doi:10.3772/j.issn.2095-915x.2020.04.004

开放科学（资源服务）标识码（OSID）

# 面向少量标注数据的命名实体识别研究

石教祥 朱礼军 望俊成 王政 魏超

中国科学技术信息研究所 北京 100038

**摘要：**作为语义知识库、知识图谱的基本组件，命名实体识别对智能系统建设和科技情报服务都起到重要作用。近年来，深度学习方法在特征抽取深度和模型精度上表现优异，已经超过了传统方法，但无论是传统机器学习还是深度学习方法都依赖大量标注数据来训练模型，而现有的研究对少量标注数据学习问题探讨较少。鉴于此，本文全面总结了少量标注数据命名实体识别方法。具体地，按照数据、模型、特征、知识的学习逻辑区分为 4 类：基于数据增强、模型迁移、特征变换、知识链接的方法，并对这些方法进行分析和比较。此外，我们整合了数据资源以及典型方法评测，最后对未来可能的发展方向进行预测。

**关键词：**命名实体识别；深度学习；迁移学习；科技情报

**中图分类号：**G35

# Research on Named Entity Recognition from Sparsely Labeled Data

SHI Jiaoxiang ZHU Lijun WANG Juncheng WANG Zheng WEI Chao

Institute of Scientific and Technical Information of China, Beijing 100038, China

**Abstract:** As a basic component of the semantic knowledge base and knowledge graph, named entity recognition plays an important role in intelligent system construction and science & technology intelligence services. In recent years, the deep learning method that excels in feature extraction depth and model accuracy has surpassed the traditional method, but both traditional

**基金项目：**中国博士后科学基金第 65 批面上项目“流形正则化自编码政策文本表示及主题词抽取方法研究”（2019M650804）。

**作者简介：**石教祥（1995-），硕士研究生，研究方向：知识工程与知识发现；朱礼军（1973-），博士，研究员，研究方向：Semantic Web，Web Service 和知识技术在科技信息服务中的应用；望俊成（1984-），博士，副研究员，研究方向：科技政策与科技管理、文本数据可视化、大数据治理；王政（1992-），硕士，助理研究员，研究方向：知识图谱、主题模型、智能问答；魏超（1985-），博士，助理研究员，研究方向：文本表示、知识图谱、智能问答，E-mail: weichaolx@gmail.com。

doi:10.3772/j.issn.2095-915x.2020.04.004

machine learning and neural network methods rely on a large amount of labeled resources to train the model. Since the existing research rarely discusses the sparsely labeled data problem, this paper comprehensively summarizes the sparsely labeled data named entity recognition methods. Specifically, according to the learning logic of data, model, feature, knowledge, we divide these methods into four categories: the methods based on data augmentation, model migration, feature transformation, knowledge linkage. And then, we analyze and compare these categories. In addition, we integrated data resources and evaluation of typical methods, and finally predict the possible future development direction.
**Keywords:** Named entity recognition; deep learning; transfer learning; science & technology Intelligence

## 引言

新时期，情报工程化是创新科技情报工作的有效途径，如何从海量无结构数据中抽取出语义信息成为情报工程化亟待解决的问题[1]。其中，命名实体识别（Named Entity Recognition，NER）是解决这一问题的基本手段，特别是近些年来，以知识驱动的人工智能应用不断发展，带来了对精炼和专业知识的更多需求，领域 NER 成为文本挖掘技术的重要发展趋势，面向特定领域的 NER 也对下游知识库和知识图谱构建起到举足轻重的作用。

与大多数自然语言处理（Natural Language Processing，NLP）任务一样，命名实体的准确识别依赖足量标注样本，当有大量标注数据集可用时，NER 任务可以得到高质量地解决。但在现实世界里，标准数据集非常稀缺，例如在生物医药等专业领域，往往缺乏可直接用于模型训练的数据集[2]，特定领域由于专业性强，数据标注依赖领域专家，这种劳动密集的缺陷制约了 NER 的快速发展。因此，对少量标注数据情况下的 NER 进行深入研究极为必要。

目前，面向少量标注数据的 NER 研究还处于发展阶段，相关工作主要集中于传统机器学习、深度学习方法。近些年，一些研究对少量标注数据 NER 进行了探索，但是缺乏系统性归纳和总结。鉴于此，本文基于文献调研和统计分析方法，全面综述了相关文献并进行相关分析，以期为学者深入研究提供参考。本文按时间跨度 1995-2019 年，以 Web of Science(Wos) 核心合集、万方数据库为数据源，制定检索式分别为：“TS=(“named entit* recog*” or ner) AND TS=(“transf* learning” or tl or “few* shot” or “zero* shot” or “small sample*” or “small set*” or “small data*” or “few sample*” or “few set*” or “few data*” or “little sample*” or “little set*” or “little data*” or “unannotated data*” or “spars* labeled data*” or “semi-supervised” or “distant supervised” or unsupervised or “weakly supervise”)”、“题名或关键词 :( 命名实体识别、NER、少量标注数据、少样本、零样本 )”，共遴选出 124 篇文献，本文的研究内容、分析比较主要据此展开。

## 1 研究内容

### 1.1 研究领域、语种统计分析

NER 指识别出文中具有唯一标识的专有名词，一个命名实体通常是在给定数据集中具有相似属性的词语或概念[3]。如图 1 所示，自从

1995 年 MUC-6 会议首次提出以来，NER 的相关文献逐渐上升，特别是 2013 年以来，随着深度学习的兴起，NER 等 NLP 任务在更多领域得到广泛应用。

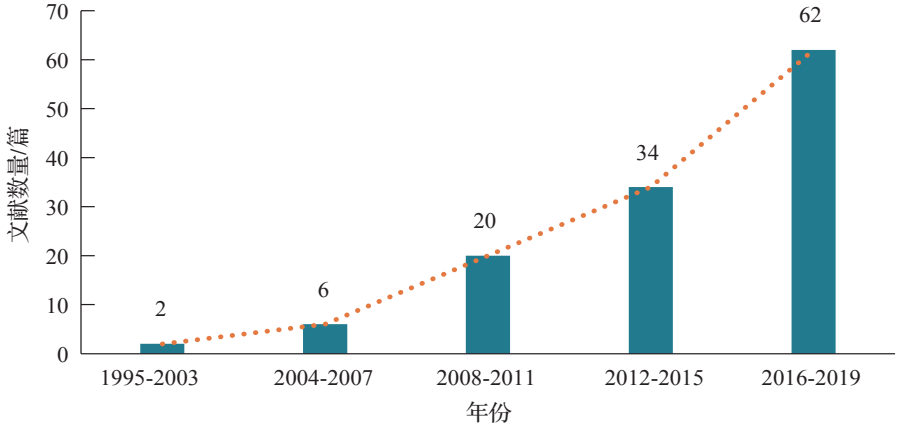


图 **1** 发文数量随时间变化

如图 2，从标注数据来看，由于通用领域标注资源丰富，机构、人名和时间等实体结构相对简单，在少量标注数据实验中性能更为优异而成为人们研究的主要对象。但不容忽视的是，在生物医药等专业领域，实体资源稀缺也更为关键，相关的研究占比也很大。

如图 3，在 NER 任务中，英语是研究的主体，其次是汉语，而窄域语种资源相对匮乏，但语种之间互相借鉴对目标语种的 NER 具有启发性意义。例如东亚 - 太平洋语系较为相近，而印欧语系可互相参考。

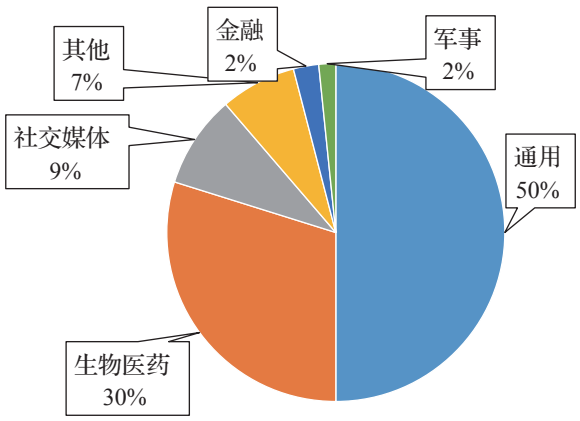


图 **2** **NER** 任务领域占比

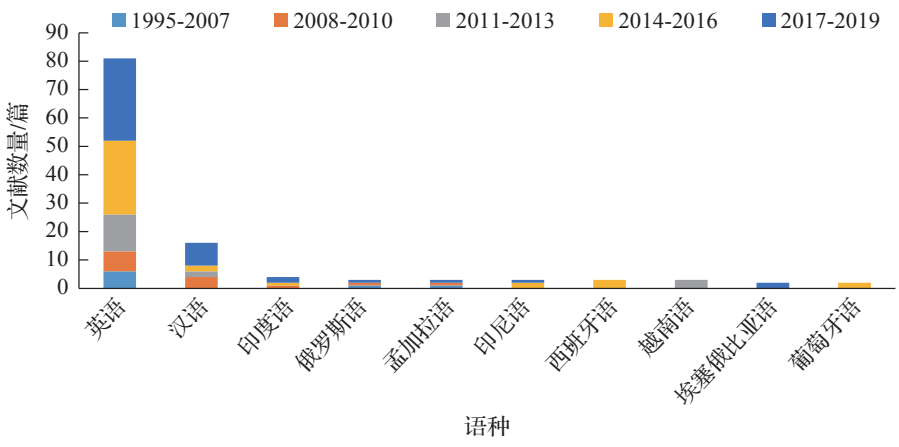


图 **3** **NER** 任务数量排名前 **10** 语言

doi:10.3772/j.issn.2095-915x.2020.04.004

### 1.2 NER基础方法

NER 基础方法可分为三类：基于规则、机器学习和深度学习方法，它们的评测指标通常基于准确率（Precision，P）、召回率（Recall，P）和 F1 值（F1 score，F1）。如图 4 所示，早期的研究使用规则和机器学习方法，其中，基于规则的方法利用信息列表以及句法 - 词法模式等规则来分类命名实体，但是这些方法依赖于限定语言、领域和文本风格，可移植性、鲁棒性较差的缺陷让研究者寻求新的思路，逐渐将兴趣转向机器学习方法。基于机器学习的方法能自动学习复杂模式或序列标记，这种方式能进行自适应特征学习。但如何定义包含丰富信息的特征是一项困难的工作，而深度学习非线性拟合的能力自动构建语义特征，结合分布式词表示（Distributed Representation）技术有效克制高维空间数据稀疏的特性，为 NER 任务提供新的思路 [4]。

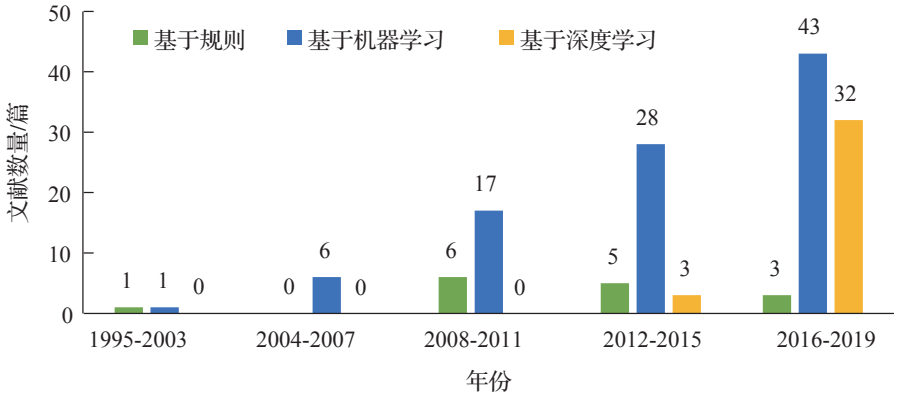


图 4 面向少量标注数据的 NER 基本方法

解决 NER 任务通常基于以上三种方法，其中，传统机器学习和深度学习方法占据主流地位，两者呈上升趋势。特别地，深度学习方法在近几年增长趋势明显，在面向少量标注数据的 NER 研究中，人们倾向于使用具有丰富表征能力的深度学习方法。

### 1.3 少量标注数据的NER

基于规则、统计机器学习和深度学习的方法在通用语料上能取得良好的效果，但在特定领域、小语种等缺乏标注资源情况下，NER 任务往往得不到有效解决。迁移学习（Transfer Learning）NER[5] 为此提供契机，迁移学习利用领域相似性，在领域之间进行数据共享和模型共建，为少量标注数据相关任务提供理论基础。本文从迁移的方法出发，按照知识的表示形式不同，将少量标注数据 NER 方法分为基于数据增强、基于模型迁移、基于特征变换、基于知识链接的方法。如图 5 所示，在这 20 多年间，四种方法的发文数量基本呈上升趋势，整体而言，当前的研究以数据增强、模型迁移为主，而其他的方法通常配合前两种方法使用，在研究中也值得关注。

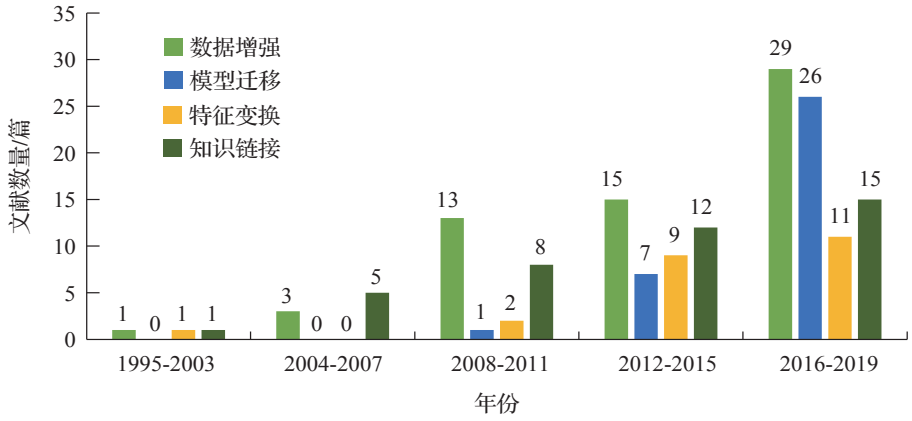


图 5 面向少量标注数据 NER 方法分类

## 2 面向少量标注数据的 NER 方法

### 2.1 基于数据增强的NER

数据增强的方法即：在少量数据集训练模型导致过拟合时，通过样本选择、权重调整等策略以创建高质量样本集，再返回分类器中迭代学习，使之能够较好地完成学习任务的方法[6]。

（1）样本选择。在面向少量标注数据时，最直接的策略是挑选出高质量样本以扩大训练数据。其中，样本选择是数据增强式 NER 的核心模块，它通过一定的度量准则挑选出置信度高、信息量大的样本参与训练，一种典型的思路为主动学习（Active learning）[7] 采样，例如 Shen 等 [7] 利用基于“不确定性”标准，通过挖掘实体内蕴信息来提高数据质量。在实践中，对于给定的序列 $X=(x_1, x_2,\cdots x_i)$ 和标记序列 $Y=(y_1, y_2,\cdots y_i)$，$x$ 被预测为 $Y$ 的不确定性可以用公式（1）来度量，其中 $P(y)$ 为预测标签的条件分布概率，$M$ 为标签的个数，$n$ 为序列的长度：

$$\varnothing^{IE}(x)=-\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M}P(y_i=m)\log P(y_i=m) \quad (1)$$

本文为验证主动学习采样的性能，在人民日报（1998 年）语料中进行实验，共迭代十次，其中 Random 为迭代中随机采样，ALL 为一次训练完所有数据的结果，Active-U 为利用数据增强的结果。实验结果（如图 6）表明，利用数据增强方法在第 7 次迭代中就能达到拟合，节省了 30% 的标注成本。

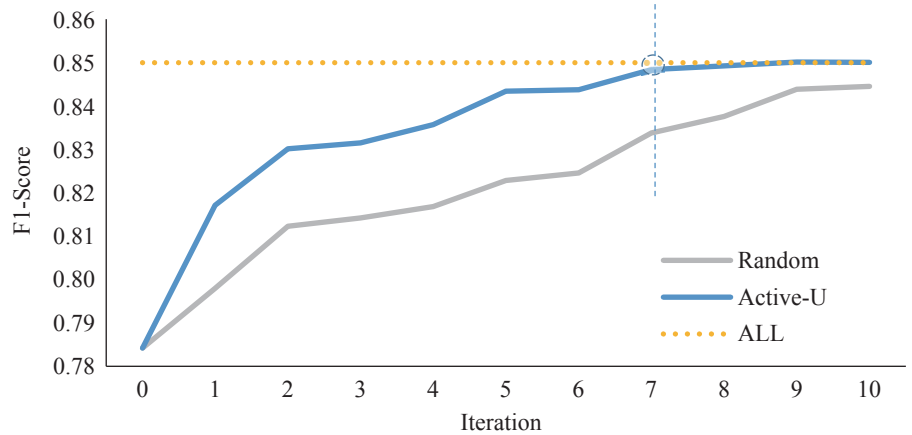


图 6 基于数据增强方法的实例

doi:10.3772/j.issn.2095-915x.2020.04.004

也有不同学者利用其他的度量准则，例如高冰涛等人[8]通过评估源域样本在目标领域中的贡献度，并使用单词相似性和编辑距离，在源域样本集和目标样本集上计算权值来实现迭代学习。Zhang 等人[9]充分考虑领域相似性，分别进行域区分、域依赖和域相关性计算来度量。这些方法利用无监督模式通过降低统计学习的期望误差来对未标记样本进行优化选择，能够有效减少标注数据的工作量。此外，半监督采样也是一种新的思路。例如在主动学习的基础上加入自学习（Self-Training）[10]、自步学习（Self-Paced Learning，SPL）[11]过程，这些方式通过对噪声样本增大学习难度，由易到难地控制选择过程，让样本选择更为精准。

（2）分类器集成。在数据增强中，训练多个弱分类器来获得一个强分类器的学习方式也是一种可行的思路。其中典型的为 Dai 等人[6]提出集成式 TrAdaBoost 方法，它扩展了 AdaBoost 方法，在每次迭代的过程中，通过提高目标分类样本的采样权重、降低误分类实例样本的权重来提高弱分类器的学习能力。TrAdaBoost 利用少量的标签数据来构建对源域标签数据的样本增强，最后通过整合基准弱分类器为一个强分类器来进行训练，实现了少样本数据的学习。之后的研究针对 TrAdaBoost 进行了相应的改进也取得了不错的效果。例如，王红斌等人[12]在分类器集成中增加迁移能力参数，让模型充分表征语义信息，在 NER 中提高精度也能显著减少标注成本。

## 2.2 基于模型迁移的NER

基于模型迁移的基本框架如图 7 所示，其核心思想是利用分布式词表示构建词共享语义空间，然后再迁移神经网络的参数至目标领域，这是一种固定现有模型特征再进行微调(Fine-Tuning) 的方法，在研究中共享词嵌入和模型参数的迁移对 NER 性能产生较大影响。

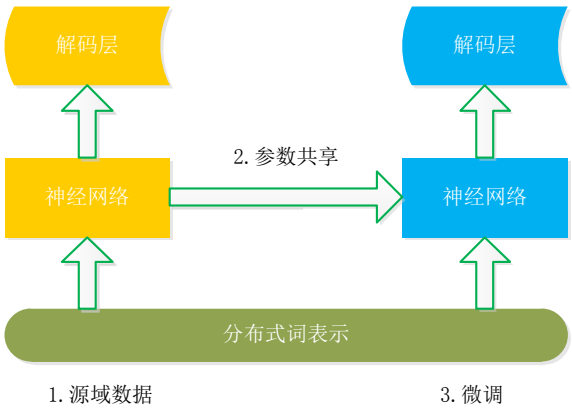


图 7 模型迁移基本结构

（1）共享词嵌入。在 NLP 中，前期工作通常会借助语言预训练模型学习文本的词义信息，这种方式构建了公共的词嵌入表示空间，词嵌入在 NER 中通常作为输入。词向量是共享词嵌入的初步形式，此后，ELMo（Embedding from Language Models）[13]模型利用上下文信息的方式能解决传统词向量不擅长的一词多义问题，还能在一定程度上对词义进行预测逐渐受到人们关注。而 2018 年谷歌提出的 BERT（Bi-directional Encoder Representations from Transformers）[14]预训练模型更是充分利用了词义和语义特性，BERT 是以双向 Transformer[15]为编码器栈的语言模型，它能强有力地捕捉潜在语义和句子关系，基于 BERT 的 NER 在多个任务上也取得 state-of-the-art[16-17]，其基本网络结构如图 8 所示。

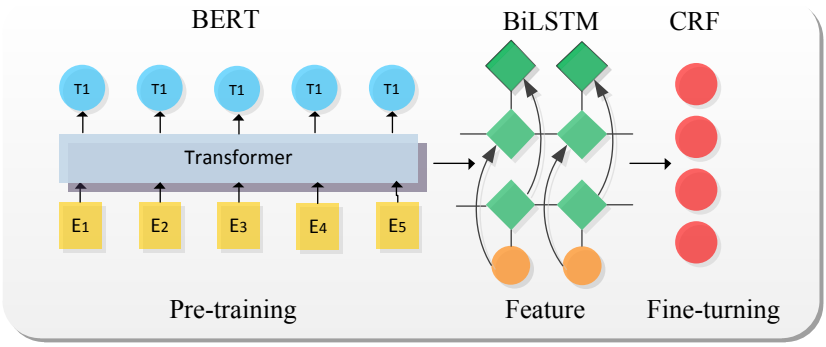


图 8 模型迁移的基础方法 -BERT-BiLSTM-CRF

其中 BERT 作为语义表示输入，BiLSTM 抽取特征，CRF 获取概率最大标签。与传统的 NER 模型相比，该模型最关键的是 BERT 语言模型的引入，BERT 通过无监督建模的方式学习海量互联网语义信息，能充分表征实体的语义信息。在人民日报（1998 年）语料中进行实验，实验结果（如表 1）表明，基于 BERT 的预训练迁移学习模型能有效提高分类的准确率。

表 1 BERT-BiLSTM-CRF 与其他方法的比较 [17]

| 任务 | 准确率 | 召回率 | F1值 |
|---|---|---|---|
| CRF | 85.17 | 85.56 | 85.36 |
| CNN | 86.92 | 85.16 | 86.03 |
| LSTM | 87.25 | 85.76 | 86.49 |
| BiLSTM | 88.34 | 87.61 | 87.97 |
| BiLSTM-CRF | 90.45 | 89.72 | 90.08 |
| BERT-BiLSTM-CRF | 94.73 | 94.99 | 94.86 |

（2）共享参数。共享词嵌入侧重于词义的表示，而共享参数则侧重于模型参数的迁移。例如，Jason 等人 [18] 从神经网络迁移机制以及迁移哪些层进行大量实验，实验结论显示浅层网络学习知识的通用特征，具有很好的泛化能力，当迁移到第 3 层时性能达到饱和，继续迁移会导致“负迁移”的产生。Giorgi 等人 [19] 基于 LSTM 进行网络权重的迁移，首先将源领域模型参数迁移至目标领域初始化，之后进行微调使适应任务需要。而 Yang 等人 [20] 从跨领域、跨应用、跨语言迁移出发测试模型迁移的可行性，在一些 benchmarks 上实现了 state-of-the-art。整体而言，在处理 NER 任务时良好的语义空间结合深度模型将起到不错的效果，在迁移过程中模型层次的选择和适应是难点。

## 2.3 基于特征变换的NER

在面向少量标注数据 NER 任务时，我们希望迁移领域知识以实现数据的共享和模型的共建，在上文中我们从模型迁移的角度出发，它们在解决领域相近的任务时表现良好，但当领域之间存在较大差异时，模型无法捕获丰富、复杂的跨域信息。因此，在跨领域任务中，一种新的思路是在特征变换上改进，从而解决领域数据适配性差的问题。基于特征变换的方法是通过特征互相转移或者将源域和目标域的数据特征映射到统一特征空间 [21]，来减少领域之间差异的学习过程，本节主要从特征选择和特征映射的角度进行探讨。

doi:10.3772/j.issn.2095-915x.2020.04.004

（1）特征选择。即通过一定的度量方法选取相似特征并转换，在源域和目标域之间构建有效的桥梁的策略。例如 Daume 等人 [22] 通过特征空间预处理实现目标域和源域特征组合，在只有两个域的任务中，扩展特征空间 $R^F$ 至 $R^{3F}$，对应于域问题，扩展特征空间至 $R^{(K+1)F}$。然而当 $Y_i$ 与 $Y_j$ 标签空间差异较大时，这种线性组合效果可能不理想，Kim 等人 [23] 从不同的角度出发，进行标签特征的变换，第一种是将细粒度标签泛化为粗粒度标签。例如源域标签中 <start-time> 泛化为 <time>，因为在对其他领域学习时，泛化标签更具一般性。第二种是将简单标签增加领域属性，例如给定“Sunny”词语，可以在原标签 <condition> 增加为 <condition, weather>。

（2）特征映射。即为了减少跨领域数据的偏置，在不同领域之间构建资源共享的特征空间，并将各领域的初始特征映射到该共享空间上 [24]。利用预测的源标签嵌入至目标领域是一种常见策略。例如，Qu 等人 [25] 从领域和标签差异出发，首先训练大规模源域数据，再度量源域和目标域实体类型相关性，最后通过模型迁移的方式微调。其基本步骤为：1、通过 CRF 学习大规模数据的知识；2、使用双层神经网络学习源域与目标域的命名实体的相关性；3、利用 CRF 训练目标域的命名实体。

实验结果显示相较于 Baseline 方法 DeepCRF，TransInit 方法能提高 160% 的性能。

标签嵌入的方式在领域之间有较多共享标签特征时迁移效果不错，但是这种假设在现实世界中并不普遍。一种新的思路是在编解码中进行嵌入适配（如图 10），这种方式利用来自预训练源模型的参数初始化 Bi-LSTM-CRF 基础模型，并嵌入词语、句子和输入级适配。具体而言，在词级适配中，嵌入核心领域词组以解决输入特征空间的领域漂移现象。在句子级适配中，根据来自目标域的标记数据，映射学习过程中捕获的上下文信息。在输出级适配中将来自 LSTM 层输出的隐藏状态作为其输入，为重构的 CRF 层生成一系列新的隐藏状态，进而减少了知识迁移中的损失。

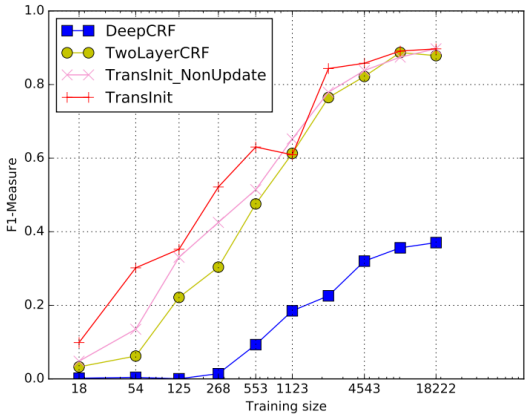


图 9 特征变换方法 **TransInit** 实验结果 [25]

## 2.4 基于知识链接的NER

基于知识链接 [27] 的 NER，即使用本体、知识库等结构化资源来启发式地标记数据，将数据的结构关系作为共享对象，从而帮助解决目标 NER 任务，其本质上一种基于远程监督的学习方式，利用外部知识库和本体库来补充标注实体。例如 Lee 等人 [28] 的框架（如图 11），在 Distant supervision 模块，将文本序列与 NE 词典中的条目进行匹配，自动为带有 NE 类别的大量原始语料添加标签，然后利用 bagging 和主动学习完善弱标签语料，从而实现语料的精炼。一般而言，利用知识库和本体库中的链接信息和词典能实现较大规模的信息抽取任务，这种方法有利于快速实现任务需求。

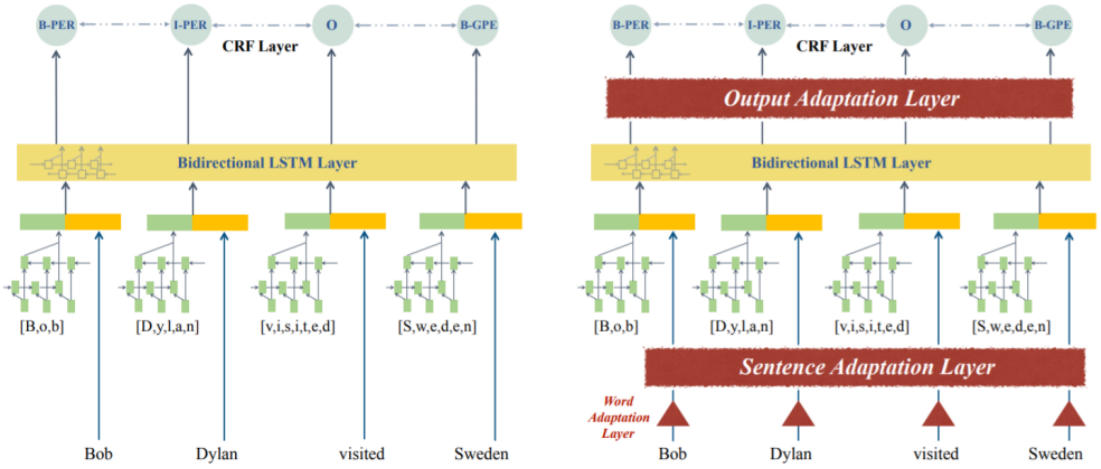


图 **10** 跨域模型对比 [26]

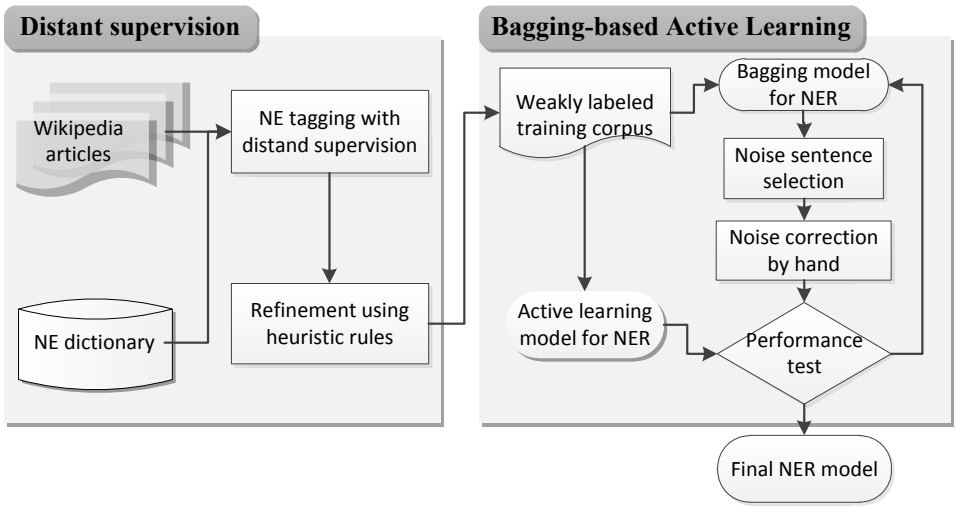


图 **11** 知识链接与数据增强结合模型 [28]

（1）基于知识库。这种方式通常借用外部的知识库来处理 NER、关系抽取、属性抽取等任务，在现实世界中如 Dbpedia、YAGO、百度百科等知识库存在海量结构化信息，利用这些知识库的结构化信息框、日志信息可以抽取出海量知识。例如，Richman 等人 [29] 利用维基百科知识设计了一种 NER 的系统，这种方法利用维基百科类别链接将短语与类别集相关联，然后确定短语的类型。类似地，Pan 等人 [30] 利用一系列知识库挖掘方法为 200 多种语言开发了一种跨语言的名称标签和链接结构。在实践中，较为普遍的是联合抽取实体和实体关系。例如 Ren 等 [31] 的做法，该方法重点解决领域上下文无关和远程监督中的噪声问题，其基本步骤为：

1、利用 POS 对文本语料进行切割一获得提及的实体；

2、生成实体关系对；

3、捕获实体与实体关系的浅层语法及语义特征；

4、训练模型并抽取正确的实体及关系。

doi:10.3772/j.issn.2095-915x.2020.04.004

在 NYT 等语料上进行实验（如表 2），基于知识库的方法相较于基线方法有显著提高。

表 2 不同语料下实体的 F1 值 [31]

| Dataset | NYT | Wiki-KBP | BioInfer |
|---|---|---|---|
| FIGER segmenter | 0.751 | 0.814 | 0.652 |
| Our Approach | 0.837 | 0.833 | 0.785 |

（2）基于本体系统。该方式通过一定的规则，将本体库中的概念映射为实体。例如史树敏等人 [32] 通过构建的 MPO 本体，首先利用 CRF 获得高召回率的实体，再融合规则过滤噪声，最终获得较为精确的匹配模式。相似地，Lima 等人 [33] 通开过发出 OntoLPER 本体系统，并利用较高的表达关系假设空间来表示与实体—实体关系结构，在这个过程中利用归纳式逻辑编程产生抽取规则，这些抽取规则从基于图表示的句子模型中抽取特定的实体和实体关系实例。同样地，李贯峰等人 [34] 首先从 Web 网页提取知识构建农业领域本体，之后将本体解析的结果应用在 NER 任务中，使得 NER 的结果更为准确。这些方法利用本体中的语义结构和解析器完成实体的标准化，在面向少量标注的 NER 中也能发挥出重要作用。

## 2.5 四种方法比较

上述所介绍的 4 种面向少量标注的 NER 方法各有特点，本文从领域泛化能力、模型训练速度、对标注数据的需求和各方法的优缺点进行了细致地比较，整理分析的内容如表 3 所示。

表 3 面向少量标注数据的 NER 4 种方法比较

| 方法 | 泛化能力 | 速度 | 标注需求 | 优点 | 缺点 |
|---|---|---|---|---|---|
| 数据增强 | 弱 | 中 | 需要少量的目标标注数据，需要领域相关标注/无标注数据 | 方法实现简单，准确率较高 | 度量准则的选择依赖经验，有一定的标注成本 |
| 模型迁移 | 中 | 快 | 需要少量的目标标注数据，需要领域相关标注数据 | 模型训练节省开销，鲁棒性较好 | 参数难收敛、难以处理数据分布不一致情况 |
| 特征变换 | 强 | 中 | 需要少量目标标注数据，部分领域相关标注数据，部分生语料 | 能充分利用特征信息，部分实现零样本学习 | 优化问题难求解，易发生过适配 |
| 知识链接 | 中 | 快 | 任何可用的结构化数据 | 能处理大规模数据 | 依赖强假设条件，易造成知识库噪音，准确率低 |

面向少量标注数据 NER，最直接的方法是数据增强，通过优先挑选高质量样本参与训练，这种方法在窄域中能实现较高的准确率。但是针对不同领域所需的策略也不同，领域的泛化能力一般。模型迁移从海量无结构化文本中获取知识，这种方式对目标领域的数据需求较少，只需“微调”模型避免了重新训练的巨大开销，但是它依赖领域的强相关性，当领域差异性太大时，容易产生域适应问题。相较于模型迁移，特征变换更加注重细粒度知识表示，这种方法利用特征重组和映射，丰富特征表示，减少知识迁移中的损失，在一定程度上能实现“零样本”学习，但是这种方法往往难以求出优化解，过适配现象也会造成消极影响。知识链接能利用任何结构化信息，通过知识库、本体库中的语义关系来辅助抽取目标实体，但是这种方法

易产生噪声，实体的映射匹配依赖强假设条件，所需的知识库通常难以满足领域实体的抽取。

## 3 数据资源集合与评测

### 3.1 数据资源集合

近年来，面向少量标注数据的 NER 实验数据集主要有：CoNLL-2003[①]、i2b2-2010[②]、BioCreative-V-CDR[③]、NYT[④]、BioNLP-2016[⑤]、OntoNotes5[⑥]、MSRA[⑦]、人民日报-1998[⑧]等语料。

CoNLL 2003：CoNLL（Conference on Computational Natural Language Learning）是由 ACLSIGNLL (Special Interest Group on Natural Language Learning) 举办的学术会议，其中英语 CoNLL 2003 数据取自路透社新闻，共包含人名、组织、地名、时间和数量五类实体。

i2b2-2010：该语料来源于 49 个临床文档和 20423 个独特句子。每个句子中都标注了三种类型的医疗实体：病症（11192 个）、治疗手段（8099 个）和测试方法（6915 个）。

BioCreative-V-CDR：BioCreative 是生物和生化领域的信息提取系统，在 BioCreative V 中有实体识别（DNER）和化学诱发疾病（CID）关系抽取两项任务，其语料来源于 1500 篇 PubMed 文章，共包括 4409 种化学药品实体，5818 种疾病实体和 3116 种化学疾病实体关系。

NYT-FB：NYT 是利用 Freebase 对齐 NYT（纽约时报）文本的数据集，其中实体有人名、地名、组织名三类，训练数据集为对齐 2005 年和 2006 年纽约时报上文章，测试数据及为对齐 2007 年的文章。

BioNLP-2016：该数据集是由 BioNLP Shared Task 2016 组织负责的 Bacteria Biotope 子任务，该任务数据源为 PubMed 论文摘要，共包含细菌、栖息地和地理位置 3 种实体。

OntoNotes5：OntoNotes 是 BBN 科技公司和科罗拉多等大学共同发起的项目，OntoNotes 5.0 是该项目的最终版本，其数据来源包括新闻、网络日志、博客等文本，其中实体被标注为人名、地名和组织名等 18 个类别。

MSRA：MSRA 语料是 ACL（Association for Computational Linguistics）的一个专门的兴趣小组 SIGHAN（the Special Interest Group for Cchinese Language Processing）在 Bakeoff 2006 评测比赛中提供的语料，共计四类实体人名、地名、机构名和地理 - 政治实体。

人民日报 -1998：该语料来源于 1998 年人民日报，共包括人名、地名和机构名 3 类实体。

### 3.2 资源链接与评测方法

本文中四类面向少量标注数据的典型方法数据资源与评测信息如表 4 所示：

① https://github.com/yuanxiaosc/BERT-for-Sequence-Labeling-and-Text-Classification.
② http://www.i2b2.org.
③ http://www.biocreative.org/tasks/biocreative-v/track-3-cdr.
④ https://github.com/shanzhenren/CoType.
⑤ http://2016.bionlp-st.org/tasks/ bb2.
⑥ https://catalog.ldc.upenn.edu/LDC2013T19.
⑦ https://github.com/buppt/ChineseNER/tree/master/data.
⑧ http://www.icl.pku.edu.cn/icl_res.

doi:10.3772/j.issn.2095-915x.2020.04.004

表 4 典型方法的数据集与评测

| 方法 | Model | 数据集（URL） | 评测指标 | 数值 (%) | 发表年份 | 发表期刊 / 会议 |
|---|---|---|---|---|---|---|
| 数据增强 | CNN-CNN-LSTM[7] | OntoNotes-5.0 | F1 | 85.1 | 2018 | ICLR |
| | CRF[35] | i2b2-2010 | F1 | 82 | 2015 | Journal of Biomedical Informatics |
| | Bi-NER-CDR[9] | Cognitive Atlas（文本）⑨ | P | 75 | 2019 | ITNEC |
| 模型迁移 | BERT-IDCNN-CRF[16] | MSRA | F1 | 94.41 | 2020 | 山东大学学报 |
| | BioBERT[36] | BioCreative-V-CDR | P | 93.68 | 2018 | Bioinformatics |
| 特征变换 | RNN[20] | CoNLL-2003 | F1 | 91.26 | 2017 | ICLR |
| | HUCRFs[23] | CoNLL-2000 | F1 | 81.02 | 2015 | ACL |
| | TransInit[25] | CoNLL-2003 | F1 | 89 | 2016 | EMNLP |
| | BiLSTM-CRF[26] | OntoNotes-5.0 | F1 | 65.33 | 2018 | EMNLP |
| | LitNER[37] | Gutenberg corpus（词典）⑩ | F1 | 79.2 | 2016 | ACL |
| 知识链接 | COTYPE[31] | NYT-FB | F1 | 36.9 | 2017 | WWW17 |
| | Ontology[38] | BioNLP-2016 | F1 | 65.9 | 2019 | BMC Bioinformatics |

## 4 总结与展望

### 4.1 未来的研究方向

在面向少量标注数据的 NER 时，基本的思路为扩大样本集、迁移领域知识、提高模型特征质量以及借用外部结构化数据等，已有的成果证明了这些方法的有效性，但仍然有很多问题存在，需要持续地进行突破。本文对未来可能研究方向进行了思考，以供其他学者参考：

（1）样本选择策略的改进。在样本选择模块，度量准则的定义是难点，优良的准则函数能充分表示信息量，在评估度量准则的优劣时，应对多种度量进行比较，进而选择出对模型拟合帮助较大的准则。此外，在未来的研究中结合生成模型和样本选择是不错的策略，例如利用 GAN[39] 能迭代生成样本的优势以进行数据增强。

（2）考虑零样本学习。在跨领域的知识迁移中，域适应问题最为关键，域适应的有效手段是联合特征映射和模型迁移。而当前的研究主要集中在有监督的情况，也就是考虑目标域和源领域都有标签，这些方法对包含丰富信息的无监督数据利用较少。在未来的研究中可以考虑零样本学习，Chen 等人 [40] 已经在此作了一些尝试，后续的研究可以考虑基于预训练模型进行改进，也可以分析不同领域零样本语料对模型的影响强度。

（3）考虑语义漂移与噪声。当借用领域知识学习时，大多数研究的假设是领域之间有较高的相似性，但这种强假设条件在现实世界里并不普遍存在，如何在领域差异大的场景下进行 NER 成为难题，一种可行的策略是借助第三

⑨ https://www.cognitiveatlas.org.
⑩ http://www.gutenberg.org.

方域完成知识传递，也可以联合多任务学习完成。另外，借用外部的结构化信息进行 NER 任务时，精度问题是最为关心的，如何消减噪声和歧义是需要持续关注的问题，在未来结合语义与深度增强模型区分正例与反例是一个可选的方向。

### 4.2 结语

面向少量标注数据学习是近些年来机器学习领域的新兴方向，基本思想为迁移源领域知识以完成目标领域任务，对这一过程深入研究不仅对资源匮乏的 NLP 任务具有重要意义，也为细分场景下智能系统建设和科技情报服务起到重要作用。在 NER 的研究中，标注资源的匮乏催生面向少量标注数据的方法，基于数据增强、模型迁移、特征变换、知识链接是典型的思路，这些方法在泛化能力、模型训练速度、对标注数据的需求有不同特点。在后续的研究中，可以结合这些方法，例如在模型迁移中利用数据增强的样本来微调，还可以对知识链接中样本采用数据增强的方式精炼以适应任务需求。此外，数据资源的有效利用值得关注，一般而言，通用领域实体较为简单，但如生物医药和化学领域的资源更有意义，后续的研究应多关注专业领域。

## 参考文献

[1] 戴国强 . 推进竞跑阶段的创新情报研究 [J]. 情报学报 , 2019, 38(8):771-777.

[2] 谷威 , 田欣 . 基于条件随机场和篇章校对的有机物命名实体识别方法研究 [J]. 情报工程 , 2018, 4(5):64-72.

[3] Goyal A, Gupta V, Kumar M. Recent Named Entity Recognition and Classification techniques:A systematic review[J]. Computer Science Review, 2018(29):21-43.

[4] 高甦 , 金佩 , 张德政 . 基于深度学习的中医典籍命名实体识别研究 [J]. 情报工程 , 2019, 5(1):113-123.

[5] 刘宇飞 , 尹力 , 张凯 , 等 . 基于深度迁移学习的技术术语识别——以数控系统领域为例 [J]. 情报杂志 , 2019, 38(10):168-175.

[6] Wen Y D, Qiang Y, Gui R X, et al. Boosting for transfer learning[C]. Proceedings of the 24th international conference on Machine learning, 2007:193-200.

[7] Shen Y, Yun H, Lipton Z C, et al. Deep active learning for named entity recognition[J]. arXiv preprint arXiv:1707.05928, 2017.

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

[9] Shun Z, Shao F L, Jiang F G, et al. Recognizing Small-Sample Biomedical Named Entity Based on Contextual Domain Relevance[C]. 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). 2019:1509-1516.

[10] 钟志农 , 刘方驰 , 吴烨 , 等 . 主动学习与自学习的中文命名实体识别 [J]. 国防科技大学学报 , 2014, 36(4):82-88.

[11] 梅涛 . 基于主动自步学习的文本分类研究 [D]. 西安 : 西安电子科技大学 , 2018.

[12] 王红斌 , 沈强 , 线岩团 . 融合迁移学习的中文命名实体识别 [J]. 小型微型计算机系统 , 2017, 38(2):346-351.

[13] Matthew E P, Mark N, Mohit I, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802. 05365, 2018.

[14] Jacob D, Ming W C, Kenton L, et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[15] Ashiah V, Noam S, Niki P, et al. Attention Is All You Need[J]. arXiv preprint arXiv:1706.03762, 2017.

[16] 李妮 , 关焕梅 , 杨飘 , 等 . 基于 BERT-IDCNN-CRF 的中文命名实体识别方法 [J]. 山东大学学报 ( 理学版 ), 2020, 55(1):102-109.

[17] 王子牛 , 姜猛 , 高建瓴 , 等 . 基于 BERT 的中文命名

doi:10.3772/j.issn.2095-915x.2020.04.004

实体识别方法 [J]. 计算机科学 , 2019, 46(S2):138-142.

[18] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks?[C]. Proceedings of the 28th Conference on Neural Information Processing Systems(NIPS), 2014:3320-3328.

[19] Giorgi J M, Bader G D. Transfer learning for biomedical named entity recognition with neural networks[J]. Bioinformatics, 2018, 34(23):4087-4094.

[20] Zhi L Y, Ruslan S, William W C. Transfer learning for sequence tagging with hierarchical recurrent networks[J]. arXiv preprint arXiv:1703.06345, 2017.

[21] Sinno J P, Ivor W T, James T K, et al. Domain adaptation via transfer component analysis[J]. IEEE Transactions on Neural Networks, 2010, 22(2):199-210.

[22] Hal D. Frustratingly easy domain adaptation[C]. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007:256-263.

[23] Young B K, Karl S, Rruhi S, et al. New transfer learning techniques for disparate label sets[C]. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015, 1:473-482.

[24] 孟创纪 . 基于特征映射的深度迁移学习研究 [D]. 兰州 : 兰州大学 , 2019.

[25] Lizhen Q, Gabriela F, Liyuan Z, et al. Named Entity Recognition for Novel Types by Transfer Learning[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016:899-905.

[26] Bill Y L, Wei L. Neural Adaptation Layers for Cross-domain Named Entity Re-cognition[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018:2012–2022.

[27] Jason A F, Sen W, Alexander R, et al. SwellShark:A Generative Model for Biomedical Named Entity Recognition without Labeled Data[J]. arXiv preprint arXiv:1704.06360, 2017.

[28] Lee S, Song Y, Choi M, et al. Bagging-based active learning model for named entity recognition with distant supervision[C]. Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp). IEEE, 2016:321-324.

[29] Alexander E R, Patrick S. Mining Wiki Resources for Multilingual Named Entity Recognition[C]. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies. 2008:1-9.

[30] Xiao M P, Bo L Z, Jonathan M, et al. Cross-lingual name tagging and linking for 282 languages[C]. proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017:1946–1958.

[31] Ren X, Wu Z, He W, et al. Cotype: Joint extraction of typed entities and relations with knowledge bases[C]. proceedings of the 26th International Conference on World Wide Web. 2017:1015-1024.

[32] 史树敏 , 冯冲 , 黄河燕 , 等 . 基于本体的汉语领域命名实体识别 [J]. 情报学报 , 2009, 28(6):857-863.

[33] Rinaldo L, Bernard E, Fred F. OntoILPER: an ontology- and inductive logic programming-based system to extract entities and relations from text[J]. Knowledge and Information Systems, 2018, 56(1):223-255.

[34] 李贯峰 , 张鹏 . 一个基于农业本体的 Web 知识抽取模型 [J]. 江苏农业科学 , 2018, 46(4):201-205.

[35] Chen Y K, Lasko T A, Mei Q Z, et al. A Study of Active Learning Methods for Named Entity Recognition in Clinical[J]. Journal of Biomedical Informatics, 2015(58):11-18.

[36] Lee J, Yoon W, Kim S, et al. BioBERT:a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4):1234-1240.

[37] Brooke J, Hammond A, Baldwin T. Bootstrapped text-level named entity recognition for literature[C]. proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2:Short Papers). 2016:344-350.

[38] Karadeniz I, Özgür A. Linking entities through an ontology using word embeddings and syntactic re-ranking[J], BMC bioinformatics, 2019, 20(1):156.

[39] 程显毅 , 谢璐 , 朱建新 , 等 . 生成对抗网络 GAN 综述 [J]. 计算机科学 , 2019, 46(3):74-81.

[40] Chen J, Xiao B L, Yue Z. Cross-Domain NER using Cross-Domain Language Modeling[C]. proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:2464-2474.