阿拉伯语命名实体识别综述：过去、近期进展与未来趋势

Xiaoye Qu、Yingjie Gu、Qingrong Xia、Zechang Li、Zhefeng Wang、Baoxing Huai

摘要——随着越来越多的阿拉伯语文本出现在互联网上，从这些阿拉伯语文本中提取重要信息尤为有用。作为一项基础技术，命名实体识别（NER）是信息抽取技术的核心组成部分，同时也在许多其他自然语言处理（NLP）系统中发挥着关键作用，例如问答和知识图谱构建。本文全面回顾了阿拉伯语命名实体识别的发展，尤其是深度学习和预训练语言模型方面的最新进展。具体而言，我们首先介绍阿拉伯语命名实体识别的背景，包括阿拉伯语的特点以及现有的阿拉伯语命名实体识别资源。随后，我们系统地回顾阿拉伯语命名实体识别方法的发展。传统的阿拉伯语命名实体识别系统侧重于特征工程和设计领域特定规则。近年来，通过以连续向量表示文本，深度学习方法取得了显著进展。随着预训练语言模型的发展，阿拉伯语命名实体识别的性能进一步提升。最后，我们总结了阿拉伯语命名实体识别与其他语言的命名实体识别方法之间的方法差距，这有助于勾勒阿拉伯语命名实体识别的未来方向。

索引词—命名实体识别 、阿拉伯语文本、深度学习、预训练语言 模型

arXiv:2302.03512v3 [cs.CL] 8 Aug 2023

1 引言

阿拉伯语 RABIC 是阿拉伯世界的官方语言，也是世界上形态屈折和派生现象最丰富的自然语言之一。阿拉伯语是全球使用人数最多的十种语言之一，拥有4.2亿母语使用者，并有超过25种广为人知的方言[1]。阿拉伯语具有许多独特特征，使阿拉伯语文本的理解既充满挑战，也颇具趣味。随着社交媒体平台日益普及，互联网上可获取的阿拉伯语文本数量显著增长。为了有效处理这些文本，阿拉伯语命名实体识别（ANER）受到越来越多的关注。命名实体识别（NER）旨在识别对应于特定类型的提及，例如人名、地点和组织。除通用领域外，在医学领域等若干特定领域中，NER还可以抽取药物和临床操作。借助阿拉伯语命名实体识别，可以便捷地从不同领域抽取重要信息。NER可用于各种下游任务，例如关系抽取 [2],实体链接 [3],事件抽取 [4],共指消解 [5],和机器翻译 [6]。

Cite	Year	Type	标题
[9]	2009	报告	NERA：命名实体识别 阿拉伯语
[10]	2014	**综述**	阿拉伯语命名实体 识别与分类综述
[11]	2015	报告	命名实体识别 面向阿拉伯语社交媒体
[12]	2016	**综述**	阿拉伯语命名实体识别 —综述与分析
[13]	2017	**综述**	机器学习的比较性综述 用于阿拉伯语命名 实体识别
[14]	2019	报告	阿拉伯语命名实体识别 使用深度学习方法
[15]	2019	报告	阿拉伯语命名实体识别：哪些方法有效 以及下一步的发展方向
[16]	2020	**综述**	最近一项关于阿拉伯语命名实体 识别在社交媒体上的综述。

表1：现有的阿拉伯语命名实体识别综述和报告。“报告”指竞赛报告，或对系统实现进行详细介绍的报告。“综述”指对不同方法进行全面考察。

作为自然语言处理（NLP）的基础组成部分，命名实体识别（NER）在英语中已得到充分研究，但对阿拉伯语的全面探索仍然不足。英语命名实体识别比阿拉伯语命名实体识别更容易，因为英语中的大多数提及都以大写字母开头，而阿拉伯语没有这一特点。此外，阿拉伯语的形态结构十分复杂，原因在于其屈折变化极其丰富，并且存在许多变体 [7]。阿拉伯语分为三类：古典阿拉伯语（CA）、现代标准阿拉伯语（MSA）和方言阿拉伯语（阿拉伯方言） [8]。古典阿拉伯语是最初的阿拉伯语，已有超过 1,500 年的使用历史，通常用于大多数阿拉伯语宗教文本。现代标准阿拉伯语是联合国的官方语言之一，广泛用于当今的阿拉伯语报纸、信函和正式会议，也是研究人员关注的重点。阿拉伯方言是用于非正式日常交流的口语阿拉伯语。

阿拉伯语命名实体识别解决方案的发展，随着算法和计算能力的进步而不断演进。在阿拉伯语命名实体识别中，早期的命名实体识别系统依赖人工制定的规则。随后，机器学习方法利用人工设计的特征进行训练，以将词语分类到不同的命名实体类别中。之后，

• *X. Qu*、*Y.Gu*、*Q. Xia*、*Z. Li*、*Z. Wang* 和 *B. Huai* 就职于华为云系统人工智能创新实验室，地址：中国杭州 *310051*。电子邮箱：{*quxiaoye*、*guyingjie4*、*xiaqingrong*、*lizechang1*、*wangzhefeng*、*huaibaox- ing*}*@huawei.com*。*Zhefeng Wang* 是通信作者。

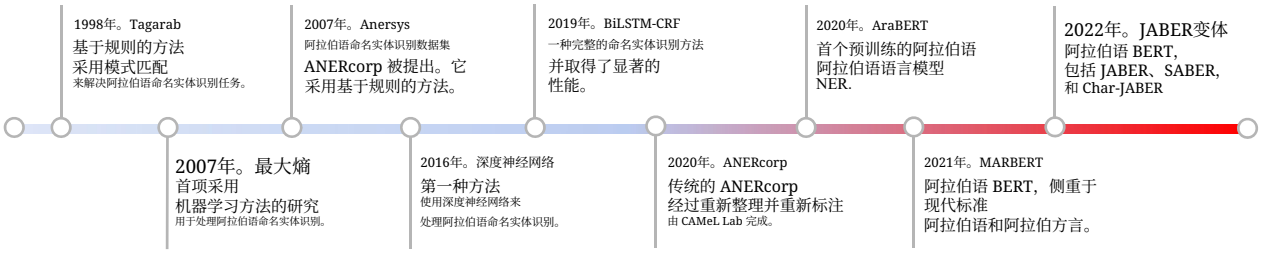


图1：阿拉伯语命名实体识别的发展里程碑。

依赖词嵌入并自动学习实体分类表示的深度学习技术逐渐占据主导地位。近年来，基于Transformer的预训练语言模型凭借强大的计算能力，进一步推动了这一领域的发展。本文主要关注阿拉伯语命名实体的发展。我们并不声称本文涵盖了所有语言中的全部命名实体识别研究。

开展本综述的动机。过去几年中，大量方法探索了将深度学习方法应用于阿拉伯语命名实体识别，并取得了不错的表现。我们注意到，这些方法采用了相似的网络，主要是可选用双向长短期记忆网络进行编码，并用条件随机场进行解码。然而，英语领域的解决方案要丰富得多，我们希望这些方法也能助力阿拉伯语命名实体识别的发展。同时，据我们所知，该领域鲜有系统性的综述。如表1所示，我们列出了以往关于阿拉伯语命名实体识别的报告和综述。Shaalan 等人 [10] 于2014年对阿拉伯语命名实体识别进行了综述，但重点关注手工特征。Dandashi 等人 [12] 和 Salah 等人 [13] 主要探讨了用于阿拉伯语命名实体识别的机器学习方法。近期，Ali 等人 [16]于2020年研究了社交媒体上的阿拉伯语命名实体识别，但实验使用的是英语数据集。总之，现有综述并未全面涵盖现代基于深度学习的命名实体识别系统。表1中还有若干介绍竞赛方法或系统架构的报告，例如 [15]，它是由 Topcoder.com 举办的阿拉伯语命名实体识别挑战赛的冠军。这些报告描述了有效的系统，但未对阿拉伯语命名实体识别的发展进行全面分析。此外，近期也有一些综述关注英语通用领域中的深度学习技术 [17], [18] 。这些综述没有考虑阿拉伯语的语言特征及其特有挑战。因此，目前缺少对阿拉伯语现有命名实体识别方法的全面回顾。我们希望本综述能系统梳理以往的阿拉伯语命名实体识别方法（如图1所示）。值得注意的是，在阿拉伯语这类资源较少的语言中发现有价值的信息，也是自然语言处理社区的重要机遇。

开展本综述的贡献。我们总结贡献如下：

- 据我们所知，这是首篇聚焦于阿拉伯语命名实体识别近期进展的全面综述，涵盖阿拉伯语命名实体识别语料库、现有资源、基于规则的方法、机器学习方法、深度学习方法以及最新的预训练语言模型方法。此外，我们还综述了应用于阿拉伯语命名实体识别问题的深度学习技术中最具代表性的方法。
- 在本综述中，鉴于不同数据划分下以往方法的比较有失公平，我们谨慎地汇总了阿拉伯语命名实体识别在 ANERcorp 和 AQMAR 数据集上的结果。这样，研究人员便能轻松比较不同方法。我们认为，公平的比较将有助于推动更具竞争力的阿拉伯语命名实体识别方法。
- 我们概述了阿拉伯语命名实体识别的潜在发展方向，包括阿拉伯语命名实体识别的替代解决方案、细粒度和嵌套式阿拉伯语命名实体识别、低资源和远程监督命名实体识别，以及面向阿拉伯语命名实体识别的跨语言知识迁移。这些未来趋势都将推动下游应用更加有效。

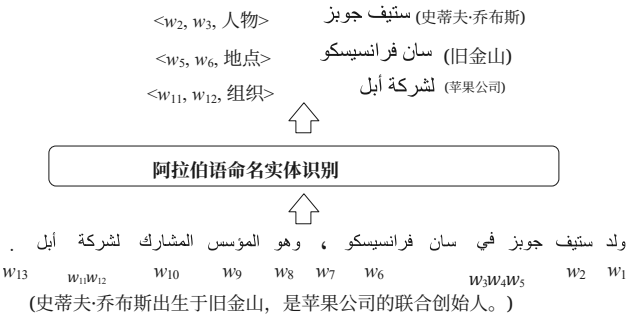


图 2：阿拉伯语命名实体识别任务示意图。

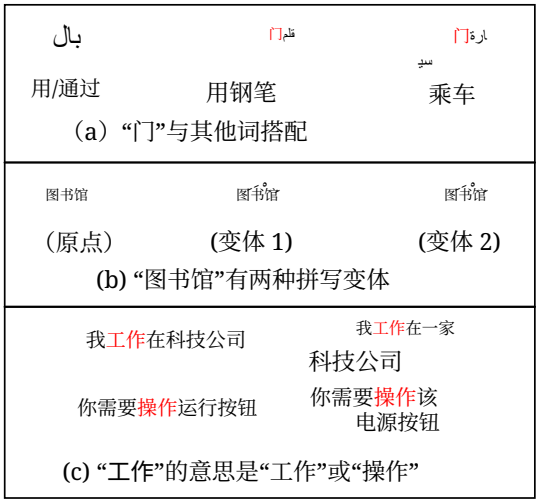
2 背景

本节首先对阿拉伯语命名实体识别进行正式定义。随后，我们介绍阿拉伯语命名实体识别面临的挑战。最后，我们汇总当前的阿拉伯语命名实体识别资源，并说明评估指标。

2.1 什么是阿拉伯语命名实体识别

阿拉伯语命名实体识别旨在定位阿拉伯语文本中的命名实体 $s =< w_n, \ldots\ldots w_2, \ w_1 >$（与中文或英语从左向右书写不同，阿拉伯语从右向左书写），并将其划分到预定义的实体类别中。与其他常见的命名实体识别任务一样，阿拉伯语命名实体识别也可分为通用领域（包括人物、地点和组织类型）以及特定领域（如生物医学、新闻和金融）。

NER 的典型输出有两种：一种是元组列表 $< P_s, \ P_e, \ t >$，其中每个元组都是一个命名实体



(a) “门”与其他词搭配

(b) “图书馆”有两种拼写变体

(c) “工作”的意思是“工作”或“操作”

图3：阿拉伯语命名实体识别所面临挑战的三个示例。

实体的提及。这里， P_s 和 P_e 分别是命名实体提及的起始位置和结束位置， t 是该提及在预定义类别集合中的类型。另一种方式将 NER 视为序列标注任务，为每个词元分配一个特殊标签（例如 B-PER、I-PER、O）。图 2 展示了一个示例：阿拉伯语命名实体识别模型从给定的阿拉伯语文本中识别出三个命名实体。

2.2 阿拉伯语命名实体识别面临的挑战

由于阿拉伯语的特殊性和独特性，命名实体识别在处理阿拉伯语时尤其具有挑战性。阿拉伯语的主要特征给命名实体识别任务带来了以下不小的挑战：

- **无大小写**：大小写是一种极具区分性的正字法特征，例如专有名词和缩写，它有助于命名实体识别系统准确识别特定类型的实体。与英语不同，英语中的命名实体通常以大写字母开头，而大小写并非阿拉伯语词的特征。因此，仅依靠在专有名词词典中查找条目并不是解决这一问题的恰当方案 [19]。

- **黏着特性**：阿拉伯语具有很强的黏着特性，一个词可以由不同组合的前缀、词元和后缀构成，这导致其形态学非常复杂 [20]。同时，阿拉伯语中的命名实体常与其他词相关联，如介词（例如，for/to、as 和 by/with）或连词（例如，and） [10]。如图3(a)所示，介词与其他词结合在一起。常见的处理方式是获得以空格字符分隔的文本分词 [21]。

- **拼写变体**：在阿拉伯语中，一个词可能有不同的拼写形式，但仍指向同一个含义相同的词，从而造成多对一歧义，这也称为转写歧义 [22]，，如图3(b)所示。为解决这一问题，研究人员可以通过字符串距离计算 [23]等机制，将各种变体规范化为统一形式。

- **无短元音**：短元音，也称附加符号，对于发音和消歧至关重要。然而，大多数现代阿拉伯语文本不包含附加符号。因此，根据词语出现的上下文，阿拉伯语中的一种词形可能指代两个或更多不同的词或含义，从而产生一对多歧义。如图3(c)所示，该词在不同的使用上下文中含义不同。阿拉伯语母语者很容易理解这一现象，但阿拉伯语命名实体识别系统 [24]却难以理解。短元音非常重要，因为不同的附加符号表示不同的含义，有助于消歧阿拉伯语实体的结构和词汇类型。长期以来，这类由短元音引起的歧义只能借助上下文信息和一些语言知识来解决 [25]。近年来，随着阿拉伯语预训练模型（如 AraBERT [26]和 AraELECTRA [27]）的出现，附加符号带来的挑战进一步得到缓解。

- **语言资源匮乏**：充足的数据资源，包括标注语料库和高质量词典/地名词典，是构建强大的命名实体识别系统的关键。然而，与拥有大规模标注语料库的英语和中文等语言不同，由于缺乏标注，阿拉伯语命名实体识别可用的阿拉伯语语言资源数量有限 [28]。即便如此，仍有少数阿拉伯语语料库已创建并公开供研究使用，我们将在第2.3节详细介绍这些阿拉伯语资源。

2.3 阿拉伯语命名实体识别的现有资源

本节总结了一些广泛使用的阿拉伯语命名实体识别数据集和热门工具。如表2所示，大多数数据集主要通过标注新闻文章构建，且实体类型数量较少。考虑到某些阿拉伯语命名实体识别语料库仅向持有付费许可协议的用户开放，例如 ACE 2003，研究人员很难获取这些资源，因为它们并非免费。下面介绍三个具有代表性的数据集：

ANERcorp [29] 包含来自 11 个不同来源的 316 篇文章，以构建通用语料库，并由训练集和测试集组成。它包含 150,286 个词元。除了组织、地点和人物等传统实体外，ANERcorp 还引入了杂项类型，用于表示不属于其他任何类别的实体。值得注意的是，多年来，该数据集被划分为许多不同的配置，导致不同论文和系统之间难以进行公平比较 [30]–[32]。2020年，CAMeL Lab 的一组研究人员与该语料库的创建者会面并征询其意见，共同商定了一个精确的划分方案，同时接受了对原始数据集的少量修正。因此，ANERcorp[1] 发布了一个新版本，我们建议将此划分作为评估阿拉伯语命名实体识别性能的标准数据集。在这个

1. https://camel.abudhabi.nyu.edu/anercorp/

表2：阿拉伯语命名实体识别标注数据集列表。“#Tags”指实体类型的数量。

语料库	Year	文本来源	#标签	URL
ANERcorp	2007	网站、新闻、杂志	4	https://camel.abudhabi.nyu.edu/anercorp/
ACE 2003	2004	广播新闻、新闻通讯体裁	7	https://catalog.ldc.upenn.edu/LDC2004T09
ACE 2004	2004	广播新闻、新闻通讯体裁	7	https://catalog.ldc.upenn.edu/LDC2005T09
ACE 2005	2005	转录文本、新闻	7	https://catalog.ldc.upenn.edu/LDC2006T06
ACE 2007	2007	转录文本、新闻	7	https://catalog.ldc.upenn.edu/LDC2014T18
REFLEX	2009	路透社新闻	4	https://catalog.ldc.upenn.edu/LDC2009T11
AQMAR	2012	阿拉伯语维基百科	7	http://www.cs.cmu.edu/~ark/阿拉伯语命名实体识别/
OntoNotes 5.0	2013	News	18	https://catalog.ldc.upenn.edu/LDC2013T19
WDC	2014	维基百科	4	https://github.com/Maha-J-Althobaiti/Arabic NER – 维基语料库
DAWT	2017	维基百科	4	https://github.com/klout/opendata/tree/master/wiki – 标注
CANERCorpus	2018	宗教	20	https://github.com/RamziSalah/Classical-Arabic-Named- 实体识别语料库
Wojood	2022	维基百科	21	https://ontology.birzeit.edu/Wojood/

数据集的新版本中，前 5/6 词语所在的句子被划入训练集，其余句子则划入测试集。训练集划分包含 125,102 个词，测试集划分包含 25,008 个词。

AQMAR [33] 是一个基于阿拉伯语维基百科的小型命名实体识别数据集，包含约 3,000 个句子。创建者从历史、技术、科学和体育领域中选取了 31 篇文章，并依据篇幅、跨语言链接以及主观质量判断，确定这些文章适合纳入数据集。与新闻文本相比，维基百科文章的类别丰富得多。因此，传统的人物、地点和组织类别可能并不适用，而且恰当的实体类别可能因领域而异。因此，除经典类别外，AQMAR 还专门针对特定领域设计了自定义实体类别。

CANERCorpus [34] 旨在支持对古典阿拉伯语的研究；古典阿拉伯语是《古兰经》和圣训的官方语言。该语料库以伊斯兰主题为基础，收录了 7,000 多篇圣训，涵盖先知穆罕默德生活的方方面面，并为伊斯兰教提供恰当的指导。其命名实体类别也专门针对伊斯兰主题设计，涵盖真主安拉、先知等实体，这些内容此前尚未得到处理。三位在计算语言学领域拥有五年以上经验的专家完成了标注。这项工作为古典阿拉伯语的发展作出了重要贡献。

在这些阿拉伯语命名实体识别数据集中，ANERcorp 是最早发布的，也是相关研究最多的数据集，代表了传统的阿拉伯语命名实体识别。除经典的实体类别外，AQMAR 还纳入了历史、技术、科学和体育领域的领域实体，从而支持对特定领域命名实体识别的研究。不过，与 ANERcorp 相比，它的规模要小得多。在 CANERCorpus 出现之前，尚无古典阿拉伯语数据集。该数据集专注于古典阿拉伯语，并在大规模标准伊斯兰教语料库上进行标注。尽管实体类别也很丰富，但分布不够均衡，有些类别的占比甚至不到 1%。

除了数据集之外，我们还介绍一些常用的阿拉伯语自然语言处理工具，这些工具有助于阿拉伯语命名实体识别处理。与直接以阿拉伯语词元作为输入并输出命名实体识别标签的端到端阿拉伯语命名实体识别模型不同，在传统的基于特征的时代，阿拉伯语命名实体识别还需要一些重要的预处理步骤，例如分词和词性标注。MADAMIRA[35] [2] 提供了一个用于阿拉伯语形态分析和消歧的在线演示。它旨在处理一些基础任务，这些任务通常是复杂系统的首个步骤或中间步骤，包括元音符号标注、词元化、形态消歧、词性标注、词干提取、释义标注和分词。输入文本会被转换为Buckwalter表示，并传入形态分析模块。随后，支持向量机提取封闭类特征，而语言模型则获取开放类特征。排序组件会对每个词的分析列表进行评分，之后将其传递给分词组件。基本阶段的分块组件和命名实体识别器均使用支持向量机处理分析结果和分词结果。如MADAMIRA [35], 所示，阿拉伯语命名实体识别组件直接受益于前序组件的输出。

分词是阿拉伯语命名实体识别中的一项关键预处理步骤。然而，大多数分词器都专注于现代标准阿拉伯语，在处理包含方言词汇和语法的阿拉伯语文本时表现不佳。为处理阿拉伯方言文本，Monroe 等人 [36] 提出了斯坦福阿拉伯语分词器，该工具在埃及阿拉伯语语料库上的 F1 分数达到 95.1 %。斯坦福阿拉伯语分词器[3] 是字符级条件随机场模型的扩展。该分词器扩展了常规标签空间，并引入了两条阿拉伯语特有的正字法规则。该方法采用领域自适应方法，对非正式阿拉伯语和阿拉伯方言文本进行分词。

Farasa [37] [4] 是一款基于支持向量机、快速且准确的阿拉伯语分词器，它通过采用多种特征和词典，对词语可能的切分结果进行排序。Farasa 支持分词、词干提取、命名实体识别、词性标注和元音符号标注。实验表明，在信息检索任务中，该工具的表现远优于斯坦福阿拉伯语分词器和 MADAMIRA。

由于大多数阿拉伯语自然语言处理工具缺乏灵活性，因此推出了 CAMeL [38] [5] 来解决这一问题。它使用 Python 3 实现，并提供一系列预

2. https://camel.abudhabi.nyu.edu/madamira/ 3. https://nlp.stanford.edu/projects/arabic.shtml 4. https://farasa.qcri.org/ 5. https://github.com/CAMeL-Lab/camel tools –

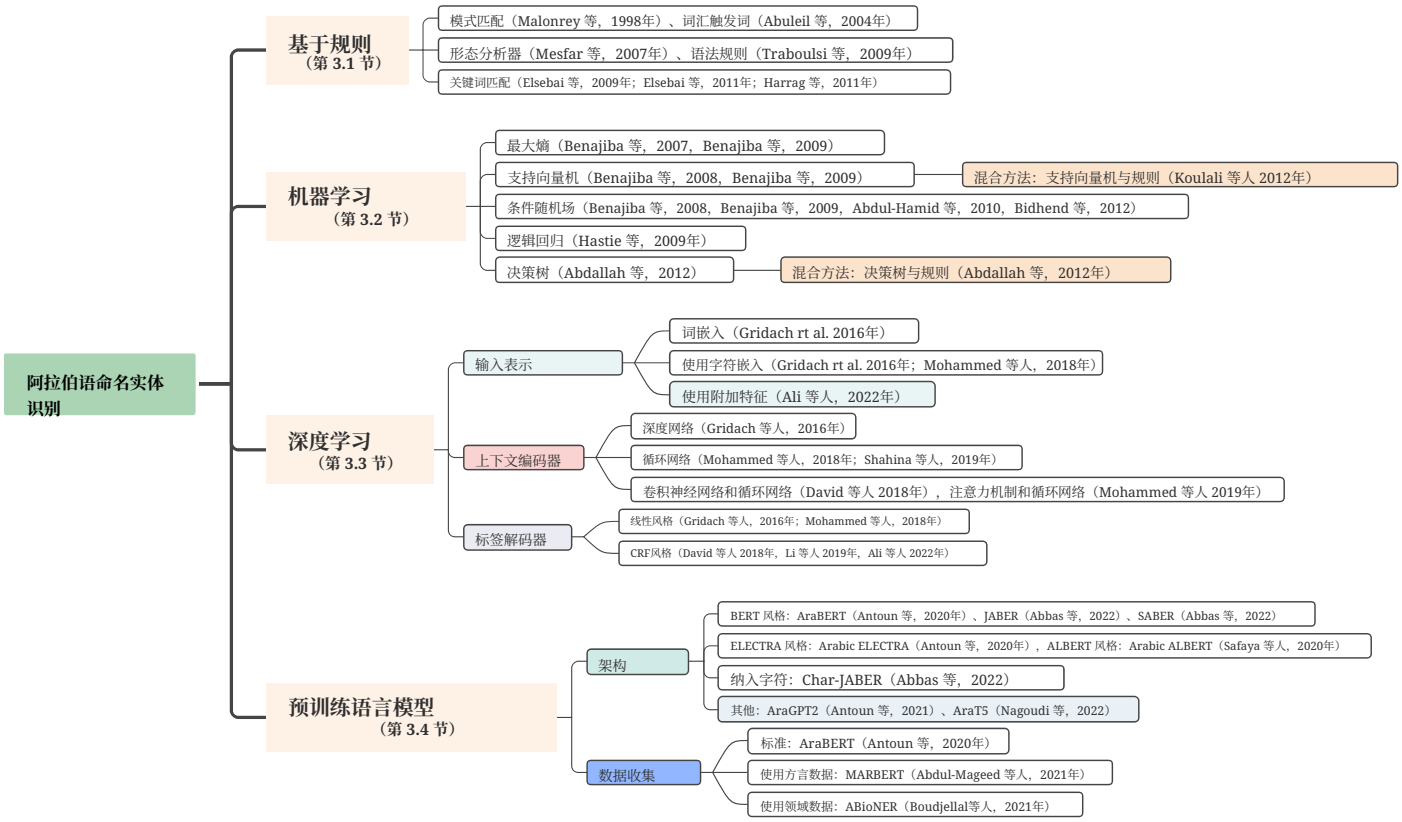


图4：基于不同方法的阿拉伯语命名实体识别方法分类体系。基于规则的方法和机器学习方法将在第3.1节和第3.2节中介绍。用于阿拉伯语命名实体识别的深度学习方法和预训练语言模型将在第3.3节和第3.4节中阐述。在深度学习方法部分，我们从三个角度介绍阿拉伯语命名实体识别的发展，包括输入表示、上下文编码器和标签解码器。在预训练语言模型部分，我们根据其架构和数据收集范围说明阿拉伯语命名实体识别方法。

处理步骤，例如转写和正字法规范化。在命名实体识别方面，CAMeL 对 AraBERT 进行了微调，实验表明，该工具的表现优于基于条件随机场的系统。具体而言，CAMeL 在 ANERCorp 数据集上取得了 83% 的总体 F1 分数，比基于条件随机场的系统高出 4 个百分点。特别是在人物类别上，CAMeL 的 F1 分数比基于条件随机场的模型提高了 14%。

2.4 阿拉伯语命名实体识别指标

我们采用的阿拉伯语命名实体识别评估指标分为精确匹配指标和宽松匹配指标。对于精确匹配指标，我们使用精确率、召回率和F分数来评估各类方法。只有当预测实体的边界和实体类型都正确时，才将其视为正确。精确率反映了提取正确实体的能力，而召回率衡量的是尽可能多地找出实体的能力。因此，F分数体现了精确率与召回率之间的平衡。不过，精确匹配指标的标准有些过于严格。一些方法也采用宽松匹配指标进行评估。例如，MUC-6 [39]提出通过实体类型评估系统，只要预测结果与真实边界存在重叠，就忽略边界。由于宽松匹配指标

较为复杂，而且不便于分析错误案例，因此至今仍未得到广泛应用。

3 方法

为了揭示阿拉伯语命名实体识别的发展，我们介绍几种具有代表性的方法。如图 4 所示，阿拉伯语命名实体识别方法可归纳为四种范式：基于规则的方法、机器学习方法、深度学习方法，以及近期的预训练语言模型方法。在接下来的章节中，我们将介绍各个范式中的重要方法。

3.1 基于规则

在机器学习和深度学习方法广泛应用于命名实体识别任务之前，最常用的是基于规则的方法。基于规则的命名实体识别系统 [9], [40]–[44]依赖于由语言学专家提取并验证、经人工编制的局部语法规则。在命名实体出现的上下文中，地名词典和词典（或字典）被用作语法规则。基本上，地名词典中的大多数词语都用作触发词，帮助在给定文本中识别命名实体。

在阿拉伯语命名实体识别任务中，采用基于规则的方法的研究寥寥无几，因为编码人类专业知识需要

大量专家语言学家的工作，而且通常只适用于单一语言。在阿拉伯语命名实体识别的早期阶段，Maloney 和 Niv [45]提出了一种名为 TAGARAB 的、基于规则的模式匹配方法，用于阿拉伯语命名实体识别任务。该系统结合高精度的形态分析和模式匹配引擎，支持识别姓名、日期、时间和数值实体。

然而，TAGARAB 只考虑阿拉伯语的形态学，这是一种低层次的规则特征。因此，Traboulsi 等人 [46] 探讨了在未标注阿拉伯语语料库中使用所谓的局部语法形式体系方法，并结合有限状态自动机识别阿拉伯语新闻文本中的人名模式。结果表明，考虑所有常见语法可以节省大量时间和精力。此外，Elsebai 等人 [47]采用了一种基于规则的方法，使用 Buckwalter 阿拉伯语形态分析器（BAMA） [48] 以及一组关键词来分析可能包含人名命名实体的文本，F值达到 89%。

最近，Harrag 等人 [49] 提出了一种基于有限状态转换器的实体提取器，用于从圣训文本中识别关键实体。具体而言，他们首先采用有限状态转换器，从一组有限的标签中为实体分配概念标签，以捕捉不同类型的命名实体。随后，对于每个检测到的概念，系统的目标是填充一个包含事件描述的模式。该模型在一组圣训文本上取得了 52.00% 的F值。

总而言之，基于规则的命名实体识别方法基于扎实的语言学知识，易于理解。不过，基于规则的系统也存在明显的缺点。首先，维护或更新这类系统需要投入大量人力和时间。其次，如果无法获得具备所需知识和背景的语言学家，问题就会更加棘手。

3.2 机器学习

一般而言，基于机器学习的命名实体识别方法[28] 利用机器学习算法来学习实体标注决策。它们将命名实体识别问题视为分类任务，并要求有大量标注数据集可供使用。

为利用机器学习方法解决阿拉伯语命名实体识别任务，Benajiba 等人（ [7], [21], [29], [50]）在早期开展了大量有价值的工作。他们首先采用最大熵（ME） [29] ，并开发了一个阿拉伯语命名实体识别系统 ANERsys 1.0。该系统使用的特征包括词汇特征、上下文特征和地名录特征。不过，该系统难以识别由多个词组成的实体。随后，他们提出了开发的 ANERsys 2.0 [21]。

为了进一步提升阿拉伯语命名实体识别系统的性能，Benajiba 等人 [50] 考虑了更多特征，例如形态特征、词性标签等。他们在 ACE 语料库和 ANERcorp 上测试了新的命名实体识别系统，并发现使用所有特征时系统取得了最先进的性能。基于条件随机场（CRF）的方法是一种强有力的基线机器学习方法 [51]。Benajiba 等人 [7] 进一步使用 CRF，在最大熵方法的基础上改进 ANERsys。结果表明，使用 CRF 方法后，模型的性能比 ANERsys 2.0 高出近两个百分点。其他机器学习技术，如逻辑回归 [52]和隐马尔可夫模型（ HMM），也广泛应用于阿拉伯语混合命名实体识别系统，与基于规则的系统相比，能够进一步提升性能。

除了单纯依赖机器学习算法之外，一些方法 [28], [44], [53] 提出将基于规则的方法与基于机器学习的方法结合起来，构建混合系统，旨在提升阿拉伯语命名实体识别任务的整体性能。具体而言，基于规则的方法依赖手工制定的语法规则，而基于机器学习的方法则利用机器学习算法，从数据集中提取特征并据此构建命名实体识别系统。随后，混合方法在流水线过程中将基于规则的方法与基于机器学习的方法结合起来，以提升系统的整体性能。例如，Abdallah 等人 [44] 将决策树与规则相结合，用于识别三类命名实体，包括人物、地点和组织，并证明该方法作为混合命名实体识别系统中的分类器具有相对较高的效率。值得注意的是，混合方法会导致复杂性增加，因为它涉及语法规则和手工特征的使用。

尽管机器学习方法取得了不错的性能，但它们无法通过对数线性 HMM 或线性链条件随机场等线性模型，从数据中学习复杂的高层特征。为了弥补机器学习的不足，研究人员提出了深度学习方法，以自动发现隐藏特征。我们将在第 3.3 节介绍深度学习方法。

3.3 深度学习

近年来，深度学习方法被广泛应用于阿拉伯语命名实体识别。与传统的基于手工设计特征的方法相比，深度学习方法能够学习到更稳健的特征，并减少设计手工特征所需的工作。此外，深度学习方法可以通过梯度下降以端到端方式进行训练。因此，我们能够设计出可能更加复杂的阿拉伯语命名实体识别系统。本节将如图 4 所示，从三个方面介绍深度学习方法在阿拉伯语命名实体识别中的发展，即输入表示、上下文编码器和标签解码器。输入表示考虑词级和字符级嵌入，以及对阿拉伯语命名实体识别有效的附加特征。上下文编码器使用卷积神经网络、循环神经网络或其他网络，从文本中捕捉上下文依赖关系。标签解码器旨在为输入序列中的词元预测相应的标签。

为了将深度学习应用于阿拉伯语命名实体识别，Gridach 等人[55] 于2016年首次提出了一种用于阿拉伯语命名实体识别的简单深度神经网络（DNN）。他们基于 Word2vec [56]构建该网络；Word2vec 可将词表示为低维实值稠密向量。在此基础上，他们进一步构建了三个神经网络层：输入层、隐藏层和输出层。在输入层中，他们将上下文明确表示为一个窗口，由中心词与其紧邻邻居串联而成。隐藏层

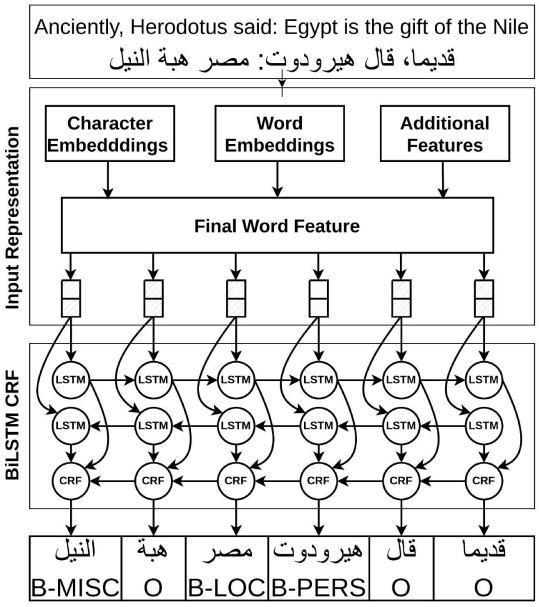


图5：一种具有代表性的阿拉伯语命名实体识别深度学习方法的架构[54]。该模型采用典型的 BiLSTM-CRF架构，并使用多种输入特征。

从窗口中的词表示学习特征组合。最后，输出层转换隐藏表示，并将其分为不同的实体类型。这项工作采用简单的线性层来预测标签序列，而后续研究大多使用条件随机场来确定标签序列。

随后，研究人员将更复杂的上下文编码器应用于阿拉伯语命名实体识别，例如循环神经网络（RNN）。RNN 能有效捕捉每个词元的上下文信息，并在许多不同的自然语言处理任务中取得了显著成功 [57], [58]。在阿拉伯语命名实体识别领域，Mohammed 等人 [59] 采用双向长短期记忆（ BiLSTM）层进行实体识别。BiLSTM 层能够有效利用过去和未来的输入特征 [60],，因此适用于命名实体识别任务。具体而言，他们首先将字符级嵌入层和词嵌入层作为模型输入。随后，他们进一步提出嵌入注意力层，以整合这两类嵌入。整合后的嵌入再输入 BiLSTM 层和最终的分类层。为评估不同循环网络的有效性，Shahina 等人 [61] 探索使用 RNN、LSTM 和 GRU 对阿拉伯语命名实体识别数据进行建模，其中序列中每个词的上下文都与任务相关。

对于阿拉伯语命名实体识别的输入表示，Gridach 等人 [62] 研究了基于常用 BiLSTM-CRF 架构引入字符级特征的有效性。具体而言，他们采用双向长短期记忆编码器对字符进行编码，并将首字符和末字符的表示拼接起来，作为对应词语的最终字符级表示。随后，将该字符表示与词嵌入拼接，作为命名实体识别模型编码器的输入。此外，该论文表明，预训练词嵌入的表现远优于随机初始化的词嵌入，差距很大。最后，Gridach 等人 [62] 在 Twitter命名实体识别数据集 [63] 上评估了所提出的模型，结果表明，引入字符级信息和预训练词嵌入是有效的。

为了编码词级和字符级输入表示，David 等人 [64] 也采用 BiLSTM 对字符级和词级表示的组合进行编码。词嵌入先经过卷积神经网络层，再输入 BiLSTM。Mohammed 等人[65] 为阿拉伯语命名实体识别任务提出了一种高效的多注意力层系统。与先前的工作类似，他们采用字符级嵌入和词嵌入。不同的是，他们在使用 BiLSTM 对嵌入进行编码后，进一步引入了自注意力层，从而提升系统的性能。

除了传统的字符嵌入和词嵌入外，如图 5 所示，Ali 等人 [54] 还从外部知识源中探索了额外的词语特征。其基本架构同样是 BiLSTM-CRF。他们将形态学特征和句法特征添加到不同的词嵌入中，以训练模型。具体而言，他们使用 2019年发布的 Madamira 工具 [35] 生成词性、大小写和词语分析特征。此外，他们还设计了第四种特征，即引号特征。所有特征都表示为独热向量，并与词嵌入和字符嵌入拼接。

除了监督训练之外，Helwel 等人 [66] 还将半监督协同训练方法引入深度学习领域，并将其命名为深度协同学习。具体而言，他们使用少量有标签数据，以及大量来自维基百科的部分标注数据。首先，他们基于完全标注的数据集和不同组的词嵌入训练两个深度学习分类器。这两个模型分别作为处理部分标注数据的不同视角。随后，他们在部分标注的语料库上分别应用每个分类器，并利用每个分类器的输出重新训练另一个分类器，从而迭代地重新训练分类器。

Liu 等人 [68] 提出了一种集成式阿拉伯语命名实体识别系统，结合了深度表示学习、特征工程、序列标注和模型集成。按照常见做法，Liu 等人 [68] 将阿拉伯语命名实体识别视为序列标注问题，并采用 BiLSTM-CRF 架构作为一个基线。为了提升模型性能，Liu 等人 [68] 引入了一种基于词典的方法来增强模型输出。具体而言，他们首先建立一个将表层词映射到相应实体类型的命名实体映射表。当表层词与映射表中的某些实体匹配，且表层词与模型预测之间不存在冲突时，该映射表将在预测层中用于添加额外实体。实验表明，基于词典的方法能够以相对较小的精确率损失提升召回率，从而提高 F1 分数。此外，这项工作还采用了手工特征，将输入词元分为三类：数字、英语词和阿拉伯语词，并将其映射为向量。实验结果表明，在 AQMAR 数据集上，F1 分数显著提高了 4 分。

与基于规则和基于机器学习的命名实体识别方法相比，基于深度学习的阿拉伯语命名实体识别在过去十年中展现出更优越的性能。此外，

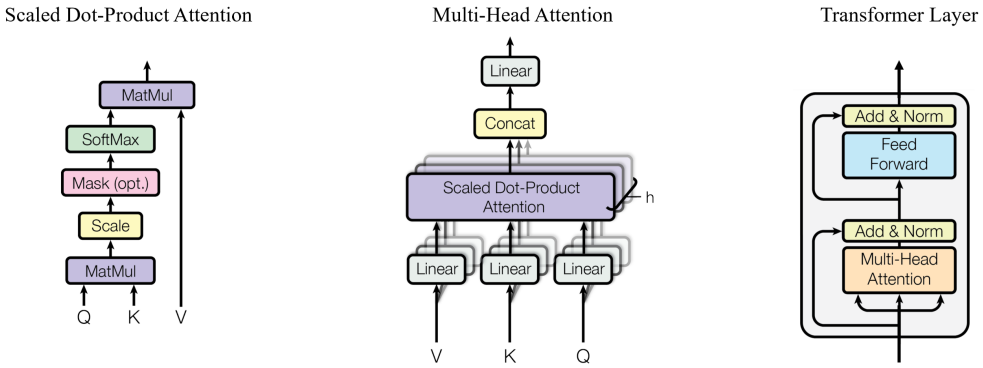


图6：缩放点积注意力（左）、多头注意力（中）和 Transformer 层（右）的概览。该图引自 [67]。

基于深度学习的阿拉伯语命名实体识别不需要手工设计特征，这简化了阿拉伯语命名实体识别的使用。

3.4 预训练语言模型

近年来，预训练语言模型（PLM）展现出卓越的表征能力，能够捕捉句子中的上下文信息，从而在包括命名实体识别（NER）在内的多种自然语言处理任务中取得显著改进。通常，大多数预训练语言模型采用 Transformer 架构作为编码器，该编码器由两层组成，即多头注意力层和一个 2 层前馈神经网络。图 6（右）展示了一个 Transformer 模块的架构，图 6（中）展示了多头注意力的详细结构 [67] ，图 6（左）则描绘了内部的缩放点积注意力。要实现强大的预训练语言模型，最重要的关键要素取决于高效的模型架构和高质量的数据收集。因此，我们将从这两个角度介绍阿拉伯语预训练语言模型及其在阿拉伯语命名实体识别中的应用。

BERT [69] 是最著名的预训练语言模型之一，它采用掩码语言模型（MLM）和下一句预测（NSP）作为学习目标。对于掩码语言模型，输入词元中的一部分会被掩码，目标是恢复掩码位置上的原始词元。下一句预测任务通过预测第二个句子是否为真正的下一句，使 BERT 学习两个句子之间的关系。通过在大规模未标注数据集上进行预训练，BERT 学习了丰富的文本信息，并能够提供有力的上下文信息，从而促进下游任务。然而，之前的大多数预训练语言模型研究都集中于英语，对阿拉伯语的关注较少。本节主要回顾面向阿拉伯语的预训练语言模型研究，展示阿拉伯语预训练语言模型对阿拉伯语命名实体识别的有效性。

Antoun 等人 [26] 预训练了一个名为 AraBERT 的阿拉伯语 BERT，旨在取得与 BERT 在英语中相同的成功。与 BERT 一样，Antoun 等人 [26] 也采用 MLM 和 NSP 作为训练目标。为了适应阿拉伯语的特点，作者提出先使用离线分词器将单词切分为词干、前缀和后缀。该预训练模型在三项阿拉伯语理解任务上进行了评估，包括情感分析、命名实体识别和问答。

Antoun 等人 [27] 构建了 AraELECTRA，其模型架构与 ELECTRA [70]相同。整个模型包含两个神经网络模块：1）生成器，以被破坏的句子为输入，并学习预测被掩码的原始词元；2）判别器，以预测后的句子为输入，并尝试判断哪些词元被掩码、哪些是原始词元。在微调阶段，输入句子会被送入判别器，并额外添加一个线性层来执行下游任务。他们使用 ANERCorp 数据集对 ArabicELECTRA 模型进行了评估，结果表明该模型表现优越。

除了训练目标外，轻量级预训练模型因推理速度更快，也日益受到关注。继英语 ALBERT 之后，Safaya 等人 [71]，使用阿拉伯语文本预训练了 Arabic ALBERT。BERT 与 ALBERT 的主要区别在于，ALBERT 在不同层之间共享相同的参数，从而大幅缩减模型参数量。考虑到阿拉伯语的语言特点，Ghaddar 等人 [73] 提出了 Char-JABER，通过在输入层加入字符级表示，增强了 BERT 风格的阿拉伯语预训练语言模型。

除了编码器式的 AraBERT，还有其他类型的阿拉伯语预训练语言模型，例如解码器式的 AraGPT2[74] 和编码器-解码器式的 AraT5 [75]。这些工作有望推动阿拉伯语自然语言处理的发展。然而，由于这些模型尚未在阿拉伯语命名实体识别任务上进行评估，我们将不再详细介绍。

基于上述阿拉伯语预训练语言模型，Norah 等人 [76] 提出了首项针对古典阿拉伯语命名实体识别的研究。他们提出在阿拉伯语 BERT 上微调 BiLSTM/双向门控循环单元-条件随机场模型，并与两个简单的基线模型进行比较，即：1）在 BERT 顶部添加带有 Softmax 函数的线性层（BERT）；2）在 BERT 顶部添加 CRF 层（BERT-CRF）。他们在经典的阿拉伯语命名实体识别数据集 CANERCorpus 上进行了实验。实验结果表明，两个简单的基线模型（BERT、BERT-CRF）已取得良好表现，F1 分数分别为 94.45 和 94.68。BERT-双向门控循环单元-条件随机场模型取得了最佳结果，F1 分数为 94.76。

为进一步探究不同预训练语言模型架构的有效性，Al-Qurishi 等人 [77] 提出了一种用于阿拉伯语命名实体识别的、基于 Transformer 的条件随机场模型，并在三个 BERT 变体上进行了微调，即 AraBERT、AraELECTRA 和 XLM-

Roberta。实验在 ANERCorp 和 AQMAR 数据集上进行，结果表明，在 AraBERT 上进行微调的效果最佳。

除了用于阿拉伯语命名实体识别的全监督方法外，Helwe等人 [78] 提出了一种基于师生学习机制的半监督框架，称为深度协同学习。其阿拉伯语命名实体识别的基础模型采用AraBERT。在训练过程中，他们首先使用已标注数据集训练教师模型。随后，利用训练好的教师模型对部分标注数据集进行预测，生成伪标签。最后，使用生成的伪数据集训练学生模型。然而，与另外两个数据集（即AQMAR和NEWS数据集 [79]）相比，该模型在TWEETS数据集 [79]上的表现较差。这可能是因为AraBERT主要基于现代标准阿拉伯语语料库训练，而TWEETS数据集大多属于埃及方言。这一现象凸显了阿拉伯语命名实体识别的另一个重要组成部分，即高质量的数据收集。

为将阿拉伯语预训练语言模型扩展到多样化的数据领域和不同的阿拉伯语方言，Abdul-Mageed等人 [80]从多个来源收集训练数据，即书籍、新闻文章、Common Crawl数据和维基百科，并训练了阿拉伯语预训练语言模型MARBERT。为了促进对阿拉伯语方言的理解，他们还使用更大的Twitter数据集对模型进行训练。该Twitter数据集来自一个包含60亿条推文的大型内部数据集，共产生了128GB的文本。在训练过程中，由于推文通常较短，该工作移除了下一句预测任务。在五个命名实体识别数据集上进行了广泛实验，即ACE03NW、ACE03BN、ACE04NW、AN-ERcorp和TW-NER。我们注意到，MARBERT在TW-NER上达到了新的最先进水平，验证了使用推文进行预训练的有效性。

除常用的新闻领域命名实体识别数据集外，还有一些工作尝试推动阿拉伯语命名实体识别在生物医学领域的发展。由于显著的领域差异，生物医学阿拉伯语命名实体识别更具挑战性。因此，典型的阿拉伯语预训练语言模型（如AraBERT）通常表现不佳。为解决这一问题，Boudjellal等人 [81]通过将生物医学知识注入AraBERT预训练模型，提出了一种阿拉伯语生物医学预训练模型。他们从医学期刊中人工收集了一组文献语料库，例如PubMed和MedlinePlusHealthInformation。所收集的语料库包含约50万个词，与用于训练AraBERT的数据相比要小得多。尽管如此，实验结果表明，训练得到的生物医学模型在生物医学命名实体识别数据上的表现优于AraBERT。这一结果还证明了针对相应领域特定任务训练领域特定预训练模型的价值。

迄今为止，随着预训练语言模型的兴起，阿拉伯语命名实体识别已取得显著进展。一方面，现有的阿拉伯语预训练语言模型为阿拉伯语文本提供了强大的表征，有效提升了阿拉伯语命名实体识别性能。另一方面，阿拉伯语命名实体识别仍迫切需要更多面向相应特定领域的阿拉伯语预训练语言模型。

3.5 讨论

在回顾阿拉伯语命名实体识别的四种范式后，我们按照发表日期的先后顺序整理了阿拉伯语命名实体识别数据集上的实验结果，以便进一步比较。如表3所示，我们总结了ANER-Corp 上的实验结果。由于不同方法采用的数据划分策略存在差异，因此很难对它们进行公平比较。换言之，阿拉伯语命名实体识别尚无标准排行榜，而 ANERcorp 已有训练集、开发集和测试集。

从表中可以看出，大多数传统方法，包括基于规则的方法和机器学习方法，其表现都显著逊于深度学习方法。这些方法通常依赖专门设计的规则或N元语法，因此泛化能力较弱。深度学习方法主要使用长短期记忆网络进行表示编码，并采用条件随机场进行标签解码。与传统循环神经网络相比，长短期记忆网络能够更有效地编码句子信息。同时，条件随机场的表现比简单的线性层更加稳健。对于输入表示，常见做法是使用卷积神经网络对字符表示进行编码，然后将字符级特征与词级特征结合起来。这一策略对于编码阿拉伯语文本十分重要，因为字符表示能够提供形态和句法信息。使用预训练语言模型后，模型性能实现了显著提升。如表3所示，对比AraELECTRA、MARBERT、Arabic-BERT、ARBERT和JABER等不同预训练模型可以发现，使用线性解码器时，JABER取得了最佳结果84.20。我们认为，由于JABER使用简单的线性层进行标签解码，其性能还有进一步提升的空间。不过，值得注意的是，这些预训练模型使用不同的训练语料库进行训练。换言之，在实际应用中，其效果取决于下游场景，尤其是在面对阿拉伯语方言时。

3.6 阿拉伯语命名实体识别的应用技术

前几节概述了阿拉伯语命名实体识别的典型网络架构。本节综述近期应用于阿拉伯语命名实体识别的技术及其探索进展。

3.6.1 用于阿拉伯语命名实体识别的深度多任务学习

多任务学习 [89], [90] 已广泛应用于自然语言处理中的许多任务 [91], , 包括命名实体识别。在多任务学习中，我们通常选择若干相似任务，并为它们采用不同的模型，同时部分共享这些模型的隐藏层。这样，模型能够提供适用于所有任务的更优表示，并避免过拟合。

对于阿拉伯语中的多任务学习，Jarrar等人 [92]将嵌套式阿拉伯语命名实体识别转换为多任务范式，并提出一种多任务模型，用于在Wojood上预测实体。Wojood是一个大规模阿拉伯语命名实体识别数据集，收集自包括网页文章、档案和社交媒体在内的多个来源，涵盖现代标准阿拉伯语和方言。具体而言，他们采用AraBert [26] 作为共享编码器，并为每种实体类型设计不同的线性层，用于预测相应的BIO标签。通过这种方式，该模型可以

表3：ANERcorp上阿拉伯语命名实体识别系统的比较。结果针对四类实体进行报告，包括人物、地点、组织和其他。“标准”数据集划分意味着这些方法采用官方数据集划分，而不是自行划分数据集。由于传统的基于规则和机器学习方法没有专门的上下文编码器，因此我们合并“输入表示”和“上下文编码器”两列。表中的“-”符号表示原论文未描述相应项目。

Work	Year	输入表示与上下文编码器	标签解码器	数据集划分	性能（F1 分数）
[29]	2007	N元语法	最大熵	标准	54.11
[7]	2008	-	CRF	标准	65.91
[30]	2010	N元语法	CRF	4:0:1	人物：81，地点：88，组织：73
[53]	2012	-	支持向量机-隐马尔可夫模型	标准	人物：94，地点：96.9，组织：80.66，其他：60.82
[31]	2012	-	Rule	标准	67.13
[32]	2013	-	CRF	3:0:1	68.05
[82]	2014	-	规则，机器学习	标准	人物：94，地点：90，组织：88
[83]	2015	N元语法	HMM	-	人物：79，地点：78，组织：67
[55]	2016	Word2vec，深度神经网络	线性	标准	88.64
[84]	2018	-	Rule	标准	人物：78，地点：67，组织：63
[59]	2018	Word2vec，基于CNN的字符模型	-	8:1:1	87.12
[85]	2019	Word2vec，基于CNN的字符模型，卷积神经网络，长短期记忆网络	CRF	8:1:1	88.77
[86]	2019	-	Rule	标准	人物：83，地点：92，组织：89
[14]	2019	双向长短期记忆网络	CRF	-	90.6
[26]	2020	AraBERT v1	线性	-	81.9
[26]	2020	AraBERT v0.1	线性	标准	84.2
[87]	2020	AraBERT v0.1，双向门控循环单元	线性	8:1:1	90.51
[38]	2020	-	-	5:0:1	83
[88]	2021	-	-	5:0:1	81.39
[27]	2021	AraELECTRA	线性	5:0:1	83.95
[77]	2021	AraBertv2	CRF	5:0:1	89.56
[77]	2021	AraELECTRA	CRF	5:0:1	87.21
[73]	2022	MARBERT	线性	5:0:1	80.5
[73]	2022	Arabic-BERT	线性	5:0:1	82.05
[73]	2022	ARBERT	线性	5:0:1	84.03
[73]	2022	JABER	线性	5:0:1	84.20

独立提取不同类型的实体。然而，它们仍无法解决同一类型中的嵌套实体问题。最终，该方法在Wojood上取得了88.40的总体F1值。

此外，艾哈迈德等人 [93]将阿拉伯语命名实体识别任务与主题建模任务相结合，并据此提出了Tafsir数据集，这是首个面向古典阿拉伯语的大规模多任务基准。这两项任务共享相同的嵌入表示。随后，他们在上层命名实体识别任务中采用BiLSTM-CRF模型，同时为主题建模任务的每个标签使用线性层。实验表明，所提出的方法在两项任务中均表现有效。

3.6.2 面向命名实体识别的迁移学习

低资源语言通常面临训练数据匮乏的问题，而迁移学习 [94], [95]能够有效缓解这一问题。尤其是，基于预训练模型的迁移学习能够惠及许多自然语言处理任务，因为知识可以跨任务或领域迁移。

对于阿拉伯语命名实体识别，为了利用多语言预训练模型的知识迁移能力，Al-Smadi 等人 [96] 提出了一种基于多语言通用句子编码器 [97] (MUSE)的 Pooled-GRU网络，如图7所示。具体而言，他们采用传统的序列标注模型和预训练的多语言通用句子编码器来执行阿拉伯语命名实体识别。通过这种方式，MUSE中的知识得以有效迁移至Pooled-GPU。得益于迁移而来的知识，其他语言场景中句子和短语之间的语义关系也能应用于阿拉伯语。实验使用阿拉伯语数据集WikiFANE进行

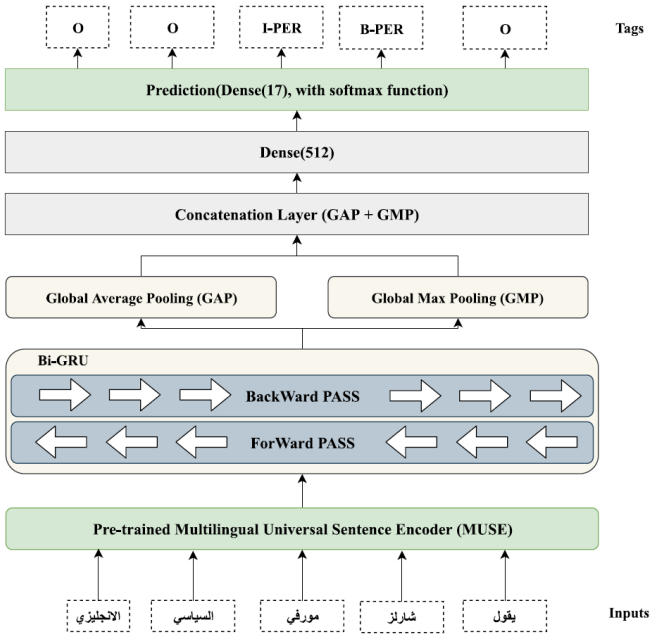


图7：**Pooled-GRU模型的架构 [96]**。

[98] 而且，该模型相比BiLSTM-CRF模型有显著提升。

除了上述多语言 MUSE 外，还有其他专为跨语言任务设计的多语言 Transformer。Lan 等人 [99] 提出了 GigaBERT，这是一种双语 Bert 模型，旨在用于阿拉伯语自然语言处理任务以及英语到

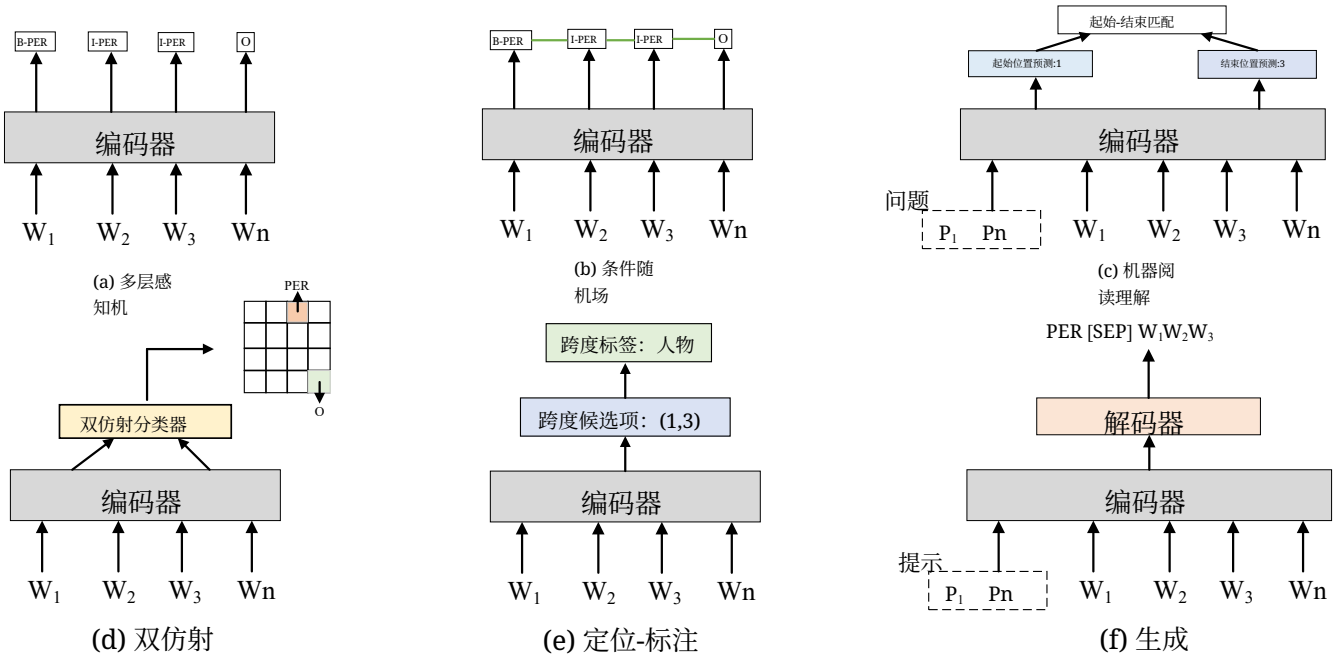


图8：不同命名实体识别解决方案范式的架构。(a)和(b)应用于以往的阿拉伯语命名实体识别方法。(c)-(f)是近期英语命名实体识别中具有代表性的解码策略。

阿拉伯语的迁移。该定制 Bert 模型在 Giga- word、维基百科和一些网页爬取数据上进行预训练。具体而言，作者定制了词汇表，并创建了代码切换数据，以进一步增强跨语言迁移能力。借此，该模型在包括命名实体识别在内的许多自然语言处理任务中，表现优于常见的多语言 BERT 模型。此外，使用阿拉伯语监督数据时，GigaBERT 的表现优于最先进的阿拉伯语专用 BERT 模型。

4 未来方向

随着建模方法的发展以及下游应用的不断增加，我们期待阿拉伯语命名实体识别在未来取得更多进展。因此，我们从以下五个方面列出阿拉伯语命名实体识别未来可能的发展方向。1）过去几年，英语或中文的命名实体识别方法取得了显著进展。这些进展主要集中在设计不同的标签解码策略，以适应多样的标签结构。这些重要的技术突破可以应用于阿拉伯语命名实体识别，但尚未得到充分探索。2）除了创新技术外，还需要根据下游应用的需求，进一步研究特定的阿拉伯语命名实体识别任务，例如嵌套命名实体识别。嵌套命名实体识别在英语领域已有深入研究，但只有少数阿拉伯语命名实体识别研究考虑了这一任务。3）阿拉伯语命名实体识别可用的标注数据较少，因此也需要低资源或远程监督方法来缓解数据匮乏的问题。这样一来，即使只有少量带标签的标注数据，阿拉伯语命名实体识别也能取得更稳健的表现。4）除了利用更多数据来构建更稳健的阿拉伯语命名实体识别模型，跨语言知识迁移也能为模型学习提供全面的信息。为此，我们将分析哪些跨语言知识迁移技术可能有助于阿拉伯语命名实体识别。5）通过将命名实体识别与其他辅助任务联合训练，两个任务都能相互受益。因此，我们还讨论了若干与阿拉伯语命名实体识别相关的任务。我们将在以下各节介绍这些主题。

可选标签解码策略。 如第3.3节所述，几乎所有阿拉伯语命名实体识别方法都采用线性分类或 CRF 层进行标签解码。如图8(a)和图8(b)所示，线性层或 CRF 为每个词元解码标签。尽管这两种解码策略在平坦命名实体识别任务中取得了不错的性能，但无法处理更复杂的命名实体识别任务，例如嵌套命名实体识别。与这两种范式不同，实际上，在英语或中文命名实体识别中，如图8(c)-(f)所示，存在许多不同的命名实体识别标签解码策略。这些新提出的框架能够统一不同形式的命名实体识别任务。下面将简要介绍这些具有代表性的解决方案：

- **机器阅读理解**[6] 如图8(c)所示，Li 等人 [100] 通过在句子前插入一个问题，将命名实体识别任务转化为机器阅读理解范式。使用机器阅读理解框架的本质，是通过设计与命名实体识别标签相关的问题，引导模型从给定文本中寻找最相关的答案。在解码过程中，他们确定问题所提及实体类型对应的起始位置和结束位置。

- **双仿射**[7] 在图8(d)中，Yu 等人 [101] 将标签解码过程重新定义为依存句法分析任务，对句子中的所有起始位置和结束位置对进行联合考虑，并将其分类为特定类型。双仿射架构非常有效，许多

6. https://github.com/ShannonAI/mrc-for-flat-nested-ner
7. https://github.com/juntaoy/biaffine-ner

近期研究都采用该架构来进一步提升性能 [102], [103]。

• **定位-标注**[8] 如图8(e)所示，与此前的单阶段范式不同，Shen 等人 [104]提出了一种两阶段方案：首先定位句子中可能的实体，然后识别这些片段的具体类别。该框架可以有效利用实体的边界信息。Xia 等人 [105]也采用类似的方法处理嵌套命名实体识别。

• **生成**[9] 除了上述基于编码器的命名实体识别范式外，近期基于编码器-解码器的方法也越来越受到关注。如图8(f)所示，Lu 等人 [106] 基于 BART 等生成式预训练语言模型，将命名实体识别任务转化为生成任务。通过这种方式，模型无需为每个词元预测实体类别，而是直接在解码后的输出句子中生成实体。除了生成实体词语外，还有一些生成式方法 [109], [110] 会生成实体在原始文本中的索引。凭借出色的生成能力，这些方法适用于各种命名实体识别任务，包括平面、嵌套和不连续命名实体识别。值得注意的是，这种范式的性能可能会受到用户构造的硬提示的显著影响。因此，如何设计恰当的提示也是阿拉伯语命名实体识别面临的一项挑战。

特定命名实体识别任务。 如方法部分所述，现有的大多数阿拉伯语命名实体识别方法只考虑通用领域中的粗粒度命名实体识别类型，并专注于平坦命名实体识别。实际上，在槽位填充[111]–[113] 等真实应用场景（如对话）中，实体具有细粒度和层次化的特点，这意味着一个命名实体可能属于多种类型。另一项挑战是实体可能嵌套，其中一个实体的一部分对应另一种实体类型。此外，文本中的实体也可能是不连续的，在临床语料库中尤其如此。在这种情况下，研究人员需要超越连续实体，提出抽取不连续实体的新方法。因此，我们可以借鉴英语命名实体识别的主流方法来应对阿拉伯语命名实体识别中的这些挑战。下面简要介绍几种解决这些挑战的方法。

• **细粒度命名实体识别**。为了探索需要识别数百种命名实体类别的细粒度命名实体识别，Mai 等人 [114] 利用词典和类别嵌入，并将其融入神经网络。Peng 等人 [115] 提出利用先前标注的数据以及新引入实体类型的实体词典，对现有命名实体识别系统进行调整。通过这种方式，他们能够快速适应细粒度实体。

• **嵌套命名实体识别**。 如表4所示，尽管在平坦命名实体识别数据集中，采用条件随机场作为解码器的阿拉伯语命名实体识别方法比线性方法取得了更好的性能，但条件随机场架构难以应用于嵌套命名实体识别。为了解决嵌套命名实体识别问题，我们可以采用上述针对命名实体识别提出的替代方案，例如机器阅读理解、双仿射、定位-标注或生成。还有一些基于图的方法，用于构建实体-实体图 [116]，或者将嵌套命名实体识别建模为潜在词汇化成分句法分析 [117]。借助这些方法，我们可以更便捷地解决阿拉伯语嵌套命名实体识别问题。

• **不连续命名实体识别**。传统命名实体识别被视为序列标注任务，难以处理这些实体。Li 等人 [118] 提出了一种基于跨度的模型，并通过关系分类判断给定的一对实体片段是重叠还是相邻。Wang 等人 [119] 为每个句子构建一个片段图，其中每个节点表示一个片段，边连接属于同一实体的两个节点。因此，他们能够识别不连续实体。此外，还有一些针对该任务提出的基于转移的方法[120] 或网格标注方法 [121] 。此外，基于生成的模型可以直接输出不连续实体。

低资源与远程监督命名实体识别。事实上，阿拉伯语命名实体识别的开源标注数据集和可用的阿拉伯语实体词典都极为稀缺。标注大量命名实体识别数据既耗时又昂贵。这一困境导致阿拉伯语命名实体识别在实际应用中的表现不尽如人意。为了充分利用有限的有标签数据，可以考虑采用低资源和远程监督方法，以取得更好的结果。下面介绍几种在英语中行之有效、可能适用于阿拉伯语命名实体识别的方法。

• **低资源命名实体识别**。 许多研究利用数据增强来提升低资源场景下的性能。Dai 等人 [122]采用标签级词元替换、同义词替换和提及替换来获取新的训练样本。Ding 等人 [123] 对线性化标注句子进行建模，以生成新的样例。Liu 等人 [124]提出了一种增强方法，可从 BERT 中提取知识，且无需大量人工干预。通过增强方法，我们可以获得更多训练样例。除了生成更多训练样本外，近期还有一些方法 [125], [126] 专注于设计低资源算法。这些方法有效利用大型预训练语言模型中的知识，并取得了显著成果。

•**远程监督命名实体识别**。 借助阿拉伯语词典或现有的命名实体识别工具，我们可以通过匹配或标注未标注数据，轻松获取大量弱标注样本。然而，这些弱标签含有噪声。为全面利用这些噪声样本，Yang 等人 [127]提出了部分标注学习范式。最近， [128] 引入正例-未标注学习，用于标注正例词元，而非直接学习弱标签。教师-学生框架 [129]–[132]被广泛用于优化噪声标签并取得更优性能。为联合利用少量有标签数据和大量弱标注数据进行学习，Jiang 等人 [133] 提出了一个多阶段计算框架，以实现稳健学习。

跨语言知识迁移。事实证明，将知识从高资源语言迁移到低资源语言能够有效缓解数据稀缺问题。为了增强阿拉伯语命名实体识别系统的能力，可以考虑从其他

8. https://github.com/tricktreat/locate-and-label
9. https://github.com/universal-ie/UIE

语言进行知识迁移。为了实现跨语言知识迁移，本文介绍了几种有效的解决方案。

•**机器翻译**。 目前，强大的机器翻译系统支持 100 多种语言。机器翻译是将其他语言的训练句子和实体译为目标语言的一种直接方式。因此，可以利用机器翻译提升跨语言命名实体识别的性能 [122], [134], [135]。

• **对抗训练**。微调期间的语言对抗任务可以显著提升跨语言迁移性能 [136]。换言之，对抗技术促使模型对齐英文文档及其目标语言译文的表示。此外，对抗学习框架可以利用未标注数据并提升结果 [137]，：判别器根据与源语言的相似度，挑选出语言依赖性较低的目标语言数据。

• **多语言预训练语言模型**。 多语言预训练语言模型（如 mBERT、XLM [138], mT5 [139], ）提供了一种通过跨语言迁移实现零样本学习的直接途径，从而无需为目标任务和目标语言提供有标签数据 [140]。对于预训练语言而言，跨语言迁移尤其高效 [141]。

将 NER 与其他任务联合训练。有许多任务在语义上与 NER 相关。探索联合执行 NER 和其他任务的方法是有价值的。下面我们介绍与 NER 关系最密切的任务。

• **NER 与实体链接**。大多数实体链接系统会忽略提及检测，仅对预先检测到的提及执行实体消歧。因此，实体链接与 NER 之间的依赖关系被忽略了。在实际场景中，我们可以对 NER 和 EL 进行联合学习，从而在每个决策环节利用这两项任务的信息 [142]–[144]。此外，通过在用于 NER 和实体链接的表示计算之间建立信息流，我们能够改进模型。

•**NER 与关系抽取**。NER 和关系抽取都旨在从非结构化文本中提取结构化信息。典型的流水线方法[145], [146] 是先识别实体提及，然后对任意两个提及进行分类，以抽取它们之间的关系。然而，这种流水线会导致错误传播问题。为此，一种替代性且较新的方法是联合执行这两项任务 [147], [148]。这样，我们便可以利用任务之间的交互，从而提升性能。

阿拉伯方言命名实体识别。 在方法章节中，大多数方法都侧重于现代标准阿拉伯语（MSA）。然而，阿拉伯语文本中存在大量方言。尽管现代标准阿拉伯语与阿拉伯方言（DA）具有一些相似特征，但这两种语言变体之间的显著差异，妨碍了现代标准阿拉伯语系统解决阿拉伯方言命名实体识别问题。因此，提升阿拉伯方言命名实体识别的性能是一个必要且前景广阔的方向。本文介绍几种有效的阿拉伯方言命名实体识别方法。

• **方言预训练语言模型**。 有若干种利用方言进行训练的阿拉伯语预训练语言模型。它们的训练语料库通常包含由多种方言构成的阿拉伯语推文。具体而言，MARBERT [10] 从一个约含 60 亿条推文的大型内部数据集中随机抽取 10 亿条阿拉伯语推文。AraBERTv0.2-Twitter [11] 使用从 1 亿条推文中筛选出的 6000 万条阿拉伯语推文进行训练。此外，还有专为阿尔及利亚方言打造的预训练模型 DziriBERT[149]。针对下游任务，继续在特定方言上进行预训练是可行的。

• **方言无监督领域自适应**。无监督领域自适应（UDA）可以作为跨方言学习的首选方案。Mekki等人 [150]提出了一个用于阿拉伯语多方言序列标注的无监督领域自适应框架 AdaSL，该框架利用无标注领域自适应数据和有标注现代标准阿拉伯语数据，在阿拉伯语多方言序列标注任务上取得了新的零样本性能最优结果。因此，通过领域自适应利用现代标准阿拉伯语处理阿拉伯语方言，是一个值得探索的方向。

5 结论

阿拉伯语命名实体识别的发展明显落后于英语命名实体识别。本综述旨在全面回顾阿拉伯语命名实体识别，帮助研究人员了解这一领域。具体而言，我们介绍了阿拉伯语命名实体识别的背景，尤其是其相较于英语命名实体识别所面临的独特挑战。随后，我们总结了当前阿拉伯语命名实体识别的资源，包括数据集和常用处理工具。借助这些资源，研究人员可以轻松开启探索，并快速了解阿拉伯语自然语言处理。为梳理阿拉伯语命名实体识别方法的发展，我们将其仔细归纳为四种范式，即基于规则的方法、机器学习方法、深度学习方法和预训练语言模型方法。针对这四种范式，我们介绍了最具代表性的方法，并描述了其架构。此外，我们还总结了阿拉伯语命名实体识别的应用，这些应用推动了下游任务的改进。最后，我们指出了阿拉伯语命名实体识别的若干潜在发展方向。我们希望本综述能为阿拉伯语命名实体识别模型的设计提供实用指导，并吸引更多研究人员投身这一领域。

参考文献

[1] I. Guellil、H. Saˆadane、F. Azouaou、B. Gueni 和 D. Nouvel，《阿拉伯语自然语言处理：概述》，《沙特国王大学学报：计算机与信息科学》，第33卷，第5期，497–507页，2021年。[2] Q. Cheng、J. Liu、X. Qu、J. Zhao、J. Liang、Z. Wang、B. Huai、N. J. Yuan 和 Y. Xiao，《Hacred：面向实际应用中疑难案例的大规模关系抽取数据集》，载于计算语言学协会论文集：*ACL-IJCNLP 2021*，2021年，2819–2831页。[3] Y. Gu、X. Qu、Z. Wang、B. Huai、N. J. Yuan 和 X. Gui，《阅读、回顾、选择：一种用于短文本实体链接的机器阅读理解框架》，载于*AAAI*人工智能会议论文集，第35卷，第14期，2021年，第12 920–12928页。

10. https://github.com/UBC-NLP/marbert
11. https://github.com/aub-mind/arabert

[4] T. Zhu、X. Qu、W. Chen、Z. Wang、B. Huai、N. J. Yuan 和 M. Zhang，《通过伪触发词感知剪枝完全图高效进行文档级事件抽取》，arXiv预印本 arXiv:2112.06013，2021年。[5] K. Clark 和 C. D. Manning，《通过学习实体级分布式表示改进共指消解》，载于计算语言学协会第54届年会论文集（第1卷：长篇论文），2016年，第643–653页。[6] A. Ugawa、A. Tamura、T. Ninomiya、H. Takamura 和 M. Oku- mura，《融入命名实体的神经机器翻译》，载于第27届国际计算语言学会议论文集，2018年，第3240–3250页。[7] Y. Benajiba 和 P. Rosso，《使用条件随机场进行阿拉伯语命名实体识别》，载于阿拉伯世界人类语言技术与自然语言处理研讨会论文集，LREC，第8卷，2008年，第143–153页。[8] A. Elgibali，《阿拉伯语研究：分析与学习中的当前参数》。Brill出版社，2005年，第42卷。[9] K. Shaalan 和 H. Raza，《Nera：阿拉伯语命名实体识别》，《美国信息科学与技术学会期刊》，第60卷，第8期，第1652–1663页，2009年。[10] K. Shaalan，《阿拉伯语命名实体识别与分类综述》，《计算语言学》，第40卷，第2期，第469–510页，2014年。[11] A. Zirikly 和 M. Diab，《阿拉伯语社交媒体中的命名实体识别》，载于第一届自然语言处理向量空间建模研讨会论文集，2015年，第176–185页。[12] A. Dandashi、J. A. Jaam 和 S. Foufou，《阿拉伯语命名实体识别——综述与分析》，载于《智能交互多媒体系统与服务2016》。施普林格，2016年，第83–96页。[13] R. E. Salah 和 L. Q. binti Zakaria，《阿拉伯语命名实体识别机器学习方法比较性综述》，《国际先进科学、工程与信息技术期刊》，第7卷，第2期，第511–518页，2017年。[14] I. El Bazi 和 N. Laachfoubi，《使用深度学习方法进行阿拉伯语命名实体识别》，《国际电气与计算机工程期刊》（2088-8708），第9卷，第3期，2019年。[15] L. Liu、J. Shang 和 J. Han，《阿拉伯语命名实体识别：哪些方法有效，未来方向何在》，载于第四届阿拉伯语自然语言处理研讨会论文集，2019年，第60–67页。[16] B. A. B. Ali、S. Mihi、I. El Bazi 和 N. Laachfoubi，《阿拉伯语社交媒体命名实体识别最新综述》，《人工智能评论》，第34卷，第2期，第125–135页，2020年。[17] J. Li、A. Sun、J. Han 和 C. Li，《深度学习命名实体识别综述》，《IEEE知识与数据工程汇刊》，第34卷，第1期，第50–70页，2020年。[18] V. Yadav 和 S. Bethard，《基于深度学习模型的命名实体识别最新进展综述》，载于第27届国际计算语言学会议论文集，2018年，第2145–2158页。[19] S. Algahtani，《阿拉伯语命名实体识别：基于语料库的研究》。曼彻斯特大学（英国），2012年。[20] S. AbdelRahman、M. Elarnaoty、M. Magdy 和 A. Fahmy，《用于阿拉伯语命名实体识别的集成机器学习技术》，IJCSI，第7卷，第4期，第27–36页，2010年。[21] Y. Benajiba 和 P. Rosso，《Anersys 2.0：结合最大熵与词性标注信息，攻克阿拉伯语命名实体识别任务》，载于IICAI，2007年，第1814–1823页。[22] K. Shaalan 和 H. Raza，《阿拉伯语人名实体识别》，载于2007年计算闪米特语族语言方法研讨会论文集：共同问题与资源，2007年，第17–24页。[23] R. Steinberger，《简化高度多语种文本挖掘应用开发的方法综述》，《语言资源与评测》，第46卷，第2期，第155–176页，2012年。[24] Alkharashi 和 I brahim，《阿拉伯语人名实体生成与识别》，载于第二届阿拉伯语资源与工具国际会议论文集，2009年，第205–208页。[25] Y. Benajiba、M. Diab 和 P. Rosso，《阿拉伯语命名实体识别：一项特征驱动研究》，《IEEE音频、语音与语言处理汇刊》，第17卷，第5期，第926–934页，2009年。[26] W. Antoun、F. Baly 和 H. Hajj，《AraBERT：基于Transformer的阿拉伯语理解模型》，载于第四届开源阿拉伯语语料库与处理工具研讨会（含冒犯性语言检测共享任务）论文集，2020年，第9–15页。[27] ——，《AraELECTRA：面向阿拉伯语理解的文本判别器预训练》，载于第六届阿拉伯语自然语言处理研讨会论文集，2021年，第191–195页。

[28] M. Oudah 和 K. Shaalan，“采用混合方法的阿拉伯语命名实体识别流水线”，载于 COLING 2012 论文集，2012年，第 2159–2176 页。[29] Y. Benajiba、P. Rosso 和 J. M. Bened´ıruiz，“Anersys：一种基于最大熵的阿拉伯语命名实体识别系统”，载于智能文本处理与计算语言学国际会议。施普林格，2007年，第 143–153 页。[30] A. Abdul-Hamid 和 K. Darwish，“用于阿拉伯语命名实体识别的简化特征集”，载于2010年命名实体研讨会论文集，2010年，第 110–115 页。[31] W. Zaghouani，“Renar：基于规则的阿拉伯语命名实体识别系统”，《ACM亚洲语言信息处理汇刊》（TALIP），第 11 卷，第 1 期，第 1–13 页，2012年。[32] A. Morsi 和 A. Rafea，“研究不同特征对基于条件随机场的阿拉伯语命名实体识别性能的影响”，载于2013年ACS国际计算机系统与应用会议（AICCSA）。IEEE，2013年，第 1–5 页。[33] B. Mohit、N. Schneider、R. Bhowmick、K. Oflazer 和 N. A. Smith，“面向召回率的阿拉伯语维基百科命名实体学习”，载于欧洲计算语言学协会第13届会议论文集，2012年，第 162–173 页。[34] R. E. Salah 和 L. Q. B. Zakaria，“构建古典阿拉伯语命名实体识别语料库（CANERCorpus）”，载于2018年第四届信息检索与知识管理国际会议（CAMP）。IEEE，2018年，第 1–8 页。[35] A. Pasha、M. Al-Badrashiny、M. Diab、A. El Kholy、R. Eskander、N. Habash、M. Pooleery、O. Rambow 和 R. Roth，“MADAMIRA：一种快速、全面的阿拉伯语形态分析与消歧工具”，载于第九届语言资源与评测国际会议论文集（LREC’14），2014年，第 1094–1101 页。[36] W. Monroe、S. Green 和 C. D. Manning，“结合领域自适应的非正式阿拉伯语分词”，载于计算语言学协会第52届年会论文集（第2卷：短论文）。马里兰州巴尔的摩：计算语言学协会，2014年6月，第 206–211 页。 [在线]。可访问：https://aclanthology.org/P

14 2034 - A. Abdelali、K. [37] Darwish、N. Durrani 和 H. Mubarak，“Farasa：快速高效的阿拉伯语分词器”，载于计算语言学协会北美分会2016年会议论文集：演示。加利福尼亚州圣迭戈：计算语言学协会，2016年6月，第 11–16 页。

[ ] 16 3003 在线。可 [38] 访问：https://aclanthology.org/N - O. Obeid、N. Zalmout、S. Khalifa、D. Taji、M. Oudah、B. Alhafni、G. Inoue、F. Eryani、A. Erdmann 和 N. Habash，“Camel Tools：用于阿拉伯语自然语言处理的开源 Python 工具包”，载于第12届语言资源与评测会议论文集，2020年，第 7022–7032 页。[39] R. Grishman 和 B. Sundheim，“消息理解会议-6：简史”，载于 COLING 1996 第1卷：第16届国际计算语言学会议，1996年。

[ ] 96 1079 在线。可 [40] 访问：https://aclanthology.org/C - S. Abuleil 和 M. Evens，“为问答系统从阿拉伯语文本中抽取人名”，载于 RIAO，2004年，第 638–647 页。[41] S. Mesfar，“利用句法语法进行阿拉伯语命名实体识别”，载于《自然语言处理与信息系统：第12届自然语言在信息系统中的应用国际会议（NLDB 2007）》，法国巴黎，2007年6月27–29日。论文集 12。施普林格，2007年，第 305–316 页。[42] K. Shaalan 和 H. Raza，“从多种文本类型中识别阿拉伯语命名实体”，载于自然语言处理国际会议。施普林格，2008年，第 440–451 页。[43] A. Elsebai 和 F. Meziane，“从阿拉伯语报纸中抽取人名”，载于2011年信息技术创新国际会议。IEEE，2011年，第 87–89 页。[44] S. Abdallah、K. Shaalan 和 M. Shoaib，“将基于规则的系统与分类相结合以进行阿拉伯语命名实体识别”，载于智能文本处理与计算语言学国际会议。施普林格，2012年，第 311–322 页。[45] J. Maloney 和 M. Niv，“Tagarab：一种利用高精度形态分析的快速、准确的阿拉伯语姓名识别器”，载于闪米特语族语言计算方法，1998年。[46] H. Traboulsi，“阿拉伯语命名实体抽取：一种基于局部语法的方法”，载于2009年国际计算机科学与信息技术多学科会议。IEEE，2009年，第 139–143 页。

[47] A. Elsebai、F. Meziane、F. Z. Belkredim 和 U. H. B. Bouali，“一种基于规则的阿拉伯语人名抽取系统”。[48] T. Buckwalter，“阿拉伯语正字法与形态分析中的问题”，载于阿拉伯文字语言计算方法研讨会论文集，2004年，第 31–34 页。[49] F. Harrag、E. El-Qawasmeh 和 A. M. Salman Al-Salman，“从先知圣训文本（圣训）中抽取命名实体”，载于国际软件工程与计算机系统会议。施普林格，2011年，第 289–297 页。[50] Y. Benajiba、M. Diab、P. Rosso 等，“阿拉伯语命名实体识别：一种基于支持向量机的方法”，载于2008年阿拉伯国际信息技术会议（ACIT）论文集。阿拉伯大学协会，约旦安曼，2008年，第 16–18 页。[51] M. A. Bidhendi、B. Minaei-Bidgoli 和 H. Jouzi，“从古代伊斯兰阿拉伯语文本中抽取人名”，载于宗教文本语言资源与评测（LRE-Rel）研讨会项目论文集、第八届语言资源与评测国际会议（LREC 2012），2012年，第 1–6 页。[52] T. Hastie、R. Tibshirani、J. H. Friedman 和 J. H. Friedman，《统计学习要素：数据挖掘、推断与预测》。施普林格，2009年，第 2 卷。[53] R. Koulali 和 A. Meziane，“对阿拉伯语命名实体识别的贡献”，载于2012年第十届信息通信技术与知识工程国际会议。IEEE，2012年，第 46–52 页。[54] A. Lotfy、C. Sabty 和 S. Abdennadher，“利用嵌入特征增强阿拉伯语命名实体识别的深度学习”。[55] M. Gridach，“阿拉伯语命名实体识别的深度学习方法”，载于智能文本处理与计算语言学国际会议。施普林格，2016年，第 439–451 页。[56] T. Mikolov 、 I. Sutskever 、 K. Chen 、 G. S. Corrado 和 J. Dean，“词语和短语的分布式表示及其组合性”，《神经信息处理系统进展》，第 26 卷，2013年。[57] Y. Zhao、Y. Shen 和 J. Yao，“采用分层多尺度密集连接进行文本分类的循环神经网络”，载于IJCAI，2019年，第 5450–5456 页。[58] J. C.-W. Lin、Y. Shao、Y. Djenouri 和 U. Yun，“ASRNN：一种用于序列标注的注意力模型循环神经网络”，《知识型系统》，第 212 卷，文章编号 106548，2021年。[59] M. N. A. Ali、G. Tan 和 A. Hussain，“用于阿拉伯语命名实体识别的双向循环神经网络方法”，《未来互联网》，第 10 卷，第 12 期，2018年。 [在线]。可访问：https://www.mdpi.com/1999 5903 10 12 123 - / / / Z. Huang、[60] W. Xu 和 K. Yu，“用于序列标注的双向长短期记忆网络-条件随机场模型”，arXiv预印本 arXiv:1508.01991，2015年。[61] K. Shahina、P. Jyothsna、G. Prabha、B. Premjith 和 K. Soman，“采用深度学习算法进行阿拉伯语命名实体识别的序列标注方法”，载于2019年数据科学与通信国际会议（IconDSC）。IEEE，2019年，第 1–6 页。[62] M. Gridach，“面向社交媒体阿拉伯语命名实体识别的字符感知神经网络”，载于第六届南亚与东南亚自然语言处理研讨会论文集（WSSANLP2016），2016年12月，第 23–32 页。[63] K. Darwish，“利用跨语言资源进行命名实体识别：以阿拉伯语为例”，载于计算语言学协会第51届年会论文集（第1卷：长论文），2013年，第 1558–1567 页。[64] D. Awad、C. Sabty、M. Elmahdy 和 S. Abdennadher，“利用深度学习进行阿拉伯语姓名实体识别”，载于统计语言与语音处理国际会议。施普林格，2018年，第 105–116 页。[65] M. N. A. Ali、G. Tan 和 A. Hussain，“利用多重注意力层提升阿拉伯语命名实体识别”，《IEEE Access》，第 7 卷，第 46 575–46 582 页，2019年。[66] C. Helwe 和 S. Elbassuoni，“通过深度协同学习进行阿拉伯语命名实体识别”，《人工智能评论》，第 52 卷，第 1 期，第 197–215 页，2019年。[67] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“注意力机制就是你所需要的一切”，《神经信息处理系统进展》，第 30 卷，2017年。[68] L. Liu、J. Shang 和 J. Han，“阿拉伯语命名实体识别：哪些方法有效，接下来该做什么”，载于第四届阿拉伯语自然语言处理研讨会论文集。意大利佛罗伦萨：计算语言学协会，2019年8月，第 60–67 页。

[69] J. D. M.-W. C. Kenton 和 L. K. Toutanova，“Bert：用于语言理解的深度双向变换器预训练”，载于 NAACL-HLT会议论文集，2019年，页码 4171–4186。[70] K. Clark、M.-T. Luong、Q. V. Le 和 C. D. Manning，“Electra：将文本编码器预训练为判别器而非生成器”，ArXiv，第 abs/2003.10555 卷，2020年。[71] Z. Lan、M. Chen、S. Goodman、K. Gimpel、P. Sharma 和 R. Sori-cut，“Albert：用于语言表示自监督学习的轻量级 bert”，arXiv预印本 arXiv:1909.11942，2019年。[72] A. Safaya，“Arabic-albert”，2020年8月。 [在线]。获取地址：https://doi.org/10 5281 4718724 。/[73] zenodo。A. Ghaddar、Y. Wu、S. Bagga、A. Rashid、K. Bibi、M. Reza-gholizadeh、C. Xing、Y. Wang、D. Xinyu、Z. Wang 等人，“重新审视预训练语言模型及其在阿拉伯语自然语言理解中的评估”，arXiv 预印本 arXiv:2205.10687，2022年。[74] W. Antoun、F. Baly 和 H. M. Hajj，“AraGPT2：用于阿拉伯语语言生成的预训练Transformer”，载于 WANLP，2021年。[75] E. M. B. Nagoudi、A. Elmadany 和 M. Abdul-Mageed，“AraT5：用于阿拉伯语语言生成的文本到文本变换器”，载于计算语言学协会第60届年会论文集（第1卷：长篇论文），2022年5月，页码 628–647。[76] N. Alsaaran 和 M. AlRabiah，“使用变体深度神经网络架构和 bert 进行古典阿拉伯语命名实体识别”，IEEE Access，第9卷，页码 91 537–91 547，2021年。[77] M. S. Al-Qurishi 和 R. Souissi，“基于Transformer-条件随机场模型的阿拉伯语命名实体识别”，载于第四届自然语言与语音处理国际会议（ICNLSP 2021）论文集。意大利 Trento：计算语言学协会，2021年11月12–13日，页码 262–271。[78] [ ] 2021 1 31 在线。获取地址：https://aclanthology.org/。icnlsp-。C. Helwe、G. Dib、M. Shamas 和 S. Elbassuoni，“一种用于阿拉伯语命名实体识别的半监督 BERT 方法”，载于第五届阿拉伯语自然语言处理研讨会论文集，2020年，页码 49–57。[79] M. El-Haj 和 R. Koulali，“Kalimat：多用途阿拉伯语语料库”，载于第二届阿拉伯语语料库语言学研讨会（WACL-2），2013年，页码 22–25。[80] M. Abdul-Mageed、A. Elmadany 和 E. M. B. Nagoudi，“Ar-bert 与 marbert：用于阿拉伯语的深度双向变换器”，载于计算语言学协会第59届年会暨第11届自然语言处理国际联合会议论文集（第1卷：长篇论文），2021年，页码 7088–7105。[81] N. Boudjellal、H. Zhang、A. Khan、A. Ahmad、R. Naseem、J. Shang 和 L. Dai，“ABioNER：一种基于 BERT 的阿拉伯语生物医学命名实体识别模型”，Complex.，第2021卷，页码 6 633 213:1–6 633 213:6，2021年。[82] K. Shaalan 和 M. Oudah，“一种混合式阿拉伯语命名实体识别方法”，Journal of Information Science，第40卷，第1期，页码 67–87，2014年。[83] F. Dahan、A. Touir 和 H. Mathkour，“用于自动阿拉伯语命名实体识别的一阶隐马尔可夫模型”，International Journal of Computer Applications，第123卷，第7期，2015年。[84] F. B. Mesmia、K. Haddar、N. Friburger 和 D. Maurel，“CasaNER：阿拉伯语命名实体识别工具”，载于《智能自然语言处理：趋势与应用》。施普林格，2018年，页码 173–198。[85] M. Khalifa 和 K. Shaalan，“使用长短期记忆网络进行阿拉伯语命名实体识别的字符卷积”，Computer Speech & Language，第58卷，页码 335–346，2019年。[86] H. M. Elsherif、K. M. Alomari、A. Q. M. AlHamad 和 K. Shaalan，“使用门控机制的阿拉伯语基于规则的命名实体识别系统”，载于 MLDM (1)，2019年，页码 1–15。[87] N. Alsaaran 和 M. Alrabiah，“阿拉伯语命名实体识别：一种 bert-bgru 方法”，Comput. Mater. Contin.，第68卷，页码 471–485，2021年。[88] A. Rom 和 K. Bar，“利用预训练语言模型处理无点阿拉伯语”，载于第四届自然语言与语音处理国际会议（ICNLSP 2021）论文集，2021年，页码 89–94。[89] S. Ruder，“深度神经网络多任务学习概述”，arXiv电子印本，页码 arXiv–1706，2017年。[90] Y. Zhang 和 Q. Yang，“多任务学习概述”，National Science Review，第5卷，第1期，页码 30–43，2018年。

[91] J. Worsham 和 J. Kalita，“2020年代自然语言处理中的多任务学习：我们将走向何方？”，《模式识别快报》，第136卷，页码 120–126，2020年。[92] M. Jarrar、M. Khalilia 和 S. Ghanem，“Wojood：嵌套阿拉伯语命名实体语料库及其基于 bert 的识别”，载于语言资源与评测国际会议论文集（LREC 2022），法国马赛，2022年6月。[93] S. Ahmed、R. van der Goot、M. Rehman、C. Kruse、 ¨O. ¨Ozsoy、A. Mehler 和 G. Roig，“Tafsir数据集：面向古典阿拉伯文学命名实体识别和主题建模的新型多任务基准”，载于第29届国际计算语言学会议论文集，2022年，页码 3753–3768。[94] N. Houlsby、A. Giurgiu、S. Jastrzebski、B. Morrone、Q. De Larous-silhe、A. Gesmundo、M. Attariyan 和 S. Gelly，“面向自然语言处理的参数高效迁移学习”，载于国际机器学习会议。PMLR，2019年，页码 2790–2799。[95] F. Zhuang、Z. Qi、K. Duan、D. Xi、Y. Zhu、H. Zhu、H. Xiong 和 Q. He，“迁移学习综合综述”，IEEE会议录，第109卷，第1期，页码 43–76，2020年。[96] M. Al-Smadi、S. Al-Zboon、Y. Jararweh 和 P. Juola，“利用深度神经网络进行阿拉伯语命名实体识别的迁移学习”，IEEE Access，第8卷，页码 37 736–37 745，2020年。[97] Y. Yang、D. Cer、A. Ahmad、M. Guo、J. Law、N. Constant、G. H. Abrego、S. Yuan、C. Tar、Y.-H. Sung 等人，“用于语义检索的多语言通用句子编码器”，arXiv预印本 arXiv:1907.04307，2019年。[98] F. Alotaibi 和 M. Lee，“一种用于细粒度阿拉伯语命名实体识别的特征表示混合方法”，载于 COLING 2014（第25届国际计算语言学会议）技术论文集，2014年，页码 984–995。[99] W. Lan、Y. Chen、W. Xu 和 A. Ritter，“预训练Transformer用于阿拉伯语信息抽取的实证研究”，arXiv预印本 arXiv:2004.14519，2020年。[100] X. Li、J. Feng、Y. Meng、Q. Han、F. Wu 和 J. Li，“用于命名实体识别的统一机器阅读理解框架”，载于计算语言学协会第58届年会论文集，2020年，页码 5849–5859。[101] J. Yu、B. Bohnet 和 M. Poesio，“将命名实体识别视为依存句法分析”，载于计算语言学协会第58届年会论文集，2020年，页码 6470–6476。[102] J. Li、H. Fei、J. Liu、S. Wu、M. Zhang、C. Teng、D. Ji 和 F. Li，“将统一命名实体识别视为词间关系分类”，载于AAAI人工智能会议论文集，第36卷，第10期，2022年，页码 10 965–10 973。[103] Y. Gu、X. Qu、Z. Wang、Y. Zheng、B. Huai 和 N. J. Yuan，“深入探究规律性：一种简单而有效的中文命名实体识别方法”，载于计算语言学协会论文集：NAACL 2022。美国西雅图：计算语言学协会，2022年7月，页码 1863–1873。 [在线]。获取地址：https://aclanthology.org/2022.findings-naacl。[104] Y. Shen、X. Ma、Z. Tan、S. Zhang、W. Wang 和 W. Lu，“定位与标注：一种用于嵌套命名实体识别的两阶段标识器”，载于计算语言学协会第59届年会暨第11届自然语言处理国际联合会议论文集（第1卷：长篇论文），2021年，页码 2782–2794。[105] C. Xia、C. Zhang、T. Yang、Y. Li、N. Du、X. Wu、W. Fan、F. Ma 和 S. Y. Philip，“多粒度命名实体识别”，载于计算语言学协会第57届年会论文集，2019年，页码 1430–1440。[106] Y. Lu、Q. Liu、D. Dai、X. Xiao、H. Lin、X. Han、L. Sun 和 H. Wu，“面向通用信息抽取的统一结构生成”，载于计算语言学协会第60届年会论文集（第1卷：长篇论文），2022年，页码 5755–5772。[107] C. Raffel、N. Shazeer、A. Roberts、K. Lee、S. Narang、M. Matena、Y. Zhou、W. Li、P. J. Liu 等人，“通过统一文本到文本Transformer探索迁移学习的极限”，J. Mach. Learn. Res.，第21卷，第140期，页码 1–67，2020年。[108] M. Lewis、Y. Liu、N. Goyal、M. Ghazvininejad、A. Mohamed、O. Levy、V. Stoyanov 和 L. Zettlemoyer，“BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练”，载于计算语言学协会第58届年会论文集，2020年，页码 7871–7880。[109] H. Yan, T. Gui, J. Dai, Q. Guo, Z. Zhang, 和 X. Qiu，《一种适用于各种命名实体识别子任务的统一生成式框架》，载于计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），2021年，第5808–5822页。[110] X. Chen, N. Zhang, L. Li, X. Xie, S. Deng, C. Tan, F. Huang, L. Si, 和 H. Chen，《Lightner：一种采用提示引导注意力的轻量级生成式框架，用于低资源命名实体识别》，*arXiv*预印本 *arXiv:2109.00720*，2021年。[111] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, 和 Y.-N. Chen，《用于联合槽位填充与意图预测的槽门控建模》，载于计算语言学协会北美分会*2018*年年会：人类语言技术论文集，第*2*卷（短篇论文），2018年，第753–757页。[112] Q. Chen, Z. Zhuo, 和 W. Wang，《用于联合意图分类与槽位填充的BERT》，*arXiv*预印本 *arXiv:1902.10909*，2019年。[113] H. Weld, X. Huang, S. Long, J. Poon, 和 S. C. Han，《自然语言理解中联合意图检测与槽位填充模型综述》，*ACM*计算调查（*CSUR*），2021年。[114] K. Mai, T.-H. Pham, M. T. Nguyen, T. D. Nguyen, D. Bollegala, R. Sasano, 和 S. Sekine，《细粒度命名实体识别的实证研究》，载于第*27*届国际计算语言学会议论文集，2018年，第711–722页。[115] M. Peng, R. Ma, Q. Zhang, L. Zhao, M. Wei, C. Sun, 和 X.-J. Huang，《迈向识别更多命名实体类型：一种仅使用实体词典的高效实现》，载于计算语言学协会论文集：*EMNLP 2020*，2020年，第678–688页。[116] J. Wan, D. Ru, W. Zhang, 和 Y. Yu，《基于跨度级图的嵌套命名实体识别》，载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），2022年，第892–903页。[117] C. Lou, S. Yang, 和 K. Tu，《将嵌套命名实体识别视为潜在词汇化成分句法分析》，载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），2022年，第6183–6198页。[118] F. Li, Z. Lin, M. Zhang, 和 D. Ji，《一种用于联合识别重叠和不连续命名实体的基于跨度的模型》，载于计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），2021年，第4814–4828页。[119] Y. Wang, B. Yu, H. Zhu, T. Liu, N. Yu, 和 L. Sun，《将不连续命名实体识别视为最大团发现》，载于计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），2021年，第764–774页。[120]X. Dai, S. Karimi, B. Hachey, 和 C. Paris，《一种有效的基于转移的不连续命名实体识别模型》，载于计算语言学协会第*58*届年会论文集，2020年，第5860–5870页。[121] J. Liu, D. Ji, J. Li, D. Xie, C. Teng, L. Zhao, 和 F. Li，《Toe：一种通过嵌入标签/词语关系并采用更多细粒度标签增强的网格标注不连续命名实体识别模型》，*IEEE/ACM*音频、语音与语言处理汇刊，第31卷，第177–187页，2022年。[122] X. Dai 和 H. Adel，《命名实体识别简单数据增强方法分析》，载于第*28*届国际计算语言学会议论文集，2020年，第3861–3867页。[123] B. Ding, L. Liu, L. Bing, C. Kruengkrai, T. H. Nguyen, S. Joty, L. Si, 和 C. Miao，《DAGA：采用生成方法进行数据增强，以应对低资源标注任务》，载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，2020年，第6045–6057页。[124] J. Liu, Y. Chen, 和 J. Xu，《通过提示进行数据增强的低资源命名实体识别》，载于第*31*届人工智能国际联合会议论文集，*IJCAI-22*，第4252–4258页。[125] L. Cui, Y. Wu, J. Liu, S. Yang, 和 Y. Zhang，《使用BART进行基于模板的命名实体识别》，载于计算语言学协会论文集：*ACL-IJCNLP 2021*，2021年，第1835–1845页。[126] D.-H. Lee, A. Kadakia, K. Tan, M. Agarwal, X. Feng, T. Shibuya, R. Mitani, T. Sekiya, J. Pujara, 和 X. Ren，《优秀示例助力快速学习者：面向低资源

命名实体识别的简单基于示例的学习》，载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），2022年，第2687–2700页。[127] Y. Yang, W. Chen, Z. Li, Z. He, 和 M. Zhang，《通过部分标注学习与强化学习进行远程监督命名实体识别》，载于第*27*届国际计算语言学会议论文集，2018年，第2159–2169页。[128] M. Peng, X. Xing, Q. Zhang, J. Fu, 和 X.-J. Huang，《使用正例-未标注学习进行远程监督命名实体识别》，载于计算语言学协会第*57*届年会论文集，2019年，第2409–2419页。[129] C. Liang, Y. Yu, H. Jiang, S. Er, R. Wang, T. Zhao, 和 C. Zhang，《BOND：借助BERT并采用远程监督的开放域命名实体识别》，载于第*26*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，2020年，第1054–1064页。[130] Y. Meng, Y. Zhang, J. Huang, X. Wang, Y. Zhang, H. Ji, 和 J. Han，《通过抗噪学习与语言模型增强的自训练进行远程监督命名实体识别》，载于*2021*年自然语言处理实证方法会议论文集，2021年，第10 367–10 378页。[131] X. Zhang, B. Yu, T. Liu, Z. Zhang, J. Sheng, X. Mengge, 和 H. Xu，《通过自协同去噪学习改进远程监督命名实体识别》，载于*2021*年自然语言处理实证方法会议论文集，2021年，第10 746–10 757页。[132] X. Qu, J. Zeng, D. Liu, Z. Wang, B. Huai, 和 P. Zhou，《通过自适应教师学习和细粒度学生集成进行远程监督命名实体识别》，*arXiv*预印本 *arXiv:2212.06522*，2022年。[133] H. Jiang, D. Zhang, T. Cao, B. Yin, 和 T. Zhao，《利用少量强标注数据和大量弱标注数据进行命名实体识别》，载于计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），2021年，第1775–1789页。[134] A. Jain, B. Paranjape, 和 Z. C. Lipton，《通过机器翻译进行实体投射，以实现跨语言命名实体识别》，载于*2019*年自然语言处理实证方法会议暨第*9*届自然语言处理国际联合会议论文集（*EMNLP-IJCNLP*），2019年，第1083–1092页。[135] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, 和 E. Hovy，《自然语言处理数据增强方法综述》，载于计算语言学协会论文集：*ACL-IJCNLP 2021*，2021年，第968–988页。[136] P. Keung, Y. Lu, 和 V. Bhardwaj，《利用上下文嵌入进行对抗学习，以实现零资源跨语言分类和命名实体识别》，载于*2019*年自然语言处理实证方法会议暨第*9*届自然语言处理国际联合会议论文集（*EMNLP-IJCNLP*），2019年，第1355–1360页。[137] W. Chen, H. Jiang, Q. Wu, B. Karlsson, 和 Y. Guan，《AdvPicker：利用对抗判别器有效利用未标注数据进行跨语言命名实体识别》，载于计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），2021年，第743–753页。[138] G. Lample 和 A. Conneau，《跨语言语言模型预训练》，*arXiv*预印本 *arXiv:1901.07291*，2019年。[139] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, 和 C. Raffel，《mT5：一种大规模多语言预训练文本到文本Transformer》，载于计算语言学协会北美分会*2021*年年会：人类语言技术论文集，2021年，第483–498页。[140] A. Ebrahimi 和 K. Kann，《如何将预训练多语言模型适配至1600种语言》，载于计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），2021年，第4555–4567页。[141] T. Pires, E. Schlinger, 和 D. Garrette，《多语言BERT究竟有多多语言？》，载于计算语言学协会第*57*届年会论文集，2019年，第4996–5001页。[142] A. Sil 和 A. Yates，《联合命名实体识别与链接的重排序》，载于第*22*届*ACM*信息与知识管理国际会议论文集，2013年，第2369–2374页。

[143] G. Luo、X. Huang、C.-Y. Lin 和 Z. Nie，“实体识别与消歧联合方法”，载于*2015*年自然语言处理实证方法会议论文集，2015年，第879–888页。[144] P. H. Martins、Z. Marinho 和 A. F. Martins，“命名实体识别与实体链接的联合学习”，*ACL 2019*，190，2019年，第。[145] D. Zelenko、C. Aone 和 A. Richardella，“关系抽取的核方法”，《机器学习研究杂志》，第3卷，2月，第1083–1106页，2003年。[146] Y. S. Chan 和 D. Roth，“利用句法语义结构进行关系抽取”，载于计算语言学协会第*49*届年会：人类语言技术论文集，2011年，第551–560页。[147] M. Miwa 和 M. Bansal，“使用序列和树结构上的长短期记忆网络进行端到端关系抽取”，载于计算语言学协会第*54*届年会论文集（第*1*卷：长篇论文），2016年，第1105–1116页。[148] J. Wang 和 W. Lu，“两个胜过一个：利用表格序列编码器联合抽取实体与关系”，载于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），2020年，第1706–1721页。[149] A. Abdaoui、M. Berrimi、M. Oussalah 和 A. Moussaoui，“DziriBERT：一种面向阿尔及利亚方言的预训练语言模型”，*arXiv*预印本 *arXiv:2109.12346*，2021年。[150] A. El Mekki、A. El Mahdaouy、I. Berrada 和 A. Khoumsi，“AdaSL：一种用于阿拉伯语多方言序列标注的无监督领域自适应框架”，信息处理与管理，第59卷，第4期，文章编号102964，2022年。