山东大学

语义计算实验室

命名实体识别综述

学生姓名 吴佳琪

指导教师 孙宇清

2021 年 9 月 6 日

目录

命名实体识别综述

摘要

命名实体识别（Named Entity Recognition, NER）指从非结构化文本中识别实体，并将实体分类为预先定义的实体类型。它是信息提取和文本理解中的一项具有挑战性的任务。早期的 NER 系统在设计特定领域的特征和规则时需要付出很大的代价。近年来，深度学习方法被广泛应用于 NER 任务，促进 NER 技术快速的发展。本文首先介绍 NER 领域相关的资源，包括 NER 语料库和现成的 NER 模型工具代码。然后，在细分领域上对 NER 的相关工作进行整理分析，并介绍最具代表性的方法。最后，讨论前沿问题和未来方向。

关键词：命名实体识别；深度学习；命名实体

第一章 背景和意义

1.1 背景

命名实体识别（Named Entity Recognition, NER）是自然语言处理领域十分重要的子任务。命名实体（Named Entity, NE）指由一个或多个严格指示符号来代表对象的实体，该定义最早可以追溯到 S. A. Kripke[50]给出的定义。例如 1949 年成立的中华人民共和国通常被称为“中国”或“中华人民共和国”。这里的“中国”和“中华人民共和国”都是中华人民共和国这个实体的严格指示符号。Kripke 认为严格的指示符包括专有名词以及一些自然种类的术语，例如生物物种或物质名称。之后在命名实体识别领域社区，大家普遍认为一些时间或数字的表达式（例如日期或金额）也可称为命名实体，这些类型的某些实例是严格指示符很好的例子，（比如，2001 年指示公历 2001 年），但也存在不符合定义的例子，（比如 6 月份，可以是去年的 6 月份，今年的 6 月份，任何一年的六月份），这样，命名实体就产生了具有争议的很松散的定义。

在早期的 NER 工作中，命名实体通常为一些通用的专有名词，如人名、地名、组织机构名。自 MUC-6[49]竞赛以来，这些类型的实体统称为“enamex”。同时，在之后的相关工作中，这些实体类型又出现了更加细粒度的划分，比如地名划分为城市、州、国家等，人名划分为政客、娱乐者等。甚至慢慢的出现一些新的命名实体类型，例如疾病名称，以及政府地名“GPE”。

之后的 CONLL 会议中，对命名实体识别任务中命名实体类型进行了扩充，传统的“enamex”类型之外，加入了“product”，“date”以及“time”等新的类型。随之而来的，是越来越多的为了处理特殊任务的边界类型实体的加入。例如电影名、科学家名、电子邮件地址、研究领域、项目名称、书名等等。

最近几年，随着对生物信息学的研究蓬勃发展，越来越多的研究致力于发现生物领域的文本实体，例如蛋白质类型、DNA、RNA、细胞类型等等。以及一些药物领域的药物名称。

随着研究领域的不断扩充，命名实体的概念和定义已经不再是最早那样的严

格的定义了，但总的来说现在命名实体还是有比较显著的特征和属性的，比如命名实体通常在文本中以名词的形式出现，并具有一定特殊的属性和含义。

近年来随着深度学习技术在命名实体识别领域的蓬勃发展，许多优秀的方法都能在数据充足的场景下表现良好。但是在一些低资源或特殊领域上，命名实体的标注数据稀少时，现有的命名实体识别模型很难有好的表现。因此如何设计和训练一个对命名实体人工标记数据有较小依赖的命名实体识别模型成为一个重要的问题。

1.2 意义

命名实体识别不仅是信息抽取的重要工具，而且为信息检索、问答系统、机器翻译等下游任务提供重要的预处理服务。

以语义搜索任务为例，语义搜索是指一组技术，这些技术使搜索引擎能够理解用户查询背后的概念、含义和意图。据统计，大约 70%的查询语句具有至少一个实体。识别出查询中的实体能够帮助我们更好的理解用户的意图，从而提供更好的搜索结果。

同时，通用 NER 转向专用 NER 是一个趋势，解决这其中遇到的低资源、领域迁移问题将能够帮助命名实体识别技术更好在各个领域适配。

第二章 问题描述和挑战

2.1 问题描述和形式化定义

通过对命名实体发展的梳理，我们给出以下的定义，一个命名实体（NE）通常是一个专有名词或短语，它清楚地从一组具有类似属性的其他项中标识一个项。

命名实体识别顾名思义即根据预先定义的命名实体集，在文本中定位并分类命名实体的过程。下面我们给出命名实体识别任务的形式化定义。对于一些给定的文本$D = \{X_1, \dots, X_k\}$，NER 任务会有一个预先定义的实体类型集合$E = \{E_1, \dots, E_M\}$，其中E_i代表一种实体类型，例如人名（Person）。对于D中的一段文本输入$X_i = w_1, \dots, w_n$，NER 需要给出实体识别三元组输出(I_s, I_e, E_i)，其中$I_s \in [1, N], I_e \in [1, N]$代表实体词元在输入中的起始和终止位置，$E_i$代表其所属于的实体类型。

2.2 主要挑战

2.2.1 高标记成本和标记误差

标注数据带来的问题是所有序列标注任务，以及监督学习任务都必须要面对的。在现有表现良好的 NER 模型中，几乎所有的模型都需要大量的标注数据进行训练。但标注数据带来的是大量的人力成本和时间成本，这使得 NER 在实际应用的过程中具有很大的代价，在一些专业领域的应用过程中，这种代价尤为明显。而且 NER 任务通常是领域独立的，不同任务训练都是基于特定领域的数据集进行，当需要进行领域迁移时，标注数据将会成为这一过程中最昂贵的成本。

标注数据除了人力成本代价大之外，还有标注数据的一致性和标注的质量问题。一致性指的是同样的一个词可能在不同数据下被标注成不同的实体，比如“Baltimore”在数据集 MUC-7 中被标注为地名，而在 CoNLL03 中被标注为组织机构名。同时实体边界在不同数据中可能也是模糊的，比如“Empire State”

和“Empire State Building”在 CoNLL03 和 ACE 数据集中都是地名，但它们的实体边界是不同的。这样，在某个数据集下训练的模型更难复用到另一个数据集上了，即使他们是同一领域的数据。

除了上述问题之外，NER 任务的标注数据中还独有一个特殊的问题，那就是嵌套实体和多类型实体的存在。一个命名实体在某个场景下可能是由多个嵌套的实体组成，同时一个实体可能具有多个类别。这些都加大了 NER 任务的难度。

2.2.2 非正式文本

相对而言目前 NER 在部分充足数据的场景下表现良好，比如新闻或百科类数据。但当出现一些不是那么正式的文本时，比如网络上的用户评论等，特殊的语法结构和新的词汇组合会对 NER 任务带来十分大的冲击，模型将更加难以理解句子的语义。

除此之外，在许多任务场景下，会出现一些实体从未在训练集中出现过，这也是 NER 需要解决一个重要问题

第三章 评价指标、数据集和开源模型

3.1 评价指标

命名实体识别通常采用准确率（accuracy）、查准率(precision)、召回率(recall)以及 F1-score 作为模型的评价指标。下面对四个评价指标进行详细的说明。首先对公式中使用的变量定义如下：

TP：数据集中实际为命名实体，模型也预测为命名实体的数量。

FP：数据集中实际为非命名实体，模型却预测为命名实体的数量。

FN：数据集中实际为命名实体，模型却预测为非命名实体的数量。

TN：数据集中实际为非命名实体，模型也预测为非命名实体的数量。

然后我们分别给出查准率、召回率和 F1-Score 的公式定义。

查准率的计算公式定义如下：

$$Precision = \frac{TP}{TP + FP}$$

查准率衡量了模型中识别的命名实体有多少为真的命名实体。

召回率的计算公式定义如下：

$$Recall = \frac{TP}{TP + FN}$$

召回率衡量了数据集中的命名实体有多少被模型预测出来。

F1-Score 的计算公式如下：

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

F1-Score 为查准率和召回率的调和平均数，表示了查准率和召回率的一个综合的分数。任何一方的数值过低都会导致 F1-Score 的数值降低。

3.2 数据集

命名实体识别任务中常用的公开数据集信息如表 3-1 所示。

表 3-1 公开命名实体识别数据集信息

数据集	任务场景	数据集描述	应用范例	URL
Conll03[43]	非嵌套	Conll03 包含英文和德文两个语言的 NER 数据。英语数据取自路透社语料库，德文数据的文本来自 ECI 多语言文本语料库。包含 14,987 个训练句子，3,466 个验证句子，3,684 个测试句子，共四种实体类型。	[20]	链接
ACE2004[41]	嵌套和非嵌套	ACE 2004 多语言训练语料库包含了 2004 年自动内容提取（ACE）技术评估的全套英文、阿拉伯文和中文训练数据	[20]	链接
ACE2005[42]	嵌套和非嵌套	ACE 2005 多语言训练语料库包含了 2005 年自动内容提取（ACE）技术评估的全套英文、阿拉伯文和中文训练数据	[20]	链接
GENIA[39]	嵌套和非嵌套	GENIA 语料库是在 GENIA 项目范围内汇编和注释的生物医学文献的主要集合。创建该语料库是为了支持分子生物学领域的信息提取和文本挖掘系统的发展。	[20]	链接
OntoNotes5.0[44]	非嵌套	OntoNotes 5.0 版本是 OntoNotes 项目的最终版本。由三种语言（英语、汉语和阿拉伯语）的各种类型的文本（新闻、电话对话、网络日志、网络新闻组、广播、谈话节目）组成的大型语料。共 18 种实体类型	[20]	链接
OntoNotes4.0[45]	非嵌套		[20] [12]	链接
MSRA[46]	非嵌套	微软提供的中文命名实体识别数据集，包含三类实体	[20] [12]	链接
Weibo[47]	非嵌套	微博 NER 数据集是一个中文命名实体识别数据集，来自于社交媒体网站新浪微博。	[12]	链接
Resume[10]	非嵌套	简历实体识别数据集，包含八类实体的标注	[12]	链接
WNUT2017[48]	非嵌套	包含六大类实体类型	[37]	链接

3.3 开源 NER 模型分析比较

近几年在命名实体识别领域的开源模型中，有着许多优秀的工作。详细信息以及性能对比如表 3-2、3-3 所示。Bi-LSTM-CRF 作为早期的经典 NER 模型，在现在的工业界仍然有着广泛的应用，其性能虽然不如直接使用 BERT 等预训练模型，但是运行速度和存储空间都要远远优于 BERT 模型。Lattice 作为中文 NER 中的一个里程碑式工作，在 BERT 提出前，在中文 NER 领域有着绝对的统治力。但因为其结构复杂而且在大规模预料上无法并行训练，限制其在实际生产环境下的使用。Flat-Lattice 是它的一种升级版，更适配大规模预料高性能场景下的训练和使用。在嵌套实体的识别上 Biaffine-NER 和 BERT-MRC 不分伯仲，BERT-MRC 的缺点在于一次只能识别一种类型实体，所以运行速度上偏慢。

表 3-2 经典工作信息

方法	核心思想	数据集	F1-score	代码
Bi-LSTM-CRF[8]	使用双向长短期记忆网络作为上下文编码器，并在序列输出上应用 CRF	Conll03	88.83	链接
Lattice[10]	设计晶状体的 LSTM 融合字符信息和词典信息进行中文 NER 任务	MSRA	93.18	链接
		Weibo	58.79	
		Resume	94.46	
		OntoNotes4.0	73.88	
BERT[17]	基于 MLM 和 NSP 的预训练模型	Conll03	92.8	链接
BERT-MRC[18]	使用机器阅读理解框架解决 NER 任务，引入实体先验知识，同时解决嵌套实体识别和非嵌套实体识别任务	Conll03	93.04	链接
		OntoNotes5.0	91.11	
		MSRA	95.75	
		OntoNotes4.0	82.11	
		ACE2005	86.88	
		ACE2004	85.98	
		GENIA	83.75	
Biaffine-NER[19]	使用双仿射注意力机制使得实体的头尾词元信息能够进行交互	Conll03	93.5	链接
		OntoNotes5.0	91.3	
		ACE2005	85.4	
		ACE2004	86.7	
		GENIA	80.5	
Flat-Lattice[23]	将 Lattice 转换为平面结构	MSRA	96.09	链接
		Weibo	68.55	
		Resume	95.86	
		OntoNotes4.0	76.45	

CL-KL[37]	通过检索外部语境来提高 NER 模型准确性，并使用合作学习方法鼓励两个输入视图产生类似的上下文表征或输出标签分布	Conll03	93.85	链接
		WNUT 2017	60.45	
LUKE[38]	基于词和实体的语境的预训练 transformer 模型	Conll03	94.3	链接

表 3-3 不同数据集上各类方法比较

数据集	方法	F1-score
Conll03	Bi-LSTM-CRF	88.83
	BERT	92.8
	BERT-MRC	93.04
	Biaffine-NER	93.5
	CL-KL	93.85
	LUKE	94.3
OntoNotes5.0	BERT-MRC	91.11
	Biaffine-NER	91.3
OntoNotes4.0	Lattice	73.88
	Flat-Lattice	76.45
	BERT-MRC	82.11
MSRA	Lattice	93.18
	Flat-Lattice	96.09
	BERT-MRC	95.75
ACE2004	BERT-MRC	85.98
	Biaffine-NER	86.7
ACE2005	BERT-MRC	86.88
	Biaffine-NER	85.4
GENIA	BERT-MRC	83.75
	Biaffine-NER	80.5

第四章 传统命名实体识别方法

纵观从 NER 提出至现今的所有研究方法和模型，我们可以将其按照研究方法的差别大体分为三个类别：1. 基于规则的方法；2. 基于特征的监督学习方法；3. 基于深度学习的方法。其中 1，2 为传统研究方法，而 3 为近年来的主流研究方法。自深度学习方法在 NER 应用以来，NER 的研究迎来巨大变革，不断刷新各大数据上的最优结果。

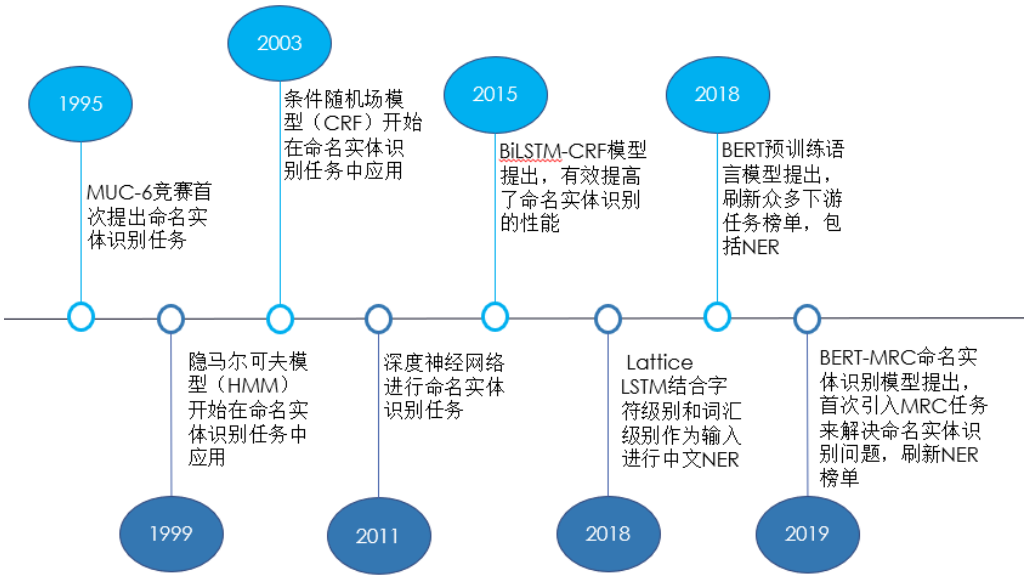

图 4-1 命名实体识别技术发展史

基于规则的 NER 系统主要依赖人工制定的规则来工作。规则的制定通常基于领域专有的词典以及句法词汇的模式匹配。

比较典型的方法研究如下：

1. Kim 等人[1]提出对文本输入使用规则推理方法，该系统能自动的生成基于词性标记的规则对实体进行匹配

2. 在生物医学领域，Hanisch 等人[2]提出 ProMiner 系统，该系统利用一个预处理的同义词词典，确定生物医学文本中的蛋白质和潜在的基因实体。

基于规则的命名实体识别方法主要是基于手工制作的语义或句法规则来进行实体识别。基于规则的 NER 系统普遍都能表现良好，因为他们的规则指定过

程通常是详尽的并且贴合任务，为人为制定的标准。但由于规则的制定过程费事费力，并且在进行不同领域的迁移时，基于规则的方法无法复用，所以该类别的方法难以大规模的流行。基于规则的方法在专业领域的命名实体识别过程中，会出现高准确率和低召回率的现象，这是因为在规则和字典中的实体通常能很好的被识别，但不在规则中的专业领域实体通常直接被判断为非实体，所以出现召回率的降低现象。

在监督学习的方法中，NER 通常被当作多分类或者序列标注任务。在基于特征的监督学习方法中涌现了很多优秀的机器学习算法的应用。之所以称之为基于特征的方法，是因为在该类算法中，输入文本通常被表示为一定的特征向量表示，如词级别的特征向量表示，或文档的特征向量表示。基于这些特征表示，许多传统的机器学习算法能在监督学习后表现良好。

比较著名的方法如下：

1. Bikel 等人[3]首次提出基于隐马尔可夫模型（HMM）的 NER 系统。
2. Borthwick 等人[4]应用最大熵原理来解决实体识别问题。
3. McNamee 和 Mayfield[5]使用 SVM 来进行实体识别。
4. SVM 虽然效果良好，但在预测实体标签时不考虑“相邻”标签的信息，而条件随机场（CRF）在计算时充分的考虑了上下文的信息。McCallum 等人[6]首次提出在 NER 中使用 CRF。此后 CRF 因为其在序列标注任务中的有效性成为了主流的序列标注算法，并且在深度学习方法提出后被广泛使用。

第五章 基于深度学习的命名实体识别方法

近年来，基于深度学习的方法出现了井喷式的发展，主要的一个原因是其在实体识别效果上比传统算法表现更为优异。和基于特征的监督学习方法相比，深度学习方法能自动的发现文本的潜在语义特征表示。深度学习方法的关键优势在于表征学习的能力，以及由向量表征和神经网络处理共同赋予的语义组合能力。这使得模型可以通过原始数据，自动学习文本的语义潜在表示。

基于深度学习的 NER 方法具有以下几个重要的优点：

1. 得益于深度网络激活函数的非线性变换，能够产生输入到输出的非线性映射，而传统机器学习方法如 HMM 和 CRF 只能实现线性的变换过程。

2.基于深度学习方法可以大大节省针对 NER 特性进行设计的工作，传统的基于特征的监督学习方法需要大量的工程技术和领域专业知识。另外，深度学习方法能够有效的从原始数据中自动学习有用的潜在知识。

3.深度学习的方法可以通过梯度下降等优化算法实现端到端的训练，这个特性使研究者能专注于设计更加复杂有效的 NER 系统。

对于众多的深度学习命名实体识别方法，如何进行分类是一个重要的问题。在我看来结合实际使用场景或数据特征下的 NER 目标，所有的 NER 方法可以分为 Flat NER 和 Nest NER 两类。这种分类对应了实际命名实体识别场景下的两种情况，即非嵌套的命名实体识别和嵌套的命名实体识别。基于实体的嵌套数据情况，也有与之对应的两类方法，即基于序列标注的命名实体识别方法和基于 span(实体块)预测的命名实体识别方法。基于序列标注的命名实体识别方法非常适合解决非嵌套实体的识别，而对于嵌套实体的识别就显的无能为力。而基于 span 预测的命名实体识别方法则能同时应对嵌套实体和非嵌套实体两种情况。

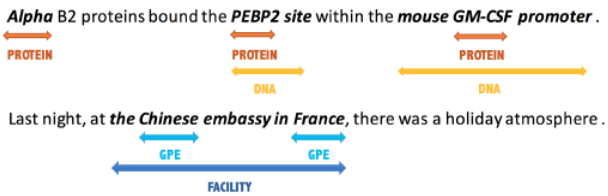

图 5-1 嵌套实体示例[18]

5.1 基于序列标注的命名实体识别方法

基于序列标注的命名实体识别方法将命名实体识别问题视为序列标注任务，对于序列输入中的每个词元（token）给出对应的标记输出。在序列标注方法中，数据集通常以 BIO 或 BIOES 形式进行标注。对于属于实体的词元给出对应的起始终止标记，对于非实体词元给出“O”标记。绝大部分的命名实体识别方法都是基于序列标注的，常见的设计方法为将深度神经网络和条件随机场结合使用。

早期的序列标注任务中，设计的模型常常能够同时解决 POS,NER,Chunk 等任务。常见的输入序列编码结构有 CNN, LSTM, Transformer 三种。Collobert 等人[7]提出了基于 CNN 的序列标注模型，结构如图 5-2 所示。该模型首次提出使用 CNN 来解决序列标注任务。而后有很多工作使用 CNN 进行序列标注任务，例如[32][33]。

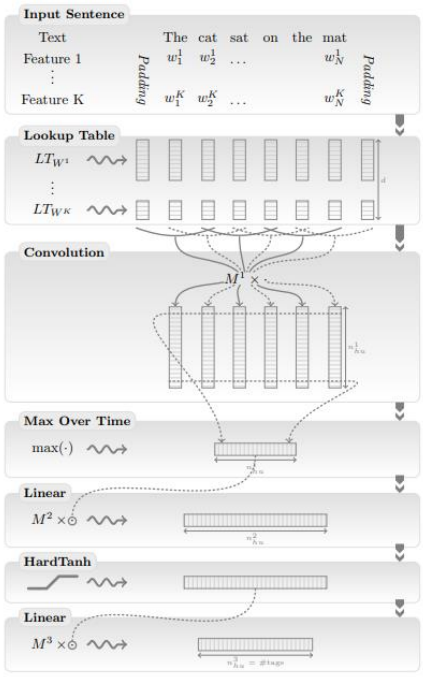

图 5-2　Collobert 工作结构图[7]

Huang 等人[8]首次在 2015 年提出了经典的 BiLSTM-CRF 的序列标注模型，利用 BiLSTM 对输入文本进行上下文编码，使用 CRF 进行输出序列预测。其模型结构如图 5-3 所示。在此之后，许多的 NER 模型都采用 BiLSTM 作为输入序列的编码器，如[15][30][31]。Ma 等人[16]在 2016 年提出 LSTM-CNN-CRF 模型，

使用 CNN 对输入英文字符进行编码，解决未登录词的问题，然后使用 LSTM 作为序列编码器。

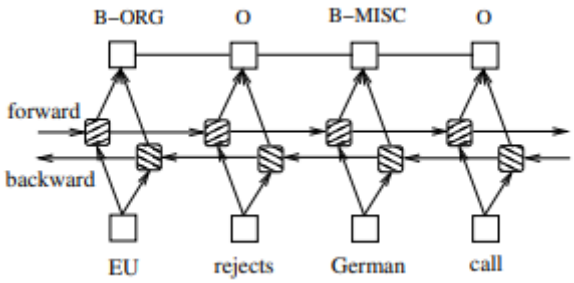

图 5-3　BiLSTM-CRF 模型[8]

Devlin 等人于 2018 年提出 BERT 模型[17]，基于 Transformer[9]编码器对输入序列进行编码，能够非常方便高效的迁移到下游任务，包括 NER。而后许多方法都以 BERT 作为 NER 的编码器，例如[18]和[19]。

5.2 基于块预测的命名实体识别方法

嵌套实体识别问题在 2003 年首次被 Kim 等人[39]提出，传统的单一序列标注模型无法进行嵌套实体识别，只能通过多个模型或输出层的叠加才能完成该任务。为了解决非嵌套实体识别，近年来，许多基于 span 预测的命名实体识别方法被提出。基于 span 的命名实体识别方法的优势在于能够同时解决嵌套实体和非嵌套实体的识别。其中代表性的工作有 BERT-MRC 和 Biaffine-NER。

Li 等人[18]提出的 BERT-MRC 模型基于机器阅读理解框架来解决命名实体识别任务。输入序列包含实体查询文本和待查询文本。其核心思想再于将实体识别拆分为单个类型的识别，并通过实体查询文本引入实体类型的语义信息。BERT-MRC 模型的数据集一个实体的标注为一个三元组$(q_y, x_{start,end}, X)$，即(QUERY, ANSWER, CONTEXT)，表示输入文本X中$x_{start,end}$是一个实体，且实体类型为q_y查询的类型。模型主体基于 BERT，输入形式为$\{[CLS], q_1, \ldots, q_m, [SEP], x_1, \ldots, x_n\}$。模型输出包含实体 Start 和 End 的预测。损失函数由 Start、End 和 span 三部分的交叉熵组成。其 attention 可视化的实验结果表明实体 Query 的引入能够在输入序列的编码上引入实体先验知识，如图 5-4 所示。

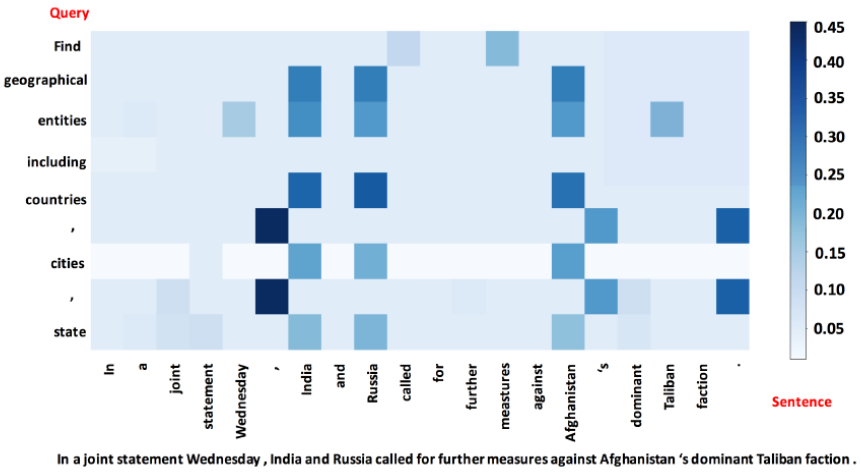

图 5-1　BERT-MRC 注意力可视化[18]

同年，Yu 等人[19]提出 Biaffine-NER 模型架构来进行实体块预测，如图 5-4 所示。其和 BERT-MRC 模型主要区别在于输出层使用了双仿射注意力机制。Biaffine-NER 模型可以同时识别句子中所有类型的实体。

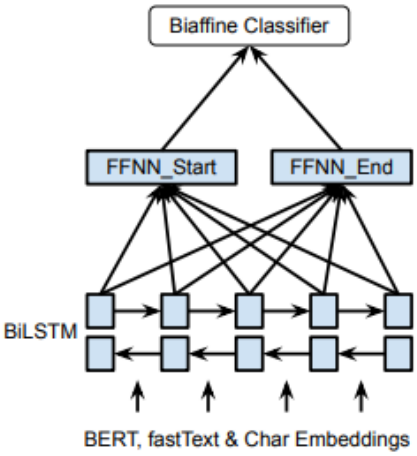

图 5-2　Biaffine-NER 模型[19]

5.3 针对低资源场景的命名实体识别方法

低资源的 NER 一直是一个重要且待解决的问题。NER 作为一个对标记数据和质量依赖很高的任务，在一些低资源环境下，大部分监督模型性能都会急剧下降。近几年的 NLP 顶级会议，越来越多的工作开始关注低资源或 ZERO-SHOT 情况下的命名实体识别问题。

Kruengkrai 等人[20]在结合句子级别标注和词元级别标注的角度上进行了尝试，

实验结果表明两种标注的联合训练能够有效改善低资源环境下 NER 模型的性能。该模型共享句子级别任务和词元级别任务的编码网络，然后进行联合训练。其模型结构如图 5-6 所示。

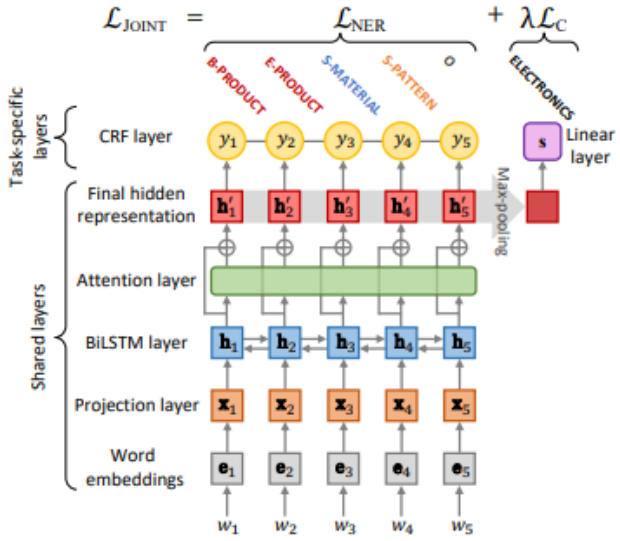

图 5-3　引入句子标注的低资源场景 NER 模型[20]

Gazetteers 被证明是一个很好的 NER 辅助知识，但是在低资源领域，很难找到合适的 Gazetteers,因此 Rijhwani 等人[21]提出引入“soft gazetteers”来改善低资源环境下的 NER 性能。该方法通过跨语言的实体链接（例如维基百科），为低资源的语种提供 soft gazetteer 信息。其模型结构如图 5-7 所示。

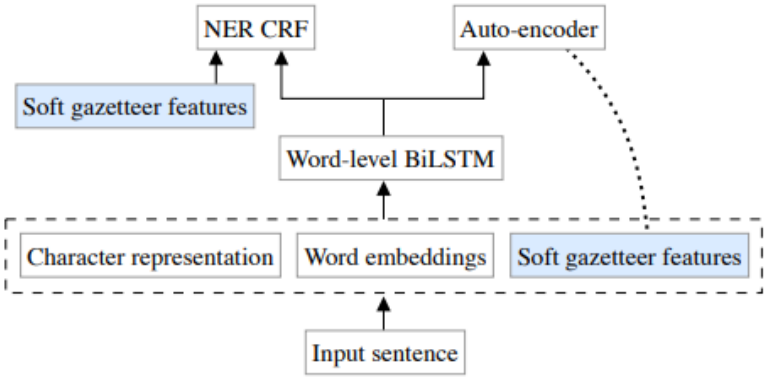

图 5-4　soft gazetteer 模型[21]

5.4 中文命名实体识别方法

中文 NER 相较于英文 NER 因为没有天然的词级别结构一直是一个具有挑战性的任务，所以相同结构的 NER 模型在英文任务下通常表现的比中文更加好。中文 NER 方法中，一个重点就是如何合理的引入外部词级的信息，例如词典或 gazetteers。

Zhang 等人[10]于 2018 年提出经典的 Lattice LSTM 结构用于解决中文 NER 问题。作者通过 Lattice LSTM 结构引入词典信息，并让模型在自动学习合适的词级别信息，而非人工干预。其模型结构如图 5-8 所示。

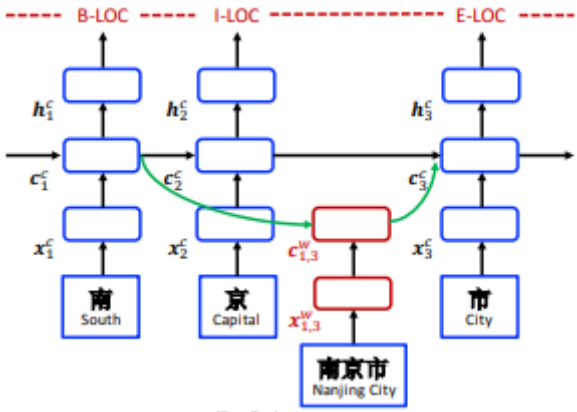

图 5-5 Lattice LSTM[10]

与 Lattice LSTM 不同的是，Ding 等人[22]通过一个独立的图神经网络来学习并引入 gazetteers 的知识。由于 Lattice LSTM 结构过于复杂且无法并行，Li 等人[23]提出 Flat-lattice Transformer 模型，将 Lattice 转换为平面结构来将词级别信息直接融入 Transformer 的训练中。其模型结构如图 5-9 所示。

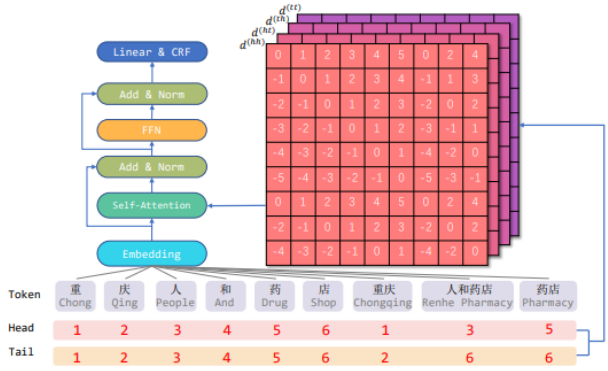

图 5-6 Flat-Lattice[23]

5.5 基于数据层面增强的命名实体识别方法

NER 任务中有许多能够利用的外部知识能够帮助提升 NER 模型的性能。最被广泛使用的即 gazetteers。无论在低资源环境（[21]）、中文 NER 任务上（[10]，[22]，[23]）或是其它场景下（[34][35][36]），gazetteers 都是优质的 NER 外部知识。另一种有效提升NER模型性能的方法为使用句子级别标注，例如[20]和[24]。

通常情况下，NER 训练数据都是脱离上下文语境的，Wang 等人[37]提出通过搜索引擎来寻找句子的外部语境，检索一组和原始句子语义相关的文本，如图 5-10 所示。并通过合作学习的方式使得两个不同的输入视图产生类似的上下文表征或输出标签分布，以此降低 NER 的使用代价。

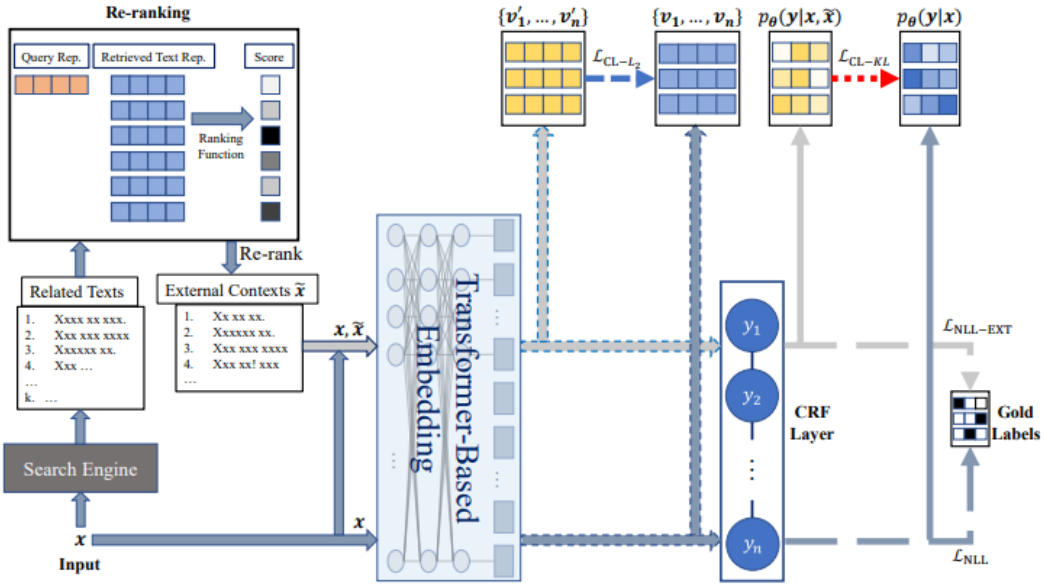

图 5-7 检索 NER 输入的上下文语境[37]

除了上述和实体相关的额外数据的引入，还有一类方法通过增加标记数据的数量来提升 NER 的性能。这些方法通过一定的方式筛选额外的弱标记数据，保证弱标记数据的质量。例如，jiang 等人[40]提出一个多阶段计算框架 NEEDLE，如图 11 所示，通过强标记数据和由知识库得到的弱标记数据进行联合学习来提升 NER 的性能，该框架有效减少了弱标记数据带来的噪声。

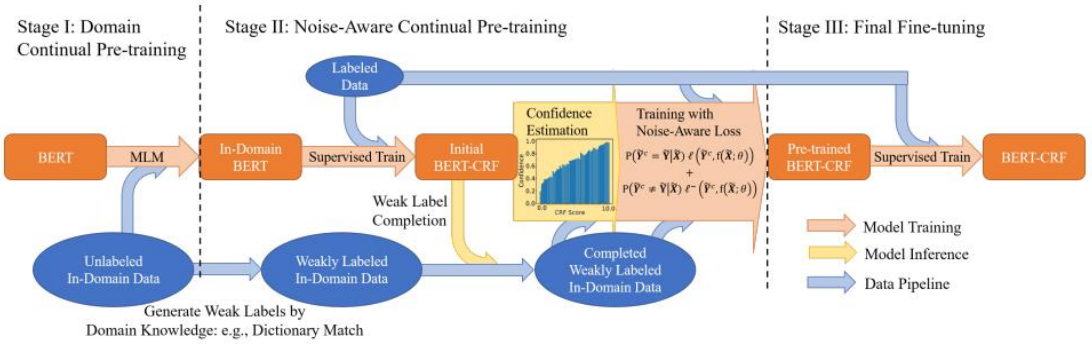

图 5-8　NEEDLE 框架[40]

5.6 基于预训练角度增强的命名实体识别方法

Wikipedia 中的大量锚文本被证明是一种很好的外部知识，Xue 等人[25]使用 Wikipedia 中大量的锚文本对 BERT 进行实体相关的预训练，有效增强了模型的性能。中文任务上，Meng 等人[27]证明使用字形信息对 BERT 进行预训练，能够有效增强该模型在下游中文 nlp 任务上的性能。

Yamada 等人[38]提出的 LUKE 模型基于词和实体的语境的进行预训练，将给定文本中的词和实体视为独立的标记，并输出它们的上下文表示，有效的增强了基于 BERT 的 NER 模型的性能。LUKE 模型的结构如图 5-12 所示。

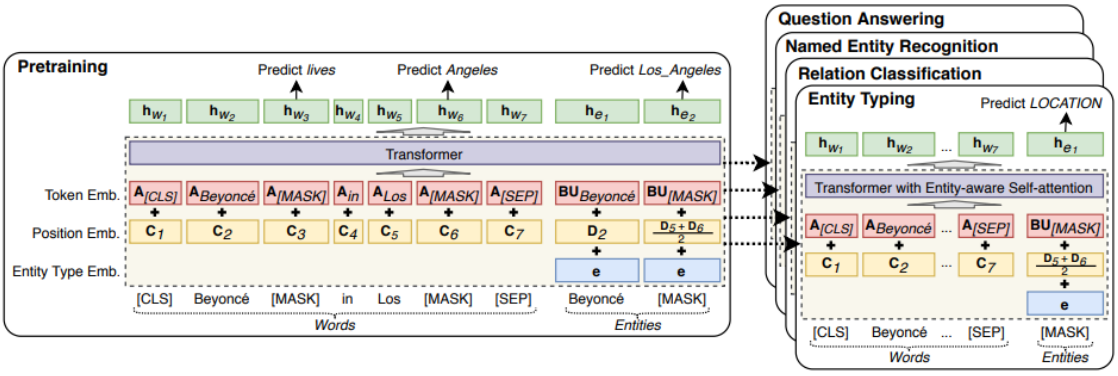

图 5-9　LUKE 模型[38]

5.7 基于迁移学习或跨域学习的命名实体识别方法

迁移学习的目的是将源领域的知识迁移到目标领域，使得目标领域的模型训练表现良好。对于 NER 这类目前需要大量标注数据以表现良好的任务来说，迁

移学习带来的帮助十分重要，特别是面向专业领域的 NER，常常因为标注数据匮乏，而表现不好。而迁移学习带来的领域知识迁移能大大改善这样的现象。

Pan 等人[11]提出了一种跨域 NER 的转移联合嵌入(TJE)方法。TJE 采用标签嵌入将多分类问题转化为低维潜在空间回归的问题。实验结果证明了 TJE 在 ACE 2005 数据集上具有跨领域的有效性。

Lin 等人[12]提出了一种 fine-tuning NER 的方法，如图 5-13 所示，引入了三个神经网络适配层:单词适配层，句子适配层，输出适配层。

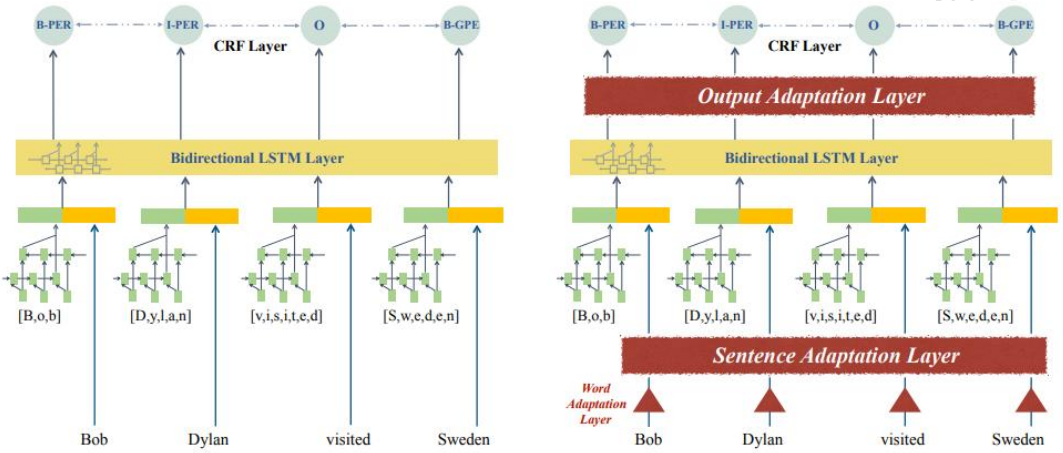

图 5-10　迁模型结构[12]

Jia 等人[29]提出跨域 NER 模型，如图 5-14 所示。通过跨域的 LM 和 NER 共同学习，使得在新领域不需要 NER 标记数据也能具有一定的标注能力。

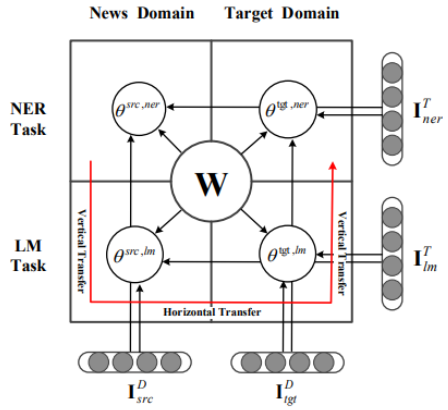

图 5-11　Cross-domain NER 模型[29]

5.8 基于持续学习的命名实体识别方法

NER 模型在实际生产环境下常常会遇到需要不断增加新实体类型的情况，所以如何有效的学习新的实体并且不遗忘旧的实体知识就显的格外重要。Monaikul 等人[26]提出使用知识蒸馏框架来进行持续学习，如图 5-15 所示。通过教师网络将旧实体知识交给新的学生网络，使得新网络在学习到新的实体类型知识的同时步遗忘旧实体的信息，并且无需旧实体类型的标记数据。

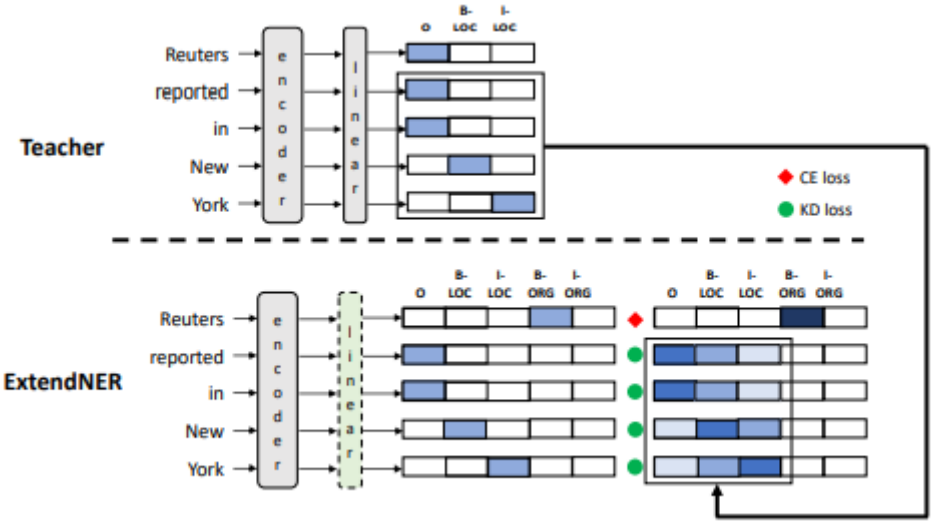

图 5-12 持续学习模型[26]

5.9 NER 中的针对性 loss

NER 问题都存在数据不平衡的问题，比如，对于命名实体识别任务，采用 BIO 标注，如果把 O 视为负例，其它为正例。交叉熵“平等”地看待每一个样本，无论正负，都尽力把它们推向 1（正例）或 0（负例）。但实际上，对分类而言，将一个样本分类为负只需要它的概率小于 0.5 即可，完全没有必要将它推向 0。基于 Dice loss，Li 等人[28]提出 DSC，训练时推动模型更加关注困难的样本，降低简单负例的学习程度。

5.10 基于对抗学习的命名实体识别方法

对抗学习的目的是在模型的训练过程中引入对抗样本，使得模型具有更好的鲁棒性。DATNet[13]是最近在 NER 上比较出色的对抗学习工作。该模型旨在使用对抗学习和迁移学习来解决少量数据的 NER 问题。该作者通过引入对抗样本，来缓解模型的过拟合问题。

第六章 前沿问题和未来工作

6.1 专业领域更加细粒度的 NER

目前更多的 NER 研究工作专注于通用领域 NER 的效果提升，榜单也在不断刷新，但通用领域 NER 带来的微小的效果提升对于整个 NER 问题的解决其实帮助不大。未来的 NER 应用场景必然是更加领域化，每个领域的 NER 识别需求都大相径庭，而现有的专注于领域细粒度的 NER 研究还很少。

6.2 实体边界的检测

现有的 NER 任务都将实体边界的检测和实体类型的检测和为一个任务进行。但我们会发现实体边界的检测效果的提升会大大提高整个命名实体识别任务的效果，实体类型的检测相对比较简单。现有的 NER 任务通常使用 BIOES 等标记数据策略，将边界和类别和为一体，使得实体边界的检测方面没有争对性的策略。我们知道实体边界是可以领域独立的，但实力类型不行，我相信未来如果能在实体边界检测上取得一定的针对性进展，那 NER 任务必然也会迎来曙光。

6.3 多任务学习和迁移学习

尽管有许多的工作在多任务学习和迁移学习的上进行了尝试，但我觉的目前来看，多任务学习和迁移学习在 NER 上的应用还远远不够，还有很多值得探索的空间。多任务学习和迁移学习对于 NER 来说具有得天独厚的优势，特别是在专业领域的 NER 任务上。未来设计一个跨领域的多任务以及迁移学习模型将是一个重要的突破口。

6.4 无监督学习

无监督学习不仅在 NER，乃至整个机器学习领域，我认为都是未来研究的

一个重要方向，如何更加方便有效的领域互联网上海量的原始数据，将会给整个机器学习领域带来革命性的改变。

参考文献

[1]. Kim, Ji-Hwan, and Philip C. Woodland. “A Rule-Based Named Entity Recognition System for Speech Input.” INTERSPEECH, 2000, pp. 528–531.

[2]. Hanisch, Daniel, et al. “ProMiner: Rule-Based Protein and Gene Entity Recognition.” BMC Bioinformatics, vol. 6, no. 1, 2005, pp. 1–9.

[3]. Bikel, Daniel M., et al. “Nymble: A High-Performance Learning Name-Finder.” Proceedings of the Fifth Conference on Applied Natural Language Processing, 1997, pp. 194–201.

[4]. Borthwick, Andrew, et al. “NYU: Description of the MENE Named Entity System as Used in MUC-7.” Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998, 1998.

[5]. McNamee, Paul, and James Mayfield. “Entity Extraction without Language-Specific Resources.” COLING-02 Proceedings of the 6th Conference on Natural Language Learning - Volume 20, 2002, pp. 1–4.

[6]. McCallum, Andrew, and Wei Li. “Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons.” CONLL ’03 Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, 2003, pp. 188–191.

[7]. Collobert, Ronan, et al. “Natural Language Processing (Almost) from Scratch.” Journal of Machine Learning Research, vol. 12, no. 76, 2011, pp. 2493–2537.

[8]. Huang, Zhiheng, et al. “Bidirectional LSTM-CRF Models for Sequence Tagging.” ArXiv Preprint ArXiv:1508.01991, 2015.

[9]. Vaswani, Ashish, et al. “Attention Is All You Need.” Proceedings of the 31st International Conference on Neural Information Processing Systems, vol. 30, 2017, pp. 5998–6008.

[10]. Zhang, Yue, and Jie Yang. “Chinese NER Using Lattice LSTM.” Proceedings of

the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, 2018, pp. 1554–1564.

[11]. Pan, Sinno Jialin, et al. “Transfer Joint Embedding for Cross-Domain Named Entity Recognition.” ACM Transactions on Information Systems, vol. 31, no. 2, 2013, p. 7.

[12]. Lin, Bill Yuchen, and Wei Lu. “Neural Adaptation Layers for Cross-Domain Named Entity Recognition.” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2012–2022.

[13]. Zhou, Joey Tianyi, et al. DATNet: Dual Adversarial Transfer for Low-Resource Named Entity Recognition. 2018.

[14]. Lample, Guillaume, et al. “Neural Architectures for Named Entity Recognition.” 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference, 2016, pp. 260–270.

[15]. Chiu, Jason P. C., and Eric Nichols. “Named Entity Recognition with Bidirectional LSTM-CNNs.” Transactions of the Association for Computational Linguistics, vol. 4, no. 1, 2016, pp. 357–370.

[16]. Ma, Xuezhe, and Eduard H. Hovy. “End-to-End Sequence Labeling via Bi-Directional LSTM-CNNs-CRF.” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, 2016, pp. 1064–1074.

[17]. Devlin, Jacob, et al. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2018, pp. 4171–4186.

[18]. Li, Xiaoya, et al. “A Unified MRC Framework for Named Entity Recognition.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5849–5859.

[19]. Yu, Juntao, et al. “Named Entity Recognition as Dependency Parsing.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 6470–6476.

[20]. Kruengkrai, Canasai, et al. “Improving Low-Resource Named Entity Recognition Using Joint Sentence and Token Labeling.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5898–5905.

[21]. Rijhwani, Shruti, et al. “Soft Gazetteers for Low-Resource Named Entity Recognition.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8118–8123.

[22]. Ding, Ruixue, et al. “A Neural Multi-Digraph Model for Chinese NER with Gazetteers.” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1462–1467.

[23]. Li, Xiaonan, et al. “FLAT: Chinese NER Using Flat-Lattice Transformer.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 6836–6842.

[24]. Patra, Barun, and Joel Ruben Antony Moniz. “Weakly Supervised Attention Networks for Entity Recognition.” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6267–6272.

[25]. Mengge, Xue, et al. “Coarse-to-Fine Pre-Training for Named Entity Recognition.” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6345–6354.

[26]. Monaikul, Natawut, et al. “Continual Learning for Named Entity Recognition.” AAAI, 2021, pp. 13570–13577.

[27]. Meng, Yuxian, et al. “Glyce: Glyph-Vectors for Chinese Character Representations.” Advances in Neural Information Processing Systems, vol. 32, 2019, pp. 2742–2753.

[28]. Li, Xiaoya, et al. “Dice Loss for Data-Imbalanced NLP Tasks.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 465–476.

[29]. Jia, Chen, et al. “Cross-Domain NER Using Cross-Domain Language Modeling.” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2464–2474.

[30]. Nguyen, Thien Huu, et al. “Toward Mention Detection Robustness with Recurrent Neural Networks.” ArXiv Preprint ArXiv:1602.07749, 2016.

[31]. Zheng, Suncong, et al. “Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme.” Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, 2017, pp. 1227–1236.

[32]. Yao, Lin, et al. “Biomedical Named Entity Recognition Based on Deep Neutral Network.” International Journal of Hybrid Information Technology, vol. 8, no. 8, 2015, pp. 279–288.

[33]. Wu, Yonghui, et al. “Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network.” Studies in Health Technology and Informatics, vol. 216, 2015, pp. 624–628.

[34]. Agarwal, Oshin, and Ani Nenkova. “The Utility and Interplay of Gazetteers and Entity Segmentation for Named Entity Recognition in English.” ACL 2021: 59th Annual Meeting of the Association for Computational Linguistics, 2021, pp. 3990–4002.

[35]. Liu, Tianyu, et al. “Towards Improving Neural Named Entity Recognition with Gazetteers.” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5301–5307.

[36]. Lin, Hongyu, et al. “Gazetteer-Enhanced Attentive Neural Networks for Named Entity Recognition.” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on

Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6231–6236.

[37]. Wang, Xinyu, et al. “Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning.” ACL 2021: 59th Annual Meeting of the Association for Computational Linguistics, 2021, pp. 1800–1812.

[38]. Yamada, Ikuya, et al. “LUKE: Deep Contextualized Entity Representations with Entity-Aware Self-Attention.” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6442–6454.

[39]. Kim, Jin-Dong, et al. “GENIA Corpus—a Semantically Annotated Corpus for Bio-Textmining.” Bioinformatics, vol. 19, 2003, pp. 180–182.

[40]. Jiang, Haoming, et al. “Named Entity Recognition with Small Strongly Labeled and Large Weakly Labeled Data.” ACL 2021: 59th Annual Meeting of the Association for Computational Linguistics, 2021, pp. 1775–1789.

[41]. Doddington, George R., et al. “The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation.” Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004), 2004.

[42]. Walker, Christopher, et al. ACE 2005 Multilingual Training Corpus LDC2006T06. Web Download. Philadelphia: Linguistic Data Consortium, 2006.

[43]. Sang, Erik F.Tjong Kim, and Fien De Meulder. “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.” CONLL ’03 Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, 2003, pp. 142–147.

[44]. Pradhan, Sameer, et al. “Towards Robust Linguistic Analysis Using OntoNotes.” Proceedings of the Seventeenth Conference on Computational Natural Language Learning, 2013, pp. 143–152.

[45]. Pradhan, Sameer. Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. 2011.

[46]. Levow, Gina-Anne. “The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition.” Proceedings of the

Fifth SIGHAN Workshop on Chinese Language Processing, 2006, pp. 108–117.

[47]. Peng, Nanyun, and Mark Dredze. “Named Entity Recognition for Chinese Social Media with Jointly Trained Embeddings.” Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 548–554.

[48]. Derczynski, Leon, et al. “Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition.” Proceedings of the 3rd Workshop on Noisy User-Generated Text, 2017, pp. 140–147.

[49]. Grishman, Ralph, and Beth Sundheim. “Message Understanding Conference-6: A Brief History.” COLING ’96 Proceedings of the 16th Conference on Computational Linguistics - Volume 1, vol. 1, 1996, pp. 466–471.

[50]. Kripke, Saul Aaron. Naming and Necessity. Semantics of natural language. Springer, 1972, pp. 253–355.