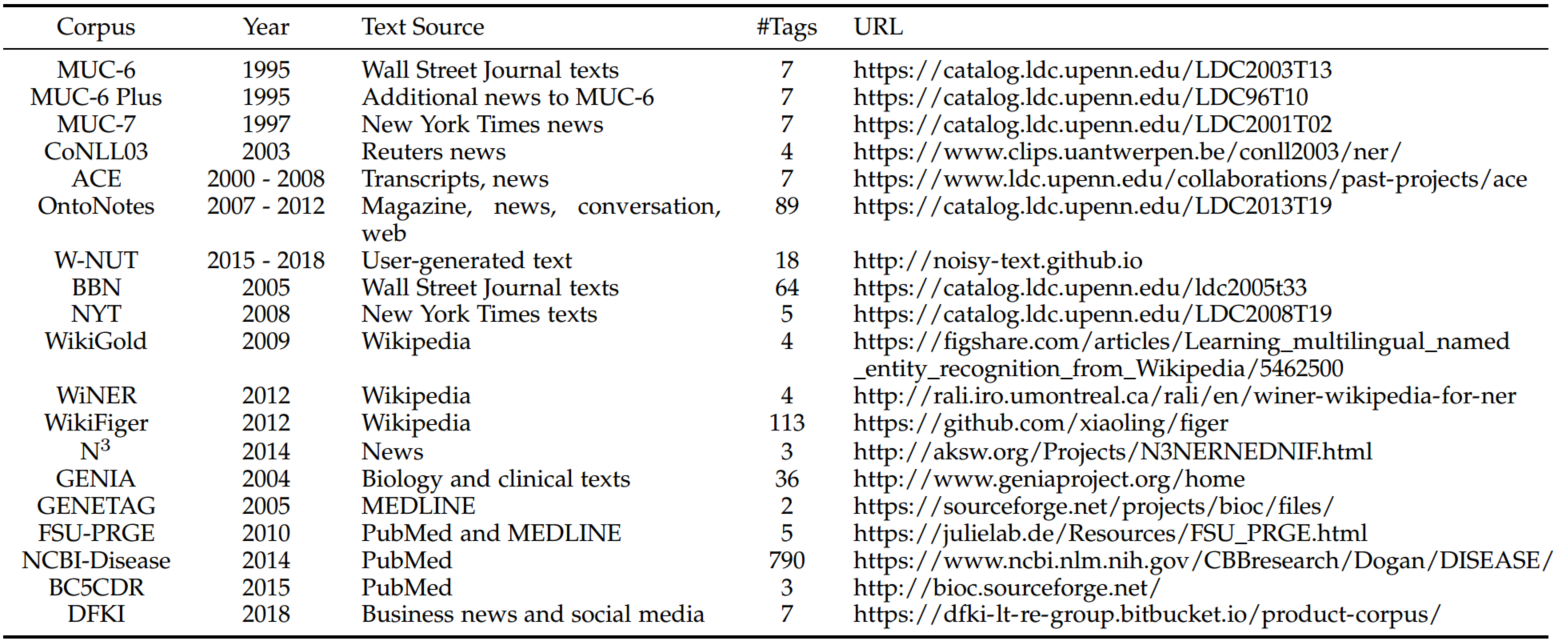


表1所列的数据集中，2005年之前数据主要来自于新闻领域，其规模均较小且实体类别较粗。此后随着互联网技术的飞速发展，Wikipedia数据集、tweets、YouTube中用户评论数据集（W-NUT）相继被创建，其实体类别也从早期的MUC-6的7类扩展至OntoNotes中的89类。

中文的命名实体识别任务相较于英文，其更为复杂和困难。这主要是因为：第一，中文词语间并没有像英文单词的明显边界，其命名实体识别效果很大程度上受制于分词任务的好坏[24]。中文命名实体识别任务最早于2003年举办的863评测，当时最优的结果其评测F1值仅为82.38%[25]。此后，2006年由ACL兴趣小组发起的SIGHAN（the special interest group for Chinese language processing）评测任务成为了研究中文命名实体识的主要途径之一[26]。

1. 命名实体识别的主要方法

早期命名实体识别任务均是基于规则匹配和统计的方法。其中基于规则匹配的方法主要是根据专家即语言学家的专业知识，手工编写规则匹配字符串。而基于统计的方法则主要是指隐马尔可夫模型、最大熵模型、条件随机场以及支持向量机等。2001年Bengio首次提出利用神经网络解决自然语言处理领域的问题，同时给出了词向量训练的最初模型NNLM。其后随着深度学习的不断发展，深度神将网络在NLP的各个任务中被广泛应用。其中对于命名实体识别任务而言，较为经典方法为Zhiheng Huang, Wei Xu[47]等2015年提出的双向LSTM结合条件随机场的网络模型。此后，又不断有新的模型被提出，包括基于CNN、GRU、Transformer以及最近利用Bert预训练等。

3.1 基于规则的方法

基于规则的方法多采用语言专家手工构造特征模板和词典进行字符串匹配识别命名实体，其中特征的选择主要包括：统计信息、标点符号、关键字、指示词和方位词、位置词、中心词等。规则主要包括主要词形、词性、词汇规则[27]，短语规则[28]。如：

Word + {City, Forest, Center, River} / Sherwood Forest

Word + {Street, Boulevard, Avenue, Road} / Portobello Street

……

在早期的MUC-6的评测中，几乎所有的研究均采用了基于规则的方法，即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[29]。此外，也有学者试图利用模型算法自动构造规则，如Kim提出利用Brill规则推断方法自动构造规则[30]。



图1 Bootstrapping流程图

Collins等提出DL-CoTrain方法，其首先根据预先定义的小部分种子规则集Decision List对语料进行无监督训练生成更多规则，再将生成的规则并入至规则集中进行下一轮训练，如此反复迭代至规则集满足需求，并最终将其应用至命名实体识别的任务中，实验表明该方法对人名、地名、组织机构这三类命名实体的识别准确率均超过了91%[31]。此类方法还包括利用Bootstrapping进行规则自动生成方法[32]。此类方法为无监督学习，其将规则的生成转化为聚类问题，在专有领域这类规则缺乏且规则编写、开发成本较高的命名实体识场景中得到了广泛的应用，然而该方法有一个较为显著的缺点，即后一次规则的确定需依赖之前生成的规则，若在迭代的初始阶段，规则的生成明显有误，则该方法生成的规则质量将较差，即此方法的自我纠错能力较差。与此同时，也有学者提出了规则、统计相结合的方法，如NER系统[33]。

对于专有领域的命名实体识别，如生物医学，Hanisch等人提出ProMiner模型，根据预处理的同义词词典识别生物医学文档中蛋白质和潜在基因[34]。Quimbaya提出了一个基于字典的命名实体识别模型对电子健康档案中的专有名词进行识别[35]。其它一些基于规则的命名实体识别系统还包括LaSIE-II[36]，NetOwl[37]，Facile[38]，SAR[39]，FASTUS[40]，LTG[33]等。

基于规则的方法主要的缺点在于规则的编写需要依赖大量的专家知识，且该方法耗时费力。此外规则的“严格”程度将直接影响最终结果的查准率与查全率。若规则较为“严格”则命名实体的识别准确率将较高，而覆盖率则将下降。反之，其准确率将较低而查全率将较高。总之在Precession和Recall间存在“跷跷板”现象。此外，虽然目前深度神经网络这种方法以被普遍采用，但是在专有领域的命名实体识别任务，尤其是在数据量较少时，规则结合神经网络的方法仍被广泛采用。

3.2 基于统计的方法

基于统计的方法并不需要规则的驱动，而只依赖需大量的标记数据且能在短时间内完成。其代表方法主要包括，隐马尔可夫模型，最大熵模型，条件随机场和SVM支持向量机等。在CoNLL-2002会议上，几乎所有的模型方案均采用基于统计的方法，其中Carreras等人提出的AdaBoost.MH取得了最好的识别效果，其在西班牙和荷兰语上的F1值分别达到了81.39和77.05[41]。

（1）隐马尔可夫模型

马尔可夫模型主要是用于描述系统状态间的转移过程，即系统随时间或空间的推移，从某一状态转移至另一状态。在马尔可夫过程中其假设系统在t时刻的状态只与上一时刻即t-1时刻有关，而与之前的状态无关。其中马尔可夫模型的输入为状态转移序列，模型参数为状态转移矩阵（一般根据历史数据统计可得），模型的输出为状态序列的概率值。在现实生活中对于满足马尔可夫假设的状态随机转移过程，我们均可由马尔可夫模型表示。如天气的变化，产品生产线质量的变化，硬币的抛掷结果（满足二项分布）等等。这里需要注意的是马氏过程描述的是状态间的随机转移过程，其随机性由概率刻画，而对于确定的状态转移过程，如红绿灯的变化一般不用马氏过程描述。

隐马尔可夫模型作为概率图模型（也即生成式模型）（probabilistic graphical model）的典型代表（概率图模型根据变量间的相关关系一般可以分为贝叶斯网络（Bayesian network，变量间存在显示的因果依赖关系，用有向无环图表示）和马尔可夫网（Markovn network，变量间存在相关性但因果关系难以获得，用无向图表示。）是一种结构最简单的动态贝叶斯网络。其在语音识别、自然语言处理、时序数据建模、生物信息（基因序列）、模式识别以及故障诊断分析等领域均有广泛应用。隐马尔可夫创建于上世纪70年代，是关于时序的概率模型，其描述的是由一个隐藏的马氏链随机生成不可观测的状态序列，再由各个状态随机生成一个观测而产生的观测序列的过程。其实也就是说在一个系统中同时存在两条序列，一个是我们可以直接通过观测获得的，称之为观察序列（显示的），而另一条则是状态转移序列即马氏链（隐示的，不可知），该序列我们并不能通过观察直接获得。但是状态转系序列会直接影响观察序列的结果，因此我们要做的是通过观察序列去推断状态序列最可能的结果，即通过可见的事物的变化揭示深藏其后的内在的本质规律。

对于隐马尔可夫模型其一般存在二类问题：①评估问题，给定观测序列，求观测序列的概率。评估问题主要对应序列标记中分词的消歧问题，即词出现多种可能时，求解观察序列中概率最高的结果，此时观察序列则为多个分词结果，而状态序列则为句子序列。评估问题主要利用前向、后向算法求解。②解码问题，给定观察序列及模型，选择一个概率最大的状态序列，使其能够合理解释观察序列。解码问题主要对应命名实体识别，词性标注等问题。解码问题主要利用Viterbi算法求解。Bikel等人首次将隐马尔可夫模型应用至命名实体识别的任务中[42][43]，并取得了不错的效果。

然而，隐马尔可夫模型输出独立性假设要求观察序列数据严格相互独立才能保证推导正确性。而正是因为这一假设导致其在自然语言处理中不能考虑词语与词语间的上下文特征，此时在序列标注问题中将有可能出现多个介词连用等错误结果。因此，为解决这一矛盾，研究人员在最大熵模型的基础上提出了条件随机场模型。

（2）最大熵模型

最大熵模型通过指定的特征集合集合即可得到输出目标的概率分布，在评估类问题中，最大熵模型相较于其它模型如AHP层次分析法，其模型参数（特征函数的权值）通过统计学习所得而不依赖主观意志确定，因此具有较强的说服力。在命名实体识别任务中，一般通过设计特征模板抽取特征，并进行统计得到最终结果[44]。由于最大熵模型中各个输出也相互独立，互不影响。这与自然语言文本序列上下文存在依赖关系的事实并不相符。因此需引入条件随机场模型。

（3）条件随机场模型

为克服最大熵模型中输出相互独立的缺点，考虑当前位置对应的标签值与前一位置对应的标签值对后验概率的影响，对此在两个相邻输出的结果间引入边上特征函数。同时考虑当前位置对应的标签值对后验概率的影响，在各个结点上引入结点特征函数（其对应最大熵模型特征函数），进行求解。Mc-Callum和Li提出的特征归纳的条件随机场模型在CoNLL03英文语料中F1值达到了84.04%[45]。此外Krishnan和Manning提出的基于双CRF的两阶段分类器也取得了不错的性能[46]。利用CRF实现命名实体识别流程图如图2所示。



图2 条件随机场模型实现命名实体识别流程图

对比隐马尔可夫模型、最大熵、条件随机场、SVM支持向量机这四种模型，最大熵模型具有更好的通用性和可扩展性，但是其时间复杂度较高，且由于特征归一化计算的存在，其训练计算代价较大。条件随机场相对最大熵模型其结果更优，但是也存在收敛速度慢、训练时间长的缺点。相较于隐马尔可夫模型，最大熵和支持向量机一般来说其精度更高，但是隐马尔可夫模型的训练和识别速度更快，因此其更加适用于如信息检索等这些大批量文本处理和短文本命名实体识别等任务。

此外基于统计的命名实体识别方法其效果的好坏很大程度上依赖于特征模板选择的好坏。而特征模板的确定一方面依赖研究人员的经验知识，另一方面需参考对训练语料库的统计分析以及特征挖掘。其中常见的特征主要包括：单词特征、上下文特征、词典及词性特征、停用词特征、核心词特征以及语义特征和句法信息等。相较于英文，汉语的常见特征一般还包括提示单字、常用尾字等，同时其分词结果的好坏也直接影响汉语命名实体识别。基于统计的命名实体识别对语料库的依赖程度也较大，而可以用来建设和评估命名实体识别系统的大规模通用语料库又相对较少。尤其是对于中文语料，在SIGHAN Bakeoff08测评中，中文命名实体识别使用的语料库主要包括香港城市大学语料库（训练集1772202字）、 微软亚洲研究院语料库（训练集1089050字）、北京大学语料库（训练集1833177字），上述语料库的规模均较小，无法支撑大型命名实体识别项目的开发。

3.3 基于深度学习的方法

神经网络的发展经历了“三起两落”，其最初起源于Warren McCulloch和Walter Pitts于1943年MCP神经元数学模型，该模型本质上是在进行加权求的运算。1957年，Rosenblatt在其基础上发明了单层感知机模型[48]，引起了学术界一时的轰动。然而1969年，Minsky和Seymour Papert发表专著指出，单层感知机连最基本的非线性“异或”问题也不能解决[49]，这直接宣判了“连接主义”这类方法的死刑，其后神经网络的发展进入了寒冬。1980年日本学者Fukushima受Hubel和Wiesel猫的视觉皮层实验的启发[50]，提出了Necognitron[51]，当时由于BP算法还并未被正式提出，因此网络的训练为求解能量函数最小，而其中类似局部感受野和“卷积”的思想直接启发了LeCun LeNet网络的发明。1982年John Hopfield教授提出Hopfield网络，成功解决了旅行商这类NP难问题[52]，而引起了学术界的轰动。1983年Hinton，Sejnowski设计了玻尔兹曼机[53]，为前馈神经网的提出创造了条件。1986年，Hinton等人在此发明了反向传播算法[54]为后续神经网络的发展铺平了道路。1989年LeCun首次利用神经网络实现了手写数字的识别[55]并成功应用至美国的邮政系统中，并于1989年正式提出LeNet5[56]，这成为了其后卷积神经网模型的基础。同年，Hornik K提出万能逼近定理，即只要隐层神经元数目足够多，即使单层的网络也能逼近任意的函数[57]，这为神经网络的发展提供了一定的理论基础。然而受限于计算机硬件的性能和训练数据的缺乏，神经网络的方法并没有得到广泛的关注，其后又被兴起SVM支持向量机这种浅层模型所取代。2006年被称为深度学习的“元年”，Hinton在Science上发文指出深度信念完可以通过逐层训练的方式完成，其相较于与传统的PCA、SVM有更好的特征表示能力[58]。有趣的是，同年Benjio和LeCun也分别在NIPS上发表有关深度网络的方法[59][60]为其证明。直至2012年，得益于计算机硬件即GPU性能的提升和数据的积累在ImageNet竞赛中，AlexNet[61]以领先第二名传统方法近11个百分点的优势一招横扫了整个CV领域，深度学习开始强势崛起。其后VGGNet[62]、GoogLeNet[63]、ResNet[64]等新模型不断被提出。深度升级网络在各个领域开始“遍地开花”。

在自然语言处理领域，传统的神经网络结构主要以RNN循环神经网络为主，该方法在处理文本序列较长时，由于前后文依赖关系不能被很好的表示，因而性能较差。因而在此基础上发明了长短时记忆神经网络LSTM模型[65]，其通过输入门、输出门和遗忘门这三门来控制信息的选择和“遗忘”，以期提升网络的性能。然而由于LSTM结构复杂，故学者们将输出门和遗忘门合并为更新门又提出了GRU模型[66]，该网络在保持与LSTM精度相同的前提下，减少了模型的参数和计算量。然而，当序列长度超过30时，LSTM的性能仍旧会急剧下降，其序列中的长时依赖仍旧不能被很好的表示，因此学者们又引入了Attention机制[67]。Attention机制是在Encoder-Decoder[68]这一框架下提出的（该框架主要是为在处理sequence2sequence这类任务如机器翻译，避免词对齐而创造的。其主要是利用RNN将序列编码为中间向量，然后在通过解码网络获得最终结果），其主要是通过Softmax与加权求和操作使中间向量根据不同的输出和上下文动态变化来解决长时依赖的问题。传统的RNN这类序列模型，下一个时刻的输出均需依赖上一时刻隐藏单元的状态，因此无法很好的支持并行，因此Google提出了Transformer的注意力机制[69]，该模型抛弃了传统的序列建模的思想，其本质上是一个新的编码机制，即让前后文词向量间发生一定的运算而产生关系，通过self-attention学习句子间的词法依存关系，而词语词间的位置关系则通过位置编码实现，transformer在解决长序列上下文依赖问题中取得了很好的性能而被广泛采用。

在计算机视觉领域基于ResNet、VGG网络的预训练模型找已得到了广泛的应用。而在自然语言处理领域，词向量即可以认为是一种简单的预训练模型。词向量最早是利用one-hot编码表示，但是由于该表示过于稀疏，存在维数灾难的问题其存储空间较大，因而并未被广泛采用，后续又逐渐提出了词袋模型（该词在文档中的出现次数）和TF-IDF（反映该词在文档中的占比），然而，其很难解决语义鸿沟的问题。基于分布式假设Bengio于2001年正式提出神经网络语言模型[70]（Neural Network Language Model， NNLM）并创造了词向量的概念，即以one-hot编码为前驱查阅look-up表将词编码为稠密的向量，而look-up表的设计即由神经网络训练所得。同时，根据输入输出的不同，词向量网络由分为NNLM、C&W、CBOW和Skip-gram等模型[71]（CBOW与Skip-gram与NNLM相比除去了隐层和词序，使模型得到简化）。此外，由于CBOW、Skip-gram等模型其Softmax的计算与参数的更新需要对整个词表进行，因此其时间成本较高，对此又提出了Hierachical softmax，通过构建哈夫曼树而每次只对少部分结点进行更新和计算以减少复杂度。对于一次多义的问题如“Bank”等，也有学者提出根据不同上下文对一个词编码多个词向量，同时通过对上下文进行idf加权进行k-mean clustering最终确定词向量。此外也有实验表明将不同结构模型如RNN与CNN等学习而得的词向量进行concatenate，对downstream task结果的提高也有帮助。然而，由于受限于语料库的规模，在自然语言处理中预训练的方法早期仅局限于词向量这一微小的方面，大规模文本的预训练模型一直无法进行。直到2017年下半年，基于ELMo[72]和GPT[73]等方法提出的Bert[74]模型使预训练在自然语言处理中广泛应用成为了可能。

对于命名实体识别这一NLP的子任务，其深度神经网络模型基本上也均为上述的方法，如早期15年的Bi-LSTM+CRF模型、CNN模型，以及最近的利用Bert预训练的网络等。

（1）Bi-LSTM+CRF for NER。基于统计的命名实体识别模型其性能与特征模板的选择密切相关。得益于深度神经网络强大的特征表示能力，深度学习在NER中也得到了广泛的应用。早期利用神经网络解决序列标注问题，虽然其在模型的构建中考虑了上下文信息，但其最终的输出结果本质上仍相互独立，其并不能很好的避免类似多个词连续标记为介词这种问题的发生，为解决这一问题，学者们在神经网络中引入CRF模型。即在传统的双向LSTM网络输出结果后再拼接一层CRF，而LSTM部分主要作用为特征的表示和抽取，以代替CRF模型中特征模板的设计工作。其网络结构如图3所示：

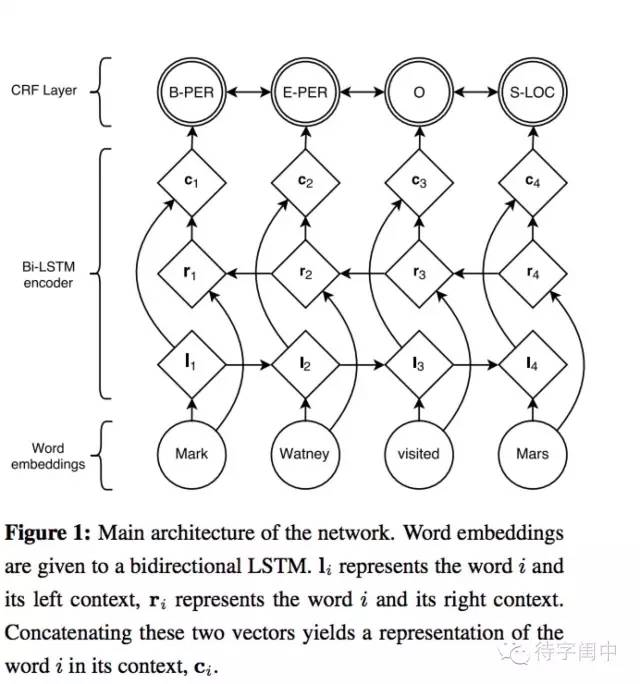


图3 Bi-LSTM+CRF网络结构示意图[75]

从图3可以看出首先对逐个单词编码为词向量，然后送入至双向LSTM，分别提取其前、后文信息，并将融合了前后文信息的编码向量进行拼接后送入CRF层，网络通过训练学习CRF边上状态矩阵最终得到输出标签。Bi-LSTM+CRF模型作为命名实体识别任务中的经典模型，在很多不同领域被广泛采用。此外之后的很多的命名实体识别网络模型虽然其在特征的抽取和表达结构上有所差异，如CNN、RNN、Transform等，但在最后一般均会外接一层CRF以提高最终结果的精度。

（2）CNN for NER。使用CNN网络提取文本特征，其主要做法是首先进行word embedding操作，然后开窗滑动提取特征（类似n-gram模型），然后在送入RNN、双向RNN或直接进行pooling操作。最终得到输出的各个标签的概率分布。Strubell等人提出ID-CNN网络[78]，其利用空洞卷积在不增加额外计算的情况下捕获句子的长时依赖（空洞卷积在CV领域也被用作扩大感受野），因此其在对长序列的处理时其精度要优于传统的CNN网络，而并没有引入巨大的计算代价，同时相较于BiLSTM-CRF网络，其在测试阶段获得了14-20倍的加速比，同时精度相当。

（3）特征的混合表示。对于英文语言我们不仅可以对逐个单词进行编码表示，我们同时也可以对各个字母进行编码表示，此外对于不同网络结构如CNN，RNN对词向量的表示也不相同，甚至可以将RNN或LSTM的隐状态特征等上述特征表示进行拼接、混合，实验表明这些操作对最后结果有一定的提高[77]。Huang等人在BiLSTM-CRF模型的基础上利用拼写特征、上下文特征、词向量以及地名特征这四类特征提高了NER的精度[76]。

（4）半监督的命名实体识别。上述的种种方法均是在有监督的条件下进行学习，这需要大量的有标记数据。然而在现实的很多情况下，我们缺少大量的有标记数据，因此，半监督或弱监督学习是未来研究的方向。在ACL2017中Peters提出的TagLM[79]模型通过使用大量无标注的语料库训练Bi-LSTM模型，并通过该模型获取当前要标注的词向量，并将其作为特征加入至BiLSTM-CRF中。实验表明在数据量较少的情况下，LM-LSTM-CRF模型对于命名实体识别任务其性能获得了较大的提升[80]。其结构如图4所示：

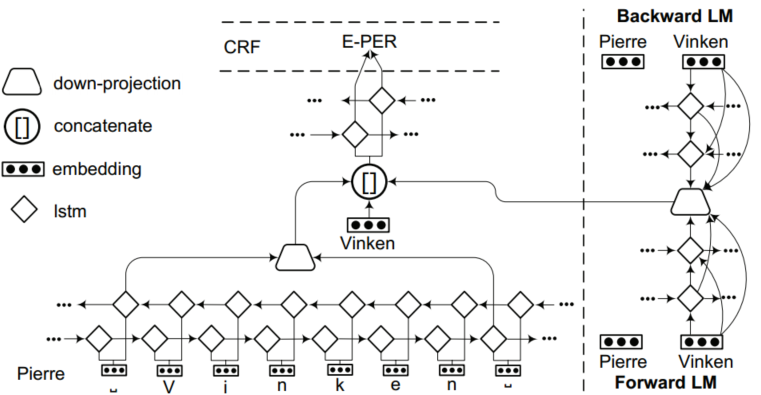
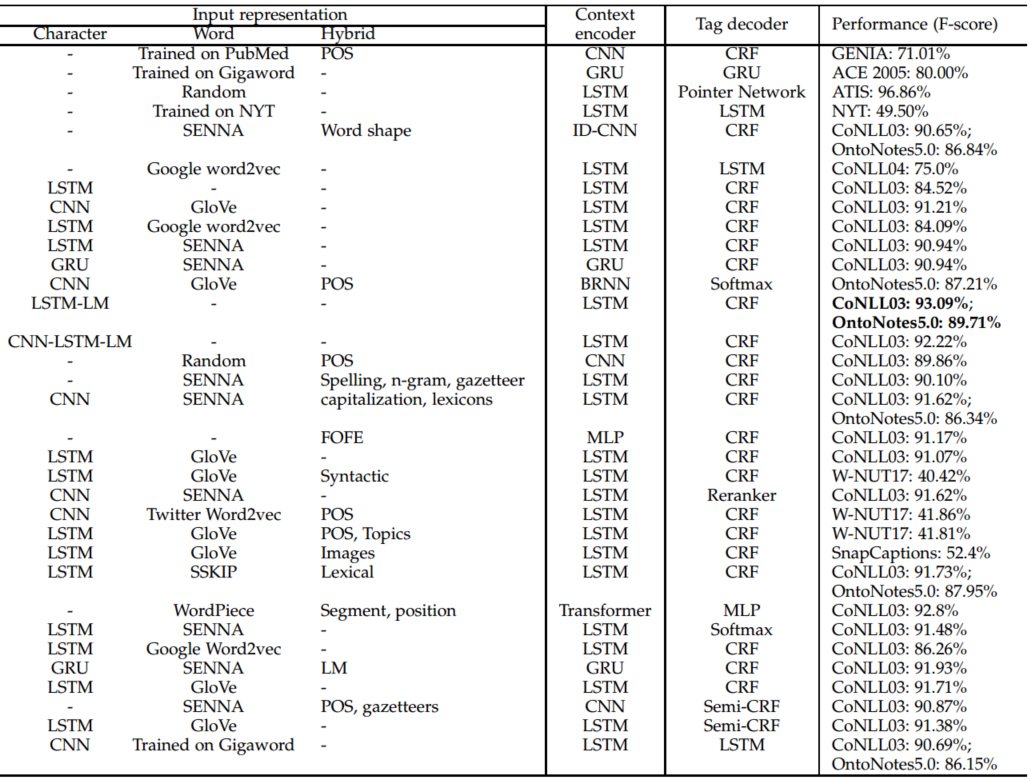


图4 LM-LSTM-CRF模型结构图

从图4可以看出，首先利用LM网络在特定语料库上训练双向的LM网络，然后利用该网络对需要完成标记的序列中的各个词进行特征表示，同时将此特征与word embedding和Bi-LSTM所得的特征表示三者进行拼接，然后再次送入到双向LSTM中，最后经过CRF层得到最终输出。此外实验表明采用深度的双向LSTM网络结构其不同的隐层将会表征不同的特征信息，且在浅层的LSTM中引入多任务（如POS tags）对更高层任务精度性能的提上将会有较大的帮助Belinkov等人[81]指出在两层的LSTM中第一层的编码比第二层在POS tag任务中效果更优。与此类似的工作还包括一些基于预训练的模型，如Bert、GPT 2.0等模型在命名实体识别中的应用。总之，目前提高任务精度的一种常见做法就是收集和融入大量的外部信息，如2019年的GPT 2.0就是在之前的基础上融入了更多高质量的信息。

神经网络在命名实体识别中的应用主要模型包括LSTM、CNN、GRU及其与CRF的结合等，如表2所示：

表2 深度学习中常见的网络模型[23]



从上图可以看出截止至2018年，Akbik等人提出的LSTM-LM模型在CoNLL03数据集上英文命名实体识别最高的F1值为93.09%。而BERT模型在NER任务中的F1值为92.8%。截止至2019年5月，命名实体识别任务精度再次被刷新至93.5%，该结果由Alexei Baevski[82]等人提出的双向Transform+pre-training模型取得，该模型与Transform模型相比并无很大差异，只是在pre-training中进行了大量的实验和调优。

3.4 命名实体识别的最近工作

（1）Contextual String Embedding for Sequence Labeling[83]

Contextual String Embedding for Sequence Labeling这篇文章对于序列标记任务最大的创新点是对传统Embedding的改进，即基于上下文利用Bi-LSTM中的hidden state结合传统Embedding编码词向量。同时与2017年Peters et al.提出的抽取hidden states 作为NLP tasks方法相比，该方法在速度上有明显提高，同时再次刷新了CoNLL03的baseline，其F1值达到了93.09超过了Bert。其模型结构如图5所示：

从图5可以看出论文提出的模型主要包括两个部分，即Character Language Model和Sequence Labeling Model。其中Character Language Model即利用双向LSTM中hidden state编码词向量，由于Bi-LSTM的使用使得该Embedding天生就融合了上下文的语义信息，因此能很好的解决一词多义的问题。同时由pre-training得到的Character Language Model可以根据不同的输入动态编码词向量，因此其能很好的嵌入至downstream任务中。Sequence Labeling Model即为传统的BiLSTM+CRF的结构。

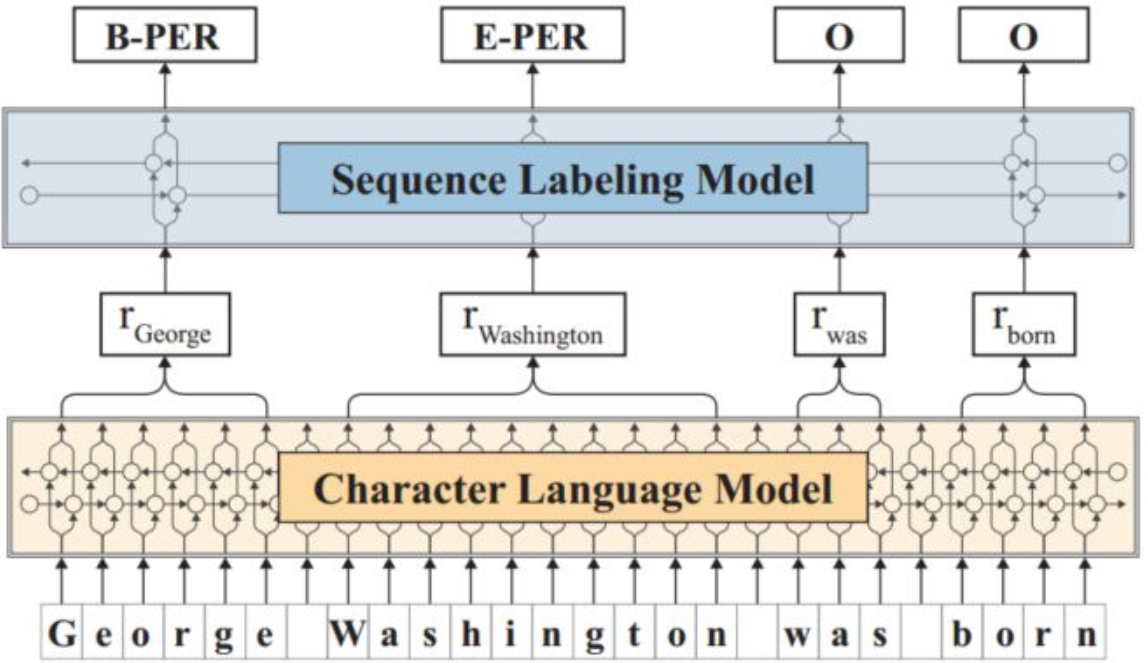


图5 模型结构图

该工作在众多sequence labeling任务中均达到了最好结果。总的来说，目前对于word Embedding主要分为两种方法：（1）利用更多的external information，如基于大量高质量数据预训练的Berth，或针对其它特定数据集预训练的word embedding；（2）利用神经网特征表示和选择的能力，根据hidden state表示word embedding，以及对各种embedding进行拼接。总之如何利用更多的外部信息以及更好的表示特征是仍需研究的方向。

（2）Chinese NER Using Lattice LSTM[84]

这篇文章为2018年ACL的工作，其主要是针对中文的命名实体识别任务。对于英文这类由拉丁字母组合而成的语言，其单词与单词间存在天然的边界，而每个单词又是由不同的字母组成，因此对于英文的NER其不需要进行分词处理，而且将各个character进行embedding的操作早已有之。然而对于中文而言首先分词质量的高低往往就直接影响了下游任务的性能（对于分词任务其主要的难点为OOB问题，即对于未登录词的处理，其次是歧义的问题），但是若直接使用单个汉字则又无法充分挖掘句子中潜在的句法结构语义信息，其效果往往不好（但是最近ACL2019有文章指出，对与中文的许多任务不分词的效果要好于分词的效果），因此对于中文NER其效果相对英文较差。之前中文的NER model一般均是直接利用word（分词的结果）或character（单个字）作为input，而本文作者则创新性的提出了Lattice LSTM model将character和word的embedding结合实现中文的NER，如图6。该方法在许多数据集上均获得了较好的结果。

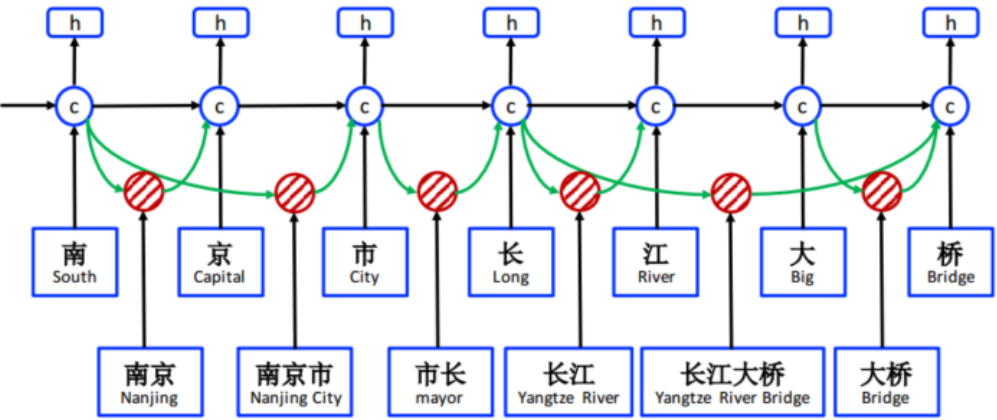


图6 模型结构图

如图6所示，该模型大体上可以分为四个部分：character-based model实现对单个字的表示；word-based model实现对词的表示；lattice model字词的联合表示；最后在再CRF得到最终结果的概率分布。其中character-based model和word-based model均是embedding+BiLSTM的结构。

对于该文章，作者利用LSTM对character和word进行特征表示，然后加权求和。该方法其本质上还是希望挖掘字、词间的潜在信息和特征。现在来看，融合更多的先验知识和外部数据信息对于精度的提升效果更加显著且目前各大公司都在第二个方向投入了大量的成本，但是对于自然语言处理，是不是说通过数据的累积就能不断的提高任务精度，而不会是向image classification COCO数据集那样存在瓶颈呢？就目前来看现在的自然语言处理的各种模型虽然在某些小部分数据集上表现得很好，但是其文字背后真正的含义却根本没有理解，而想要实现真正的人工智能，其还得需要更棒的方法，尤其是在生成式任务中要有所突破。

（3）Neural Adaptation Layers for Cross-domain Named Entity Recognition[85]

自然语言处理的很多任务中一个很普遍的问题即对于不同邻域间的迁移学习。目前的大多数模型或研究方法均是针对某一些较小邻域的数据集所设计的，而所获得的精度和结果也均是针对特定邻域调出来，当将这些模型运用到其它的数据中时其性能将大大折扣。因此对于迁移学习的研究至关重要，目前对于NER任务主要的transfer learning模型包括INIT和MULT等，如图7所示：

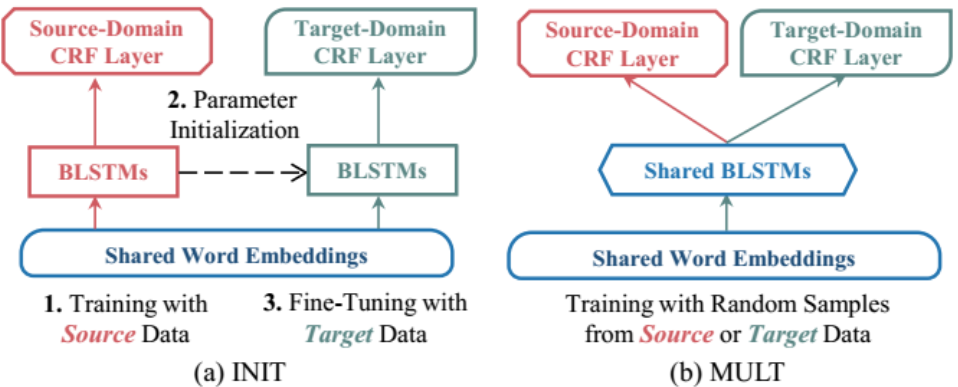


图7 INIT与MULT模型结构

从上图可以看到，对于INIT模型对于source domain和target domain均使用相同的word embedding，然后利用target label data在pre-training的model（BLSTM+CRF）进行fine-tuining得到目标邻域的模型。而MULT则是利用multi-task learning，根据source domain和target domain data同时训练source model和target model，而其除CRF层外其它参数共享。而上述这两种模型均存在以下三种缺点：

- target domain和source domain的词汇含义尤其各自专有邻域的语境决定，并不会完全相似。例如，对于模式识别邻域LDA可能是指线性判别分析，而对于文本数据挖掘邻域LDA更多是指潜狄利克雷配分模型。此外专有邻域的OOV（out-of-vocabulary）问题也是一大挑战。若简单的让source domain和target domain公用相同的word embedding显然是不太合适的。

- 为克服问题1现有的模型往往根据target domain data对source domain word embedding进行retraining。然而往往source domain corpora十分庞大，若要在此基础上进行再训练其成本往往较高，因此这将在一定程度上限制迁移学习的发展

- 现有的模型如INIT其只对最后的CRF层进行再训练，且任务通过CRF就能较好的捕获不同邻域下的语义信息，这显然并不具有十分充足的说服力。

因此针对上述三类问题，作者在原始的BiLSTM+CRF的基础上分别增加了word adaptation layer, sentence adaptation layer和output adaptation layer，最终取得了更优的结果，如下：

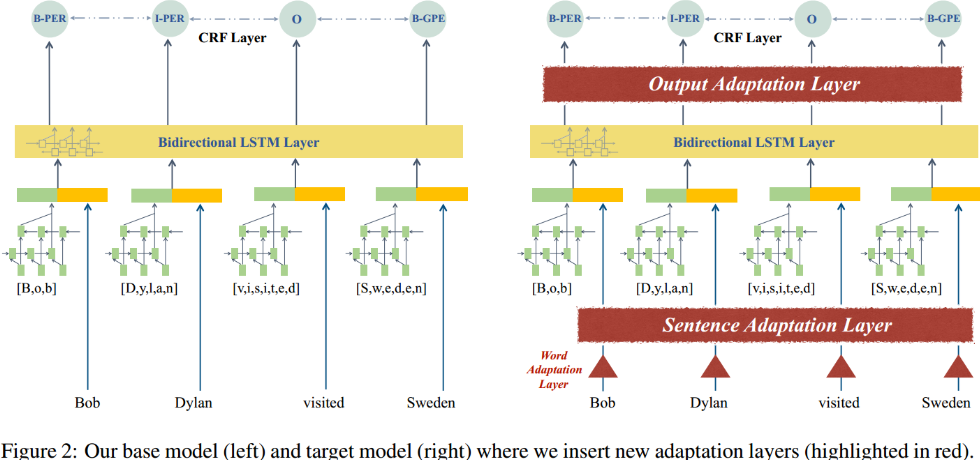


图8模型结构

对比之前模型，作者提出的新方法确实是提高了最后的结果。但是其真正有说服力和较为创新的工作为word adaptation layer的提出，其在一定程度上解决了上述的三个问题。而对于后两个tricks其说服力不足，而且在实验中作者仅是增加了模型的hidden unti的数目，即“加宽”了模型。但是实际作者的设计是“加深”模型，而本来有实验说明在同样参数数目的情况下，“加深”模型所带来的收益要比“加宽”模型所带来的收益更多，因此其不能充分的证实该结构的有效性（个人更倾向于加深模型所带来的切实收益，而非作者间的解释的获得了更多的上下文信息）。此外，对于word adaptation layer其还能应用至其它的迁移学习任务中，而非仅仅局限于NER。

1. 命名实体识别的总结

本文简单接介绍了命名实体识别任务的历史以及目前的主流模型及方法。虽然目前在CoNLL03数据集上其F1值已经达到93%以上，但是在面对专有领域或开放域的命名实体识别时其性能将大打折扣。尤其是当专有领域训练数据不足时深度神经网络仍需结合传统的基于规则、字典的方法，以及一些基于统计的方法。此外，目前来看现在提高NER精度的方法均是结合预训练的方法，然而该方法需要依赖大规模、高质量的数据，且需要算力的支撑。对于我们来说如何更好的设计网络结构，抽取和表示特征更是应该关注的方向。就目前的一些方法来看，无论是利用不同的网络结构（CNN、RNN等）学习word embedding然后进行组合，又或是利用网络的隐层单元状态与word embedding相结合提升网络性能，其本质上是给网络更多维、更丰富的特征表示，但是其精度的提升也有限。这就是说在所有的这些特征中，很多特征并不是十分有效的，而只能靠网络的特征表达能力去学习特征。另外，对比LSTM和Transform这两种模型，虽然在较长的序列上Transform的优势相较于LSTM更加明显，但是在短文本的处理上Transform的性能却比不上这种序列模型。比较Transform和RNN等，其本质上都是让前后文通过某种计算产生关联，而Transform则更像一种编码方式，其在长序列上的优势是否是因为RNN、LSTM这类网络梯度消失导致的呢？又是否能够将残差结构更好的和RNN相结合？此外相比与计算机视觉和语音处理领域，深度学习在NLP领域的应用似乎并没有那么的成功，其中一个原因是英文NLP目前还缺少像ImageNet那样大规模的数据库，另一方面是因为自然语言理解相比与图像并非那么抽象，其还涉及到词法、句法方面的知识，因此在深度学习出现之前就已经有大量的研究。个人认为相较于毫无方向的“炼丹”和各种trick的堆叠，我们其实更应该将语言方面的先验知识更好的引入至网络中，使其更好的提取和表示特征，否则现在的人工智能将很难跨越“语义鸿沟”上升至真正的感知智能，那么其又何现代统计学有什么差别呢？

参考文献：

1. Borthwick A, Sterling J, Agichtein E, et al. NYU: Description of the MENE named entity system as used in MUC-7[C]//Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. 1998.
2. Cambria E, Poria S, Gelbukh A, et al. Sentiment analysis is a big suitcase[J]. IEEE Intelligent Systems, 2017, 32(6): 74-80.
3. Culotta A, Bekkerman R, McCallum A. Extracting social networks and contact information from email and the web[R]. MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE, 2005.
4. Matheny J. Entity recommendation system using restricted information tagged to selected entities: U.S. Patent Application 11/716,012[P]. 2008-9-11.
5. Shen W, Wang J, Han J. Entity linking with a knowledge base: Issues, techniques, and solutions[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 27(2): 443-460.
6. Sasaki Y, Tsuruoka Y, McNaught J, et al. How to make the most of NE dictionaries in statistical NER[J]. BMC bioinformatics, 2008, 9(11): S5.
7. Zhou G D, Su J. Named entity recognition using an HMM-based chunk tagger[C]//proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002: 473-480.
8. Curran J, Clark S. Language independent NER using a maximum entropy tagger[C]//Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. 2003.
9. Councill I G, Giles C L, Kan M Y. ParsCit: an Open-source CRF Reference String Parsing Package[C]//LREC. 2008, 8: 661-667.
10. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
11. Li J, Sun A, Han J, et al. A Survey on Deep Learning for Named Entity Recognition[J]. arXiv preprint arXiv:1812.09449, 2018.
12. Sproat R, Emerson T. The first international Chinese word segmentation bakeoff[C]//Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. Association for Computational Linguistics, 2003: 133-143.
13. Grishman R, Sundheim B. Message understanding conference-6: A brief history[C]//COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. 1996, 1.
14. Grishman R, Sterling J. New York University: Description of the PROTEUS System as used for MUC-5[R]. NEW YORK UNIV NY DEPT OF COMPUTER SCIENCE, 1993.
15. Schank R C, Kolodner J, DeJong G. Conceptual Information Retrieval[R]. YALE UNIV NEW HAVEN CT DEPT OF COMPUTER SCIENCE, 1980.
16. Lytinen S L, Gershman A. ATRANS Automatic Processing of Money Transfer Messages[C]//AAAI. 1986, 86: 1089-1093.
17. Andersen P M, Hayes P J, Huettner A K, et al. Automatic extraction of facts from press releases to generate news stories[C]//Proceedings of the third conference on Applied natural language processing. Association for Computational Linguistics, 1992: 170-177.
18. Jacobs P S, Rau L F. SCISOR: Extracting information from on-line news[J]. Communications of the ACM, 1990, 33(11): 88-99.
19. Krupka G R. SRA: Description of the SRA system as used for MUC-6[C]//Proceedings of the 6th conference on Message understanding. Association for Computational Linguistics, 1995: 221-235.
20. Merchant R, Okurowski M E, Chinchor N. The multilingual entity task (MET) overview[R]. DEPARTMENT OF DEFENSE FORT GEORGE G MEADE MD, 1996.
21. Mikheev A, Grover C, Moens M. Description of the LTG system used for MUC-7[C]//Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. 1998.
22. Doddington G R, Mitchell A, Przybocki M A, et al. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation[C]//Lrec. 2004, 2: 1.
23. Li J, Sun A, Han J, et al. A Survey on Deep Learning for Named Entity Recognition[J]. arXiv preprint arXiv:1812.09449, 2018.
24. 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.
25. 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.
26. Wu Y C, Yang J C, Lin Q X. Description of the NCU Chinese word segmentation and named entity recognition system for SIGHAN bakeoff 2006[C]//Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006: 209-212.
27. Weischedel R, Ayuso D, Boisen S, et al. BBN: Description of the PLUM System as Used for MUC-5[C]//Proceedings of the 5th conference on Message understanding. Association for Computational Linguistics, 1993: 93-107.
28. 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. Association for Computational Linguistics, 1995: 141-155.
29. Grishman R. The NYU System for MUC-6 or Where's the Syntax?[R]. NEW YORK UNIV NY DEPT OF COMPUTER SCIENCE, 1995.
30. Kim J H, Woodland P C. A rule-based named entity recognition system for speech input[C]//Sixth International Conference on Spoken Language Processing. 2000.
31. Collins M, Singer Y. Unsupervised models for named entity classification[C]//1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 1999.
32. Cucerzan S, Yarowsky D. Language independent named entity recognition combining morphological and contextual evidence[C]//1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 1999.
33. 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. Association for Computational Linguistics, 1999: 1-8.
34. Hanisch D, Fundel K, Mevissen H T, et al. ProMiner: rule-based protein and gene entity recognition[J]. BMC bioinformatics, 2005, 6(1): S14.
35. Quimbaya A P, Múnera A S, Rivera R A G, et al. Named entity recognition over electronic health records through a combined dictionary-based approach[J]. Procedia Computer Science, 2016, 100: 55-61.
36. Humphreys K, Gaizauskas R, Azzam S, et al. University of Sheffield: Description of the LaSIE-II system as used for MUC-7[C]//Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. 1998.
37. Krupka G R, IsoQuest K H. Description of the nerowl extractor system as used for muc-7[C]//Proceedings of the 7th Message Understanding Conference, Virginia. 2005: 21-28.
38. Black W J, Rinaldi F, Mowatt D. FACILE: Description of the NE System Used for MUC-7[C]//Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. 1998.
39. Aone C, Halverson L, Hampton T, et al. SRA: Description of the IE2 system used for MUC-7[C]//Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. 1998.
40. 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. Association for Computational Linguistics, 1995: 237-248.
41. Carreras X, Màrquez L, Padró L. Named entity extraction using adaboost[C]//COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002). 2002.
42. Bikel D M, Miller S, Schwartz R, et al. Nymble: a high-performance learning name-finder[J]. arXiv preprint cmp-lg/9803003, 1998.
43. 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.
44. Borthwick A, Grishman R. A maximum entropy approach to named entity recognition[D]. New York University, Graduate School of Arts and Science, 1999.
45. 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 2003-Volume 4. Association for Computational Linguistics, 2003: 188-191.
46. 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. Association for Computational Linguistics, 2006: 1121-1128.
47. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
48. Rosenblatt F. The perceptron, a perceiving and recognizing automaton Project Para[M]. Cornell Aeronautical Laboratory, 1957.
49. Minsky M, Papert S A. Perceptrons: An introduction to computational geometry[M]. MIT press, 2017.
50. Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[J]. The Journal of physiology, 1962, 160(1): 106-154.
51. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological cybernetics, 1980, 36(4): 193-202.
52. Hopfield J J. Neural networks and physical systems with emergent collective computational abilities[J]. Proceedings of the national academy of sciences, 1982, 79(8): 2554-2558.
53. Ackley D H, Hinton G E, Sejnowski T J. A learning algorithm for Boltzmann machines[J]. Cognitive science, 1985, 9(1): 147-169.
54. Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Cognitive modeling, 1988, 5(3): 1.
55. LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989, 1(4): 541-551.
56. LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
57. Hornik K. Approximation capabilities of multilayer feedforward networks[J]. Neural networks, 1991, 4(2): 251-257.
58. Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786): 504-507.
59. Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks[C]//Advances in neural information processing systems. 2007: 153-160.
60. Poultney C, Chopra S, Cun Y L. Efficient learning of sparse representations with an energy-based model[C]//Advances in neural information processing systems. 2007: 1137-1144.
61. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
62. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
63. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.
64. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
65. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
66. Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.
67. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
68. Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.
69. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
70. Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(Feb): 1137-1155.
71. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
72. Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.
73. Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf, 2018.
74. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
75. Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. arXiv preprint arXiv:1603.01354, 2016.
76. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
77. 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.
78. Strubell E, Verga P, Belanger D, et al. Fast and accurate entity recognition with iterated dilated convolutions[J]. arXiv preprint arXiv:1702.02098, 2017.
79. Peters M E, Ammar W, Bhagavatula C, et al. Semi-supervised sequence tagging with bidirectional language models[J]. arXiv preprint arXiv:1705.00108, 2017.
80. Liu L, Shang J, Ren X, et al. Empower sequence labeling with task-aware neural language model[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
81. Belinkov Y, Durrani N, Dalvi F, et al. What do neural machine translation models learn about morphology?[J]. arXiv preprint arXiv:1704.03471, 2017.
82. Baevski A, Edunov S, Liu Y, et al. Cloze-driven Pretraining of Self-attention Networks[J]. arXiv preprint arXiv:1903.07785, 2019.
83. Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 1638-1649.
84. Zhang Y, Yang J. Chinese ner using lattice lstm[J]. arXiv preprint arXiv:1805.02023, 2018.
85. Lin B Y, Lu W. Neural adaptation layers for cross-domain named entity recognition[J]. arXiv preprint arXiv:1810.06368, 2018.