arXiv:2302.05019v1 [cs.IR] 10 Feb 2023

自动知识图谱构建综述

LINGFENG ZHONG,澳大利亚麦考瑞大学JI
A WU,澳大利亚麦考瑞大学QIAN LI, 中国北京航空航
天大学HAO PENG,中国北京航空航天大学
XINDONG WU,中国合肥工业大学

自动知识图谱构建旨在生成结构化人类知识。为此，长期以来，研究者投入大量精力，从不同数据源中提取具有信息价值的事实模式。然而，近来研究兴趣已转向获取超越信息性数据的概念化结构化知识。此外，研究人员还在探索如何在多样化场景中处理复杂构建任务的新方法。因此，人们需要对组织超越数据层面提及的知识结构的范式进行系统综述。为满足这一需求，我们全面综述了 300 多种方法，总结知识图谱构建领域的最新进展。知识图谱的构建分为三个步骤：知识获取、知识精炼和知识演化。我们详细回顾了知识获取过程，包括获取具有细粒度类型的实体及其与知识图谱的概念关联、解决共指问题，以及在复杂场景中提取实体关系。本文综述了知识精炼模型，包括知识图谱补全和知识融合。我们还系统介绍了处理知识演化的方法，包括条件知识获取、条件知识图谱补全和知识动态。我们从数据环境、动机和架构三个维度提出范式，用于比较这些方法之间的差异。此外，我们还简要介绍了可获取资源，帮助读者开发实用的知识图谱系统。最后，本文讨论了未来探索所面临的挑战和可能方向。

其他关键词和短语：知识图谱、深度学习、信息抽取、知识图谱补全、知识融合、逻辑推理

ACM 引用格式：
Lingfeng Zhong、Jia Wu、Qian Li、Hao Peng 和 Xindong Wu。2022年。自动知识图谱构建综述。*ACM*计算调查 36, 4, 第66号文章（2022年7月），50 页。https://doi.org/ 0000001.0000001

1 引言

知识图谱为搜索引擎等应用提供了组织良好的人类知识 [1], 推荐系统 [2], 和问答 [3]。

作者地址：Lingfeng Zhong，澳大利亚新南威尔士州悉尼麦考瑞大学，dmiczlf@gmail.com；Jia Wu，澳大利亚新南威尔士州悉尼麦考瑞大学，jia.wu@mq.edu.au；Qian Li，中国北京海淀北京航空航天大学，liqian@ act.buaa.edu.cn；Hao Peng，中国北京海淀北京航空航天大学，penghao@act.buaa.edu.cn；Xindong Wu，中国合肥合肥工业大学，uvmxwu@gmail.com。

允许免费为个人或课堂使用而制作本作品全部或部分内容的数字副本或纸质副本，但前提是不得为营利或商业利益而制作或分发副本，且副本首页须附有本声明和完整引文。本作品中归属 ACM 以外其他方所有的组成部分，其版权必须得到尊重。注明出处后允许摘录。以其他方式复制、再版、发布到服务器或重新分发至列表，均须事先获得明确许可和/或支付费用。请向 permissions@acm.org 申请许可。© 2022年美国计算机协会。
0360-0300/2022/7-ART66 $15.00https://doi.org/0000001.0000001

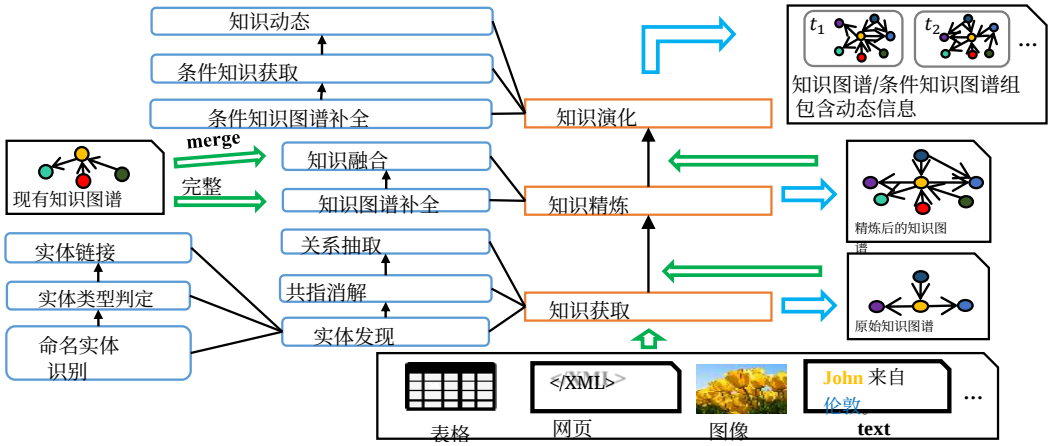


图 1. 构建知识图谱的一般流程。在该图中，通过获取知识，将半结构化或非结构化输入数据转化为原始知识图谱。随后，知识将经过精炼，以完善该知识图谱，或利用其他现有知识图谱丰富它。如果输入仅为现有知识图谱，则会直接进入知识精炼流程。最后，知识演化流程将尝试获取一组包含图谱演化动态信息的知识图谱/条件知识图谱。

许多知名的大型知识图谱系统都是通过众包构建的，例如 Freebase [4] 和 Wikidata [5]。因此，能够从非结构化或半结构化数据中自动构建知识图谱的系统化方案，可以大幅减轻实际应用中极其繁重的人工工作。

知识图谱是一种由边和节点构成的语义图，用于描述现实世界对象的知识。在这些结构中，知识元组是承载知识的最小单元。元组由两个表示概念的节点组成，二者通过一条表示关系的边相连。因此，构建知识图谱的任务是在特定领域或开放领域中发现构成知识图谱的要素。在该学科研究的早期，研究人员主要致力于从半结构化或非结构化文本数据中，以模式化知识提及的形式挖掘事实元组。TextRunner [6] 和 KnowItAll [7] 等信息抽取系统是早期知识图谱构建的里程碑，其发展由指定规则或聚类驱动。遗憾的是，这些设计未能充分利用背景知识，因此存在两大主要缺陷：1）内容贫乏：传统信息抽取系统无法创建实体，也无法区分不同表述中的实体，从而阻碍了知识聚合；2）信息不足：传统信息抽取系统只从句法结构中抽取信息，却无法捕捉给定表达中的语义指称。此外，传统的基于规则的信息抽取系统还需要大量特征工程和额外的专家知识。Wu 等人 [8] 指出，如果一个知识图谱系统没有利用关于概念的背景知识来组织节点和边，那么它就只是一个数据图。

针对这一问题，研究人员随后采用划分明确的获取子任务来组织语义知识结构。最经典的范式是流水线：首先发现并链接概念实体，解决共指提及，然后提取实体之间的关系。知识图构建的一般流程如图所示。1。

近年来，深度学习方法在自然语言处理（NLP）领域取得了巨大突破，这些突破也推动了知识图谱

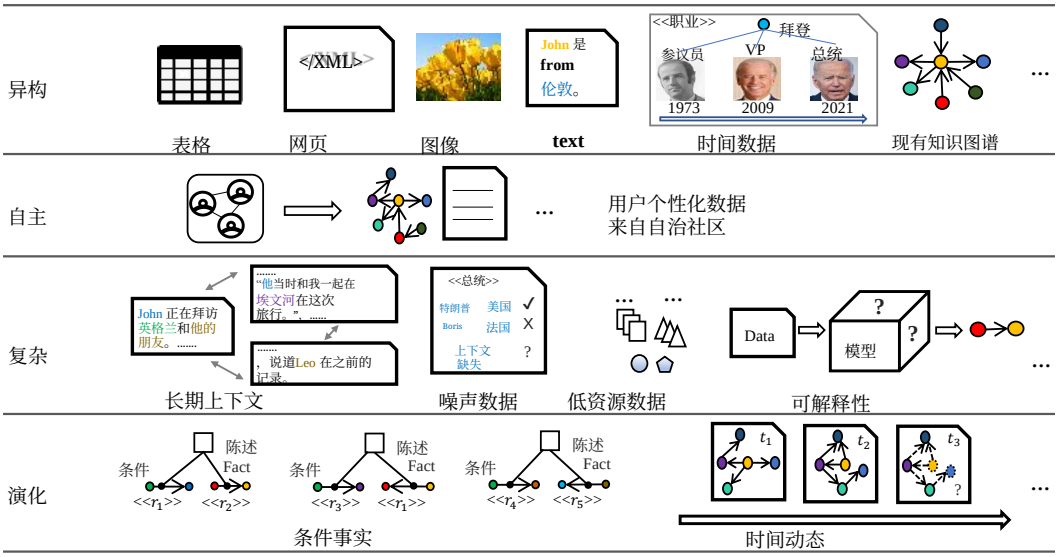


图 2. HACE 环境所带来挑战的示意图。就异构数据而言，基于半结构化和非结构化数据的知识图谱构建将在第 4 节中介绍。现有知识图谱的细化方法详见第 5 节。获取时间数据的方法将在第 6.1 节中介绍。第 8.2 节讨论多模态知识图谱。就复杂数据而言，第 4.1.2、4.2、4.3.6 节讨论长期上下文及其与多个知识图谱构建任务的关联。应对噪声数据的方法主要见第 4.3.3 节。模型可解释性将在第 5.1.3 节介绍。就演化数据而言，第 6 节介绍知识演化方面的近期工作，第 8.5 节讨论自主数据。

构建在多个方面的应用。许多深度学习模型在命名实体识别 [9][10], 实体类型判定 [11][12], 实体链接[13][14], 共指消解 [15], 关系抽取 [16][17]等任务中表现出色。此外，人们还开发了深度知识表示模型，以完善知识图谱。完善工作包括补全损坏元组、利用已构建知识图谱的内在图结构发现新元组，以及合并来自不同来源的图以构建新的知识图谱。目前，许多知识库[1]，（如 TransOMCS [18], ASER [19] 和中华谱 [20] ）已将自动知识图谱构建方法付诸实践。

此外，随着 BERT 等深度学习模型的预训练取得进展，以及大规模图卷积网络（ GCN）模型的发展，知识图谱构建任务正应用于大数据环境中的更复杂场景。除了处理网页和表格形式等异构数据的系统之外，人们也越来越关注应对复杂数据的有效方法——例如，联合统一多个获取子任务，或从长期上下文 [22],噪声数据 [23], 或低资源数据 [24]中获取知识图谱的解决方案。

就知识图谱精炼任务而言，可解释推理已成为一种普遍趋势。研究人员正在寻求将跨语言知识相融合，并通过逻辑和推理推导节点之间新关系的解决方案。研究人员还在关注条件知识的知识图谱，例如时序知识图谱 [25]以及通用条件知识图谱 [26]。主动学习 [27], 会向人类用户询问未知的有价值

[1]在本文中，知识库（KB）和知识图谱（KG）是同义术语。

表1. 现有知识图谱构建综述的比较。

综述	Year	主题	KA				KGR			目标数据					资源	
			实体	关系	CO	条件	KGC	TKGC	KF	网页	表格	句子	文档	工具	数据	集
我们的综述	2022	HACE环境中的知识图谱构建整体流程	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Paulheim [30]	2017	知识图谱细化	●	●	-	-	●	-	-	●	●	●	-	-	-	●
Yan 等人 [32]	2018	知识图谱应用、构建	●	●	●	-	●	-	-	●	●	●	●	●	●	●
Wu 等人 [31]	2019	原始知识图谱构建	●	●	●	-	-	-	-	●	-	●	-	●	●	●
Ji 等人 [35]	2020	知识图谱应用、表示、获取	●	●	-	-	●	●	●	-	-	●	-	●	●	●
Arora [36]	2020	知识图谱补全	-	-	-	-	●	-	-	-	-	-	-	-	-	●
Nayak 等人 [33]	2021	关系三元组抽取	●	●	-	-	-	-	-	●	-	●	●	●	●	●
Pawar 等人 [34]	2021	实体与关系联合抽取	●	●	-	-	-	-	-	-	-	●	-	-	-	●
Hogan 等人 [29]	2021	知识图谱概述	●	●	-	-	●	●	-	●	●	●	-	●	●	●
Cai 等人 [37]	2022	时序知识图谱	-	-	-	-	●	●	-	-	-	-	-	-	-	●

*-：未涵盖，●：涵盖 1-5 条参考文献，●：涵盖 6-14 条参考文献，●： 15+ 涵盖参考文献。*Ent：实体，Rel：关系，CO：共指，Cond：条件（时间戳或先决条件）。*Web：网页，Tab：表格形式，Sent：句子，Doc：文本文档。*KA：知识获取，KGR：知识图谱精炼，KF：知识融合。*KGC：知识图谱补全，TKGC：时序知识图谱补全。

用于数据收集，是处理来自自治社区的知识的另一个重要方向。Wu 等人 [28] 总结了大数据环境中知识发现面临的挑战，并以 HACE 定理加以阐述，如图所示2。

1.1 主要差异与贡献

许多综述概述了知识图谱及其应用。例如，Hogan 等人 [29] 对知识图谱进行了百科全书式的综述，而 Paulheim [30]则探讨了知识图谱优化与补全的方法。其他综述总结了从非结构化或半结构化数据中获取知识的方法。Wu 等人 [31] 评述了针对文本知识图谱构建子任务的竞争性工具与模型，包括关系抽取、命名实体识别和共指消解；Yan 等人 [32] 则考察了适用于不同数据类型的方法，例如网页、表格形式等。文献 [33], [34]综述了通过深度学习方法联合抽取实体及其关系的研究。一些综述也聚焦于从现有知识图谱中获取知识。诸如 [35] 和 [36] 的既有研究也涵盖了知识表示学习和知识图谱补全方法，而 Cai 等人 [37] 则深入探讨了时序知识图谱。表格 1 比较了以往研究与本综述。

与其他综述不同，我们深入探讨近期知识图谱构建模型的范式，并按照 HACE 环境的不同阶段和方面组织内容。我们还介绍实用资源，并结合数据、模型和架构讨论未来面临的挑战与发展方向。因此，我们的贡献总结如下：

- 我们通过给出正式定义和分类，介绍知识图谱构建流程及各类知识图谱。我们还汇总了知识图谱相关资源的必要信息，包括实用知识图谱项目和构建工具，并列出其发布年份、引用情况和访问链接，供读者比较。
- 我们根据任务背景和挑战，全面分析不同场景下的知识图谱构建模型，涵盖从知识获取到知识图谱精炼的各个环节。我们总结经典模型和新颖模型的动机与设计，并主要从架构和改进方面阐述其实际应用。
- 我们讨论 HACE 大数据环境中的知识图谱构建，包括含噪数据、文档级数据和低资源数据。随后，我们回顾了在实现模型可解释性和获取演化条件知识方面取得的成果。最后，我们总结了影响知识图谱构建任务的主要挑战和发展方向。

表2. 实用知识图谱项目的信息。

分类	项目	知识图谱收录内容	Year	URL
百科全书知识图谱	YAGO	20亿+ 事实，6400万+ 实体	2007	https://yago-knowledge.org
	Freebase	3.6亿+ 事实三元组	2007	https://freebase-easy.cs.uni-freiburg.de/dump/
	DBpedia	320 个类别，1,650 个不同属性，247M+ 个三元组	2007	https://github.com/DBpedia/
	CN-DBpedia	9M+ 个实体，67M+ 个三元组	2015	http://kw.fudan.edu.cn/cnDBpedia/download/
	Probase	5.4M+ 个概念	2010	https://concept.research.microsoft.com/
	Wikidata	96M+个条目	2012	https://www.wikidata.org/wiki
	CN-Probase	1700万+ 实体，3300万+ “是”关系	2017	http://kw.fudan.edu.cn/apis/cnprobase/
语言知识图谱	WordNet	117 000 个同义词集	1985	https://wordnet.princeton.edu/
	ConceptNet	3400万+ 条目	1999	https://www.conceptnet.io/
	HowNet	35、202 个概念、2、196 个义原	1999	https://openhownet.thunlp.org/download
	BabelNet	1300 万个节点	2010	http://babelnet.org/rdf/page/
	THUOCL	15.7 万个+ 词节点，涵盖 73 亿+ 篇文档	2016	http://thuocl.thunlp.org/
常识知识图谱	OpenCyc	2M+ 事实三元组	1984	https://sourceforge.net/projects/opencyc/
	ASER	438M+ 节点，648M+ 边	2020年	https://github.com/HKUST-KnowComp/ASER
	TransOMCS	1800万+ +	2020年	https://github.com/HKUST-KnowComp/TransOMCS
企业支持知识图谱	谷歌知识图谱	5000亿+ 事实，涉及50亿+ 实体	2012	https://developers.google.com/knowledge-graph
	Facebook 图谱搜索	用户的动态社交网络、用户生成内容	2013	https://developers.facebook.com/docs/graph-api/
领域专用知识图谱	PubMed	5B+三元组	2000	http://pubmed.bio2rdf.org/sparql
	DrugBank	14K+ 药物实体	2006	https://go.drugbank.com/releases/latest
	AMiner ASN	2M+ 论文节点，8M+ 引用关系	2007	https://www.aminer.cn/aminernetwork
	中华谱	17M+ 人物节点	2017	https://www.zhonghuapu.com/
	OAG	369M+ 作者，380M+ 论文 92M+ 链接关系	2017	https://www.aminer.cn/data/?nav=openData#开放学术图
	TransOMCS	18M+ +	2020年	https://github.com/HKUST-KnowComp/TransOMCS
	COVID-19 概念	4784 个实体，35172 个关系链接	2020	http://openkg.cn/dataset/covid-19-concept
	AMiner COVID-19 开放数据	COVID-19 报告、新闻、研究及其他成果	2020	https://www.aminer.cn/data-covid19/
联邦知识图谱	GEDMatch	1.2M+ DNA 档案	2010	https://www.gedmatch.com/
	OpenKG.cn	200+ 个组织的数据集	2015	http://www.openkg.cn

1.2 综述的组织结构

本文综述的组织结构如下。第2节介绍知识图谱构建的背景，包括知识图谱项目的定义和资源。第3 节介绍半结构化数据预处理方法，包括内容抽取和结构解析。第4节介绍从各种数据类型和数据环境中获取实体和关系的任务处理方法。第5 节回顾利用外部结构化数据完善知识图谱的方法。第6节介绍演化知识图谱的最新成果和发展趋势，包括条件知识图谱和时序知识图谱。第7节介绍有效存储知识图谱的解决方案。最后，我们展望未来的发展方向，并在第8 节展望后续发展，同时在第9节总结全文。

2 背景

2.1 定义

已有许多研究致力于对知识图谱进行形式化定义。Wang 等人 [38] 将知识图谱建模为多关系图，其中节点是实体，边表示不同类型的关系。然而，先前的定义未考虑知识图谱中的语义结构。Ehrlinger 和 Woß [39] 进一步强调，知识图谱将信息组织成本体，并借助“推理机”启发新的知识发现。为明确突出支撑知识层面信息的核心组成部分，Wu 等人 [31] 将知识图谱定义为一种语义图，其中节点表示概念（实体/属性/事实），边则表示连接节点的关系，同时利用有关这些概念和关系的背景知识。

定义 ***1***（知识图谱）。知识图谱 G 定义为 G ={E，R，T，F_k}，其中 E 和 R 分别表示概念集和关系集。在本文中，概念可视为实体/属性。T 是事实三元组集，其中标准的二元事实是一个三元组 (h，t，r) ∈T h，t ∈E r ∈R。n元关系三元组表示为 (e_1，..... e_n，r)，其中e_1，..... e_n ∈E。F_k 是一个集合函数，表示用于约束潜在事实、使其具有知识层面信息量的背景知识，并且有 T ⊂F_k({E，R})。在实践中，背景知识可视为规则集、模式或一组隐含数学原理。

定义 **2**（知识图谱构建）。 知识图谱构建 f 是一个将数据源映射为知识图谱的过程： f: $D \times f_k(D) \rightarrow$ G，其中 D 是数据源集， $f_k(D)$ 是数据目标的背景知识，也可以是领域知识。值得注意的是，若缺少由预先设计的规则或表示语言模型提供的背景知识，知识图谱构建通常无法继续。

2.2 实用知识图谱项目

本节回顾知识图谱的代表性实用项目（数据集），包括百科知识图谱、语言知识图谱、常识知识图谱、企业支持知识图谱、特定领域知识图谱和联邦知识图谱。详情见表2。

2.2.1 百科全书知识图谱。

百科知识图谱系统地涵盖不同领域的事实或事件知识。许多研究人员基于人工构建的在线百科全书开发了知识图谱结构。例如，DBpedia [40]（基于维基百科构建）是基础性的百科知识图谱，而 Freebase [4] 则采用自动抽取工具获取更多内容。Probase [41]（由 Microsoft Concept Graph 支持）作为事件百科知识图谱，以概率模型的形式创新性地呈现了包含冲突信息的不确定事件知识。Xlore [42]（THUKC 的子项目）作为多语言百科知识图谱，通过深度学习方法在多语言内容之间建立实体链接。在 DBpedia 项目早期尝试引入自动抽取工具之后，更多知识图谱项目也决定采用这种做法，包括 Wikidata [5] 和 CN-DBpedia [43]。马克斯·普朗克研究所开发了 YAGO [44], , 将维基百科中的时间和地理结构与 WordNet 本体相结合。Minz等人 [45] 对 Freebase 应用远程监督，以自动标注实体关系。研究界也一直关注事件知识图谱。例如，CN-Probase [46] 将中文概念融入 Probase，以理解涉及不确定事件的文本数据的一般模式。

2.2.2 语言知识图谱。

语言知识图谱提供人类语言知识，以便通过本体或外部特征提供基础语义。WordNet [47] 是一本经典且广泛使用的语言学研究知识图谱词典，提供词语之间的同义和上下位关系。借助这些工具，开发者基于构建完善的语言知识图谱创建高性能词嵌入，用于下游应用。除了 WordNet，BabelNet [48] 还扩展了 WordNet，加入了百科全书中词语的跨语言属性和关系。ConceptNet [49], 作为开放链接数据的一部分，通过众包汇集概念知识；而 HowNet [50]则人工收集词语概念和属性的义原信息（最小不可分割的语义单位）。THUOCL [51] 记录经过精心筛选的网络语料库中词语的文档频率。开发者还基于构建完善的语言知识图谱，为下游应用创建了高性能词嵌入。

2.2.3 企业支持知识图谱。

知识图谱及其相关系统一直有效地支持着企业业务活动。谷歌知识图谱 (GKG) [53], 自 2012年 起便发挥核心作用，为用户查询提供知识支持，并通过更多语义相关的内容丰富搜索结果。Facebook 图谱搜索 [2] 是 Facebook 强大的语义搜索引擎，能够提供

[2]https://developers.facebook.com/docs/graph-api/

表格 3. 现成知识图谱工具的信息。

Task	Tool	Year	URL
数据预处理	WebCollector	2016	https://github.com/CrawlScript/WebCollector
	Web Scraper（v0.4.0）	2019	https://webscraper.io/
知识获取	NLTK	2002	https://www.nltk.org/
	StanfordNLP	2002	https://stanfordnlp.github.io/stanfordnlp/
	KnowItAll	2005	https://github.com/knowitall
	TextRunner	2007	https://www.cs.washington.edu/research/textrunner/
	OpenCalais	2008	http://www.opencalais.com
	ReVerb	2011	https://github.com/knowitall/reverb
	OLLIE	2012	https://knowitall.github.io/ollie/
	spaCy	2017	https://spacy.io/
	表格挖掘器+	2017	https://github.com/ziqizhang/sti
	MantisTable	2019	http://mantistable.disco.unimib.it/
	OpenNRE	2019	https://github.com/thunlp/OpenNRE
	gBuilder	2021	http://gbuilder.gstore.cn/
知识获取	Falcon-AO	2008	http://ws.nju.edu.cn/falcon-ao/
	OpenKE	2018	https://github.com/thunlp/OpenKE
	OpenNE	2019	https://github.com/thunlp/OpenNE
	OpenEA	2020	https://github.com/nju-websoft/OpenEA

通过 Facebook 社交系统中的动态知识库，为用户提供个性化答案。与 GKG 类似，Facebook 图谱搜索是 Facebook 强大的语义搜索引擎，通过动态 Facebook 社交知识库为用户提供个性化答案。

2.2.4 常识知识图谱。

常识知识图谱描绘了广泛认可的常识知识。OpenCyc [52] 是最早的此类尝试之一，它以 CycL 的形式编码知识概念、规则和常识理念。除 OpenCyc 外，ASER [19] 还提供了一种加权知识图谱，通过对动作、状态、事件等对象及其相互关系进行建模来描述常识；它通过选择依存模式获取节点，并利用 Probase 对其进行概念化。TransOMCS [18] 构建了一个自动生成的数据集，涵盖从语言图中获得的 20 种常识关系。

2.2.5 领域特定知识图谱。

研究人员还会汇集特定知识，以服务于多个专业研究领域。在生物医学领域，Pubmed [54] 提供由 Pubmed 中心发布的开放生物医学文献数据库，而 Durgbank [55] 则通过蛋白质和药物信息提供药理学方面的洞见。许多知识图谱收集工作也致力于抗击 COVID-19 疫情，例如 COVID-19 概念数据集 [3] 和 AMiner COVID-19 开放数据 [4]。在学术活动方面，AMiner 的学术社交网络（ASN） [56]展示学者及其学术活动，其网络包含论文和引文关系。类似地，开放学术图谱（OAG） [57] 提供整合的学术社交网络及论文数据。中华谱 [20] 则别出心裁地从数字化的数千本祖先族谱书籍中构建了高质量的中文家谱语义网络，并通过融合多个网络来源建立的不同家谱树，为同一人物之间提供关系链接。

2.2.6 联邦知识图谱。

隐私保护对于拥有来自不同提供方的大量用户数据的知识库至关重要，而收集多源知识图谱数据的知识社区拥有丰富的

[3]http://openkg.cn/dataset/covid-19-concept [4]
https://www.aminer.cn/data-covid19/

特征，可用于构建集成知识模型。联邦学习旨在结合来自不同数据提供方的子知识图谱特征，并通过保护机制防止数据交换。联邦策略已应用于包含敏感数据的知识图谱系统。GEDmatch[5]是一个家谱知识图谱，收集用户提供的信息以查询 DNA，同时分发联邦知识模型来探查数据跟踪。联邦策略已应用于许多包含敏感数据的知识图谱系统。这些策略有助于构建集成知识模型，同时防止数据交换。研究人员还关注联邦知识图谱平台。OpenKG.cn [58] 是一个众包社区，提供知识共享平台，通过联邦学习开发知识应用，同时支持知识区块链的去中心化。

2.3 知识图谱构建工具

本节回顾一些常用于知识图谱构建的工具。这些工具主要支持构建子任务，例如预处理、知识获取和知识精炼。这些工具的详细信息如表3所示。

2.3.1 数据预处理。

数据预处理任务会去除广告等噪声。例如，WebCollector [59] 是一种具有代表性的数据预处理工具包，它通过集成算法自动过滤广告和布局信息等非内容噪声，同时保留网页内容。许多网络爬虫也支持数据预处理任务，可抽取有用的结构或内容。除了 WebCollector，WebScraper [6] 也是一款易于使用的手动抽取工具，可用于采集多个网页。它提供用户界面以保留重点网页结构，并提供云服务器来抽取海量内容。

2.3.2 知识获取。

早期的知识获取工具包通过规则、模式和统计特征直接抽取事实三元组。这些工具包也称为信息抽取工具包。除 KnowItAll[7], 外，TextRunner [6], ReVerb [60], 等其他工具包采用半监督设计来收集关系信息。这些工具通过句法和词汇信息生成经过细化的文本三元组。此外，OLLIE [61] 还支持发现非文本三元组。

许多自然语言处理应用可以直接执行知识获取子任务，包括命名实体识别、关系抽取和共指消解任务；或者，它们可以为相关应用提供语言学特征。NLTK [62] 和 StanfordNLP [63] 是基于统计算法（如条件随机场和MEM）的强大知识获取工具包。这些工具还可以提供词性标注和名词短语块等背景特征。与此同时，TableMiner+ [64] 和 MantisTable [65]从半结构化表格中抽取知识。

近年来，开发者们开始青睐基于深度学习的工具包。例如，spaCy [66]是一款全面实用的自然语言处理工具包，集成了用于共指消解任务的 NeuralCoref，还提供了可训练的深度学习模块，用于专门的关系抽取（在 spaCy v3 中）。OpenNRE [67] 提供了多种可扩展的神经网络模型，如卷积神经网络和长短期记忆网络，用于执行监督式关系抽取。深度学习工具包为用户提供了高性能技术。对于工业应用，用户可以通过 OpenCalais [7]定制自己的知识获取解决方案。此外，gBuilder [8] 是一种基于用户选择的神经网络

[5]https://www.gedmatch.com/ [6]
https://webscraper.io/ [7]
http://www.opencalais.com [8]
http://gbuilder.gstore.cn/

架构的端到端解决方案，堪称一项里程碑，同时还支持主动学习接口以优化性能。gBuilder 可以直接从输入的非结构化数据中输出原始知识图谱。

2.3.3 知识精炼。

Knowledge 知识精炼 用于补全现有的原始知识图谱，或将其与其他知识图谱合并。基于深度学习的工具是常用的解决方案。例如，OpenKE [68] 提供多种用于知识图谱补全的知识表示模型，而 OpenEA [69] 则采用类似方法来合并知识图谱中的结构。除集成式深度学习工具包 OpenKE 和 OpenEA 外，OpenNE[9] 还集成了 Node2VEC 等嵌入模型 [70] 和 LINE [71]，以从完整的知识图谱中生成全局表示。在知识融合（合并）任务中，Falcon-AO [72]使用多种算法测量语义相似度，从而对齐采用不同记法的概念。

3 半结构化数据预处理

现实世界的原始数据源包含多种结构的内容，其中无关部分会影响知识抽取的效果。为处理杂乱的数据环境，必须进行数据预处理。预处理子任务主要包括内容抽取和结构解析。

3.1 内容抽取

许多网页包含广告等非内容噪声。内容抽取任务旨在删除这些无关元素，同时保留知识内容。用户可以手动选择网页的主体部分，例如由“<table>”包围的内容，并借助 JSoup、BeautifulSoup 和 Web Scraper 等网络爬虫来实现这一目标。这些工具会检索并解析文档对象模型（DOM）结构中的元素，随后用户即可选择网页的主体部分。然而，数据量较大时，人工处理将难以胜任。主流的自动内容抽取方法主要包括基于包装器的方法和基于统计的方法。

基于包装器的方法是最早用于检测主体内容的尝试，它利用匹配规则捕获有用内容。现成的包装器工具可以从半结构化页面自动生成规则，例如 IEPAD [73], SoftMealy [74]。自举方法通过种子示例迭代改进抽取模板，例如 [75] 和 [76]。还有一些工具包提供用户界面以优化抽取模板，例如 NoDoSE [77] 和 DEByE [78]。基于模板的包装器易于理解，在页面结构规范时也能取得可行的效果，但难以识别被复杂、新颖元素或结构所覆盖的内部内容。

用户还可以利用基于网页统计特征的方法来获取信息丰富的内容。Finn 等人 [79] 提出了一项经验假设：网页中的信息子序列包含足够多的词语，同时标签数量尽可能少。许多模型在抽取信息丰富的内容时会考虑网页内容的统计特征，例如 CETR [80] （文本长度与标签数量之比）、CETD [81] （DOM树中每个子树结构的文本密度）和 CEPR [82] （网页链接的路径比率）。用户可以使用 WebCollector [59]，它集成了基于统计的模型，用于内容抽取。另一种启发式研究方向是基于视觉特征的方法。例如，VIPS [83] 利用页面的视觉外观（如字体和颜色类型）构建内容结构树，以进行内容抽取。

对半结构化页面执行内容抽取后，用户将获得一份焕然一新、无噪声的半结构化或非结构化文档。

[9] https://github.com/thunlp/OpenNE

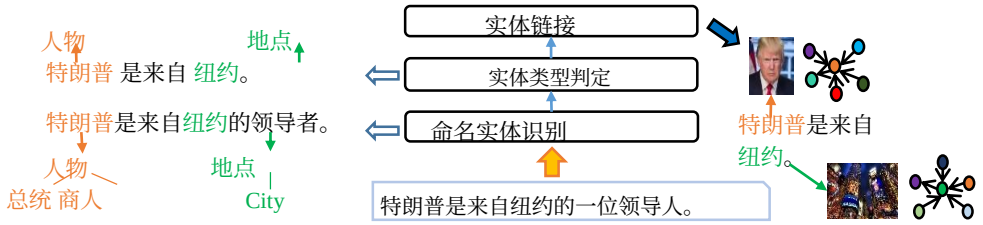


图 3. 实体发现流程。

3.2 结构解析

网页中的许多表格形式用作导航器或以样式格式化的内容容器（由内容提取器处理），并不包含关系结构。模型应在获取关系信息之前，过滤掉这些装饰性的非关系型网页表结构。

关系表解释是一项二元分类任务，用于确定表格是否包含有用信息。分类方法会分析表结构的语义特征以进行分类。Wang 和 Hu [84] 设计了一种结合支持向量机（SVM）和决策树的表分类器，该分类器基于布局和内容类型特征。类似地，WebTables [85] 开发了一种基于表格大小（行数和列数）及标签的基于规则的分类器。Eberius 等人 [86] 利用表矩阵特征开发了一个名为 DWTC 的分类系统。许多网页表格还包含数据噪声。OCTOPUS [87] 进一步将数据清洗纳入表分类任务，以筛选出有用的表格。

构建表格解读模型包括两个步骤：首先选择表格形式中的特征，然后整合学习模型来分析数据中的关系语义。我们建议读者参阅 [88]，了解更多表语法特征和高性能模型集成。

4 知识获取

知识获取是从多结构数据中收集元素以构建知识图谱的一般过程。它包括实体识别、共指消解和关系抽取。实体识别任务会发现数据中的实体提及。随后，共指消解任务会定位被指代的提及对，接着由关系抽取任务将实体与其语义关系关联起来。

4.1 实体发现

实体发现从半结构化或结构化数据中获取一部分概念，这些概念可构成知识图谱的节点。实体发现的一般流程包括命名实体识别、实体类型判定和实体链接任务。命名实体识别任务会发现指代语义实体的字符串，并将其归类为通用类型（例如人物、地点、国家、公司）。实体类型判定任务会将发现的实体归类为具体类型（例如演员、艺术家、品牌）。实体链接会将发现的实体与知识图谱中可能对应的节点关联起来。如果没有可用于链接的节点，则会创建相应的实体节点来表示新发现的实体。图 3 展示了这一一般流程的概览。

4.1.1 半结构化数据与非结构化数据中的命名实体识别。

命名实体识别任务会为半结构化或非结构化数据中的命名实体标注其位置和分类。半结构化数据由与属性结构相关的语义提示包围，而非结构化数据只包含文本。

基于规则的方法 [89]是命名实体识别的通用解决方案。对于半结构化网页数据，包装器归纳会生成规则包装器，以解析 DOM树节点等半结构。

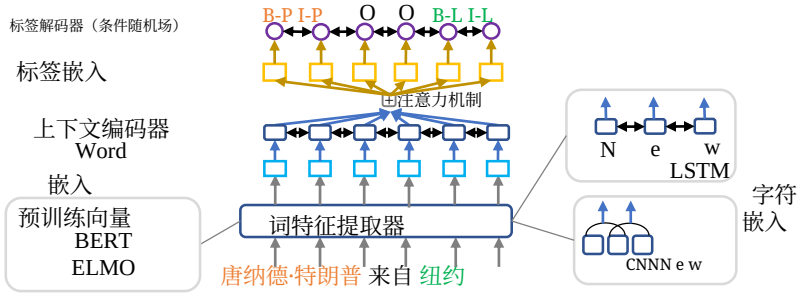


图4. 基于深度学习的命名实体识别标准架构示意图。输入句子后，此类命名实体识别模型将输出带有位置信息和粗略分类的实体标签。

和标签，从页面中提取实体。一些基于规则的解决方案是无监督的，无需人工标注，例如Omini [90]。对于表格形式中的实体，许多方法基于维基百科的属性布局提出，例如用于DBpedia 的基于规则的工具 [40][44]和 YAGO。对于非结构化数据，经典的命名实体识别系统 [91] 也依赖人工构建的规则集进行模式匹配。半监督方法通过迭代生成更精细的新模式来改进基于规则的命名实体识别，这些模式通过模式种子和评分得到，例如基于自举的命名实体识别 [92]。

基于统计的方法将命名实体识别视为序列分类标注任务，根据 BIES标注体系（开始、中间、结束、单个）及其类型为实体添加标签。对于非结构化命名实体识别，关键假设是每个词的标签仅取决于其前面的词。因此，基于隐马尔可夫模型[93] 和能够捕捉邻域依赖关系的条件随机场（CRF）模型 [94], 构建的应用，是常见的命名实体识别方案。对于半结构化表格数据，研究人员常使用条件随机场变体处理与实体相关的属性的二维特征，例如二维条件随机场 [95]。由此，他们能够以二维结构抽取每个实体的多个属性。动态条件随机场（DCRF） [96] 通过动态贝叶斯网络推断潜在的属性与实体交互；分层条件随机场 [97] 则将半结构化数据建模为分层树，以进行联合抽取。此外，Finn 和 Kushmerick [98] 开发了一种基于支持向量机的模型，用于定位文本中实体的边界。

深度学习也正成为命名实体识别领域的一种流行趋势，尤其是在文本命名实体识别方面。这些深度学习方法通常将命名实体识别视为序列到序列模型（词序列到标签序列）。这些模型会根据输入聚合上下文嵌入，上下文编码器随后通过条件随机场结构或Soft max结构等标签解码器输出词类型标签 [99]。

卷积神经网络结构主要侧重于提取局部特征以识别实体。Colloabert 等人 [100] 首次采用带条件随机场输出层的卷积神经网络，作为实体检测的统一解决方案。IDCNN 等人 [101] 通过引入膨胀卷积改进了卷积神经网络；膨胀卷积通过省略部分输入来扩大感受野，从而增强泛化能力。

RNN结构能够更好地处理长句中的全局上下文特征，例如 [102]中提出的用于生物医学实体识别的单向循环神经网络。然而，RNN可能会受到后续词语带来的上下文偏差影响 [9]。因此，许多模型采用双向循环神经网络，例如 [9]中基于双向长短期记忆网络-条件随机场的模型，以及 [103]中的基于门控循环单元的命名实体识别模型。字符和词编码器的组合也是广泛采用的结构，例如一种由用于字符嵌入的卷积神经网络和用于词嵌入的长短期记忆网络组成的结构 [10]。基于深度学习的命名实体识别模型的标准架构见图4。

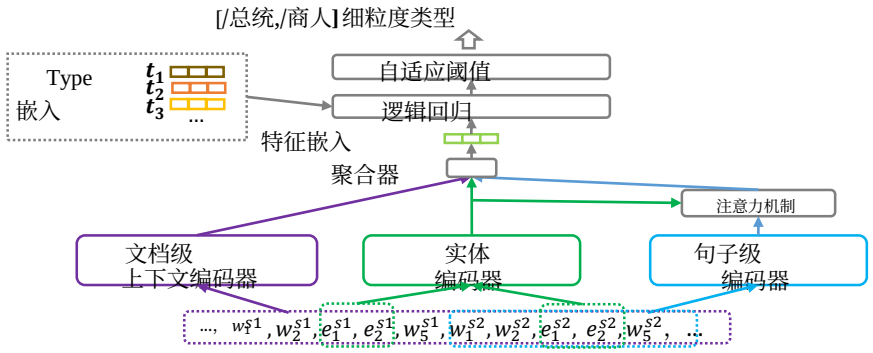


图 5. 基于深度学习的多尺度特征提取实体类型判定示意图（基于 [111]）。

另一种用于呈现全局上下文中显著交互的方向是注意力机制。Luo 等人 [104] 引入词级软注意力，以增强命名实体识别。Gregoric 等人 [105] 将词间自注意力用于命名实体识别。

图卷积网络常用于处理命名实体识别中语言图结构里的上下文。例如，Cetoli 等人 [106] 提出了一种 GCN框架，通过句法依存树，利用 GCN 结构对 LSTM 处理后的特征进行编码。能够为命名实体识别任务的训练提供背景知识表示的预训练语言模型，也在命名实体识别方面取得了突破。这类模型包括 Elmo [107], Ltp [108], 和 LUKE [109]。

4.1.2 实体类型判定。

实体类型判定（ET）任务为科学家、俱乐部和酒店等实体提供细粒度和超细粒度的类型信息。如果不执行 ET 任务，就会造成信息丢失，例如，唐纳德·特朗普既是政治家，也是商人。半结构化表格的标题为细粒度类型提供了线索。例如，“英格兰的足球运动员”暗示该实体是足球运动员。然而，对于非结构化数据，在不同上下文中标注恰当的细粒度实体类型可能很复杂。

深度学习方法应对实体类型判定的两大主要挑战：1）低频细粒度类型；以及 2）过度具体的类型判定。某些具体类型可能存在类别不平衡或出现频率较低的情况。因此，Shimaoka 等人 [110] 提出了一种用于低频实体类型判定的基于LSTM的注意力神经网络，该网络依赖于与提及和上下文表示相结合的层次化标签编码，以利用细粒度上下文特征。过度具体的类型标注虽然能得出正确类型，但并不适合当前的数据上下文。Xu 等人 [11] 对具有多个标签的实体应用了上下文外损失函数，以过滤过度具体的数据噪声；该方法假设训练期间得分概率最高的类型标签是正确标注的。为进一步探索上下文场景，Zhang 等人 [111] 引入了文档级表示，以提供用于发现实体的全局上下文。随后，利用句子级上下文表示，对出现在不同句子中的同一实体表示进行对齐。然后，通过自适应概率阈值生成实体在不同上下文中的实体类型标签。图 5 展示了一个典型的基于深度学习的实体类型判定模型。

新颖的基于嵌入的模型能够结合全局图结构特征与背景知识，通过表示来预测实体的潜在类型。研究人员报告称，经典的 TransE 模型直接用于实体类型任务时表现不佳。Moon 等人 [112]提出了 TransE-ET 模型，通过优化实体与其类型表示之间的欧氏距离来调整 TransE 模型，但该模型受限于实体类型和三元组数量不足。

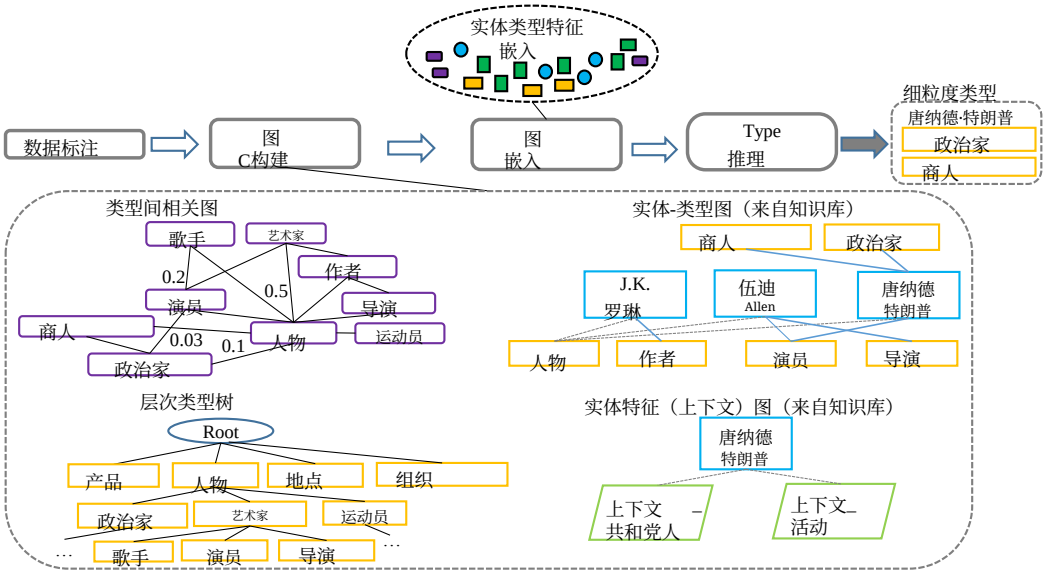


图6. 基于嵌入的实体类型识别（PTE）在异构图结构中的示意图 [12]。

特征。新的方案旨在构建各种图，以共享实体相关对象的多样化特征，并利用实体类型特征学习嵌入。PTE [12] 通过部分标签嵌入降低数据噪声。该方法构建实体与其所有类型之间的二部图，同时将实体节点与相关的已提取文本特征相连。最后，PTE 利用背景知识图谱，并根据推导出的相关性权重构建类型层级树。JOIE [113]在本体视图图和实例图中嵌入实体节点，并通过实体与类型候选项之间的前 k 名排序来收集实体类型。同样，ConnectE [114] 将实体映射到其类型，并学习知识三元组嵌入。用于实体类型任务、通过改进异构图上的嵌入而构建的实用模型（来自 Xlore 项目 [42]）还包括 [115], [116], [117]。我们在图6中展示了基于嵌入模型的实体类型任务所用的图结构。

4.1.3 半结构化数据与非结构化数据的实体链接。

实体链接（EL）任务也称为实体消歧，其目标是将实体提及链接到知识图谱中相应的对象。一个文本提及可能对应不同的指代对象，例如，“特斯拉”一词可能指汽车、公司或科学家。实体链接通过上下文信息，将不同数据背景中的提及与各自节点相连接。对于半结构化数据，实体链接利用列标题、类型标签、表格单元格文本和超链接中的语义提示来识别实体。对于非结构化文本，实体链接模型侧重于实体提及的上下文表示。

统计方法，尤其是基于概率图和支持向量机模型的方法，是处理半结构化和非结构化数据的通用方案。基于概率图的模型会构建表格中提及项的概率图，然后通过计算节点的语义因子来链接实体。例如，Limaye 等人 [118], 构建了一个用于集体实体链接的因子图，该因子图基于词频-逆文档频率算法，计算实体标签与单元格文本对之间的词频，以及类型标签与列标题对之间的词频。一些模型会整合外部知识库，以改进实体链接任务。例如，TabEL [119] 利用维基百科中的超链接来改进其因子图，并在进行消歧的集体分类之前，估算语义相关性特征。Wu 等人 [120] 提出了一种增强实体链接的方法

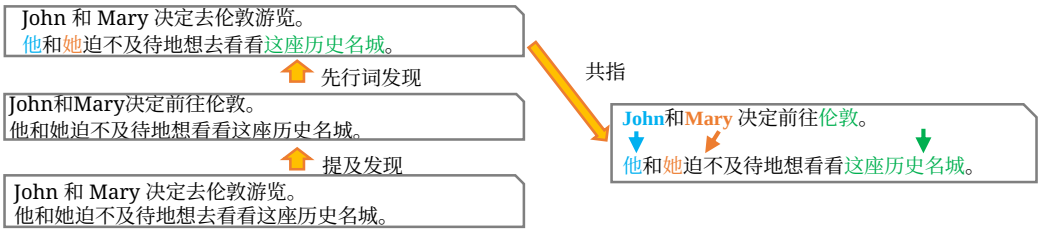


图7. 共指消解流程。首先，检测提及。然后，选择这些提及的先行词，并将其匹配为共指对。值得注意的是，共指消解任务可以在包含多个句子的文档上执行，同时处理复合句中的相同提及。

通过在多个知识库中使用“相同”边。Efthymiou 等人 [121] 系统地利用语义特征进行实体链接。其方法整合了实体上下文、最小实体上下文的向量表示，以及知识库与表格共享的示意结构。支持向量机模型将实体链接视为分类任务。在此，Mulwad 等人 [122] 提出了一种基于 SVMRanker 的模型，用于判定哪些候选节点可以链接到目标实体。同样，Guo 等人 [123] 提出了一种面向非结构化数据的概率模型，在处理非结构化数据的链接任务时利用实体的先验概率、上下文和名称。Han 等人 [124] 采用了实体引用图，并假设在同一文档中共同出现的实体在语义上相关。

基于嵌入的模型也是通过实体嵌入实现实体链接的重要方案。LIEGE [125]为网页推导分布式上下文表示，以链接实体。早期研究人员 [126] 利用词袋（BoW）为实体提及生成上下文嵌入，然后通过聚类汇集链接实体对。之后，Lasek 等人 [127] 扩展了 BoW 模型，并将语言嵌入用于实体链接任务。研究人员还致力于通过深层表示实现高性能链接。DSRM [128] 采用深度神经网络来利用语义相关性，将实体描述和关系与类型特征相结合，以提取用于链接的深层实体特征。EDKate [129] 联合学习知识库和文本数据中实体与词语的低维嵌入，捕捉超越 BoW 模型的内在实体提及特征。此外，Ganea 和 Hofmann [13]引入注意力机制进行联合嵌入，并通过语义交互实现消歧。Le 和 Titov [14] 对上下文中提及之间的潜在关系进行建模以生成嵌入，并利用提及级和关系级归一化来计算成对连贯性评分函数。

4.1.4 其他进展。

少样本/零样本实体类型判定是一项复杂且具有挑战性的问题。Ma 等人 [139]提出 Proto-HLE，为零样本细粒度实体类型判定任务建模实体标签嵌入的原型，并将原型特征与层次化类型标签相结合，以推断新类型的关键特征。Zhang 等人 [140]进一步提出 MZET，利用上下文特征和词嵌入，并结合记忆网络，为少样本实体类型判定提供语义侧信息。

结合命名实体识别与实体链接任务的联合抽取模型，可以减少基于流水线的实体识别任务中的误差传播。NEREL [141] 通过对已提取的提及—实体对进行排序，将命名实体识别与实体链接任务结合起来，从而利用实体提及及其链接之间的交互特征。图模型也是一种有效的设计，可用于整合命名实体规范化（Named Entity Normalization，NEN）标签，将实体提及转换为无歧义的形式，例如华盛顿（人物）和华盛顿州。Liu 等人 [142] 利用因子图模型将实体链接与命名实体规范化任务结合起来，为词语实体类型及其目标节点构建条件随机场链。同样，MINTREE [143]引入了一种用于集体任务的基于树的成对链接模型。

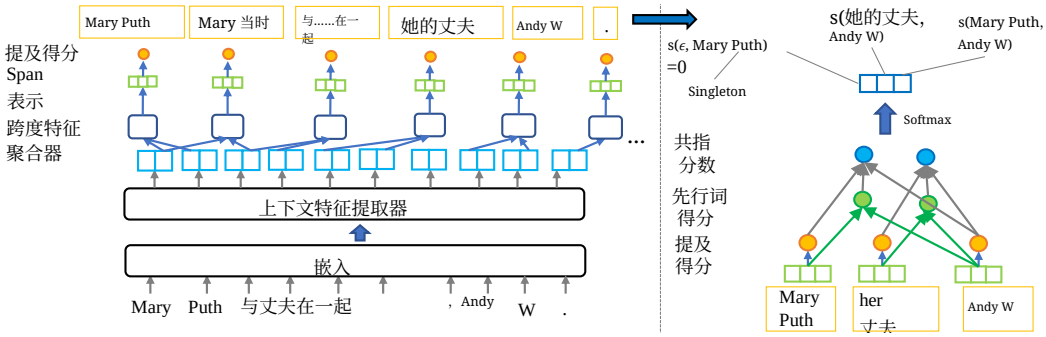


图8. 标准深度端到端模型的架构（基于 [15]）。深度学习模型通过两阶段过程来处理共指消解任务：1）提及检测，从文本中以跨度的形式发现实体提及；2）共指检测，对跨度中的先行词进行评分，以匹配并输出共指提及对。跨度包含所有词语序列的组合。本图展示了简化后的结果。

研究人员探索了更多灵活处理命名实体识别任务的策略。迁移学习在不同域或模型之间共享知识。Pan等人 [130] 提出迁移联合嵌入（TJE），将来自不同域的输出标签和输入样本共同嵌入，以融合内在实体特征。Lin 等人则采用 [131]带有适配层的神经网络，将在不同域上预训练的模型中的参数特征迁移过来。强化学习（RL）通过具有奖励策略的行为主体，使命名实体识别模型与环境域进行交互，例如基于马尔可夫决策过程（MDP）的模型 [132] 和增强型 Q网络模型[133]。值得注意的是，研究人员 [134]还利用强化学习模型减少远程监督命名实体识别数据中的噪声。对抗学习通过生成反例或扰动来增强命名实体识别模型的鲁棒性，例如，DATNet [135] 对词语表示施加扰动，并使用反例生成器（[136], [137]）。此外，主动学习通过查询用户来标注选定样本，也已应用于命名实体识别。Shen等人 [138] 在训练过程中逐步选取最具代表性的样本进行命名实体识别标注，以降低对已标注样本的依赖。

4.2 共指消解

共指表达经常出现在非结构化文本中。因此，共指消解（CO）任务会检测指向同一实体的提及（包括别名和代词）。如果没有其他提及指向某个提及，该提及便是Singleton。给定一些非结构化句子，这类任务会输出共指词跨度对。图 7 展示了这一过程。

4.2.1 基于统计的模型。

早期尝试聚焦于实体、提及和先行词的统计特征，以捕捉共指的语言对象。

基于聚类的解决方案将共指消解任务视为成对二元分类任务（是否共指）。早期聚类模型致力于提及对特征。Soon等人 [148] 提出单链接聚类策略，以检测照应对。Recasens等人 [149] 进一步开发了基于提及对的聚类模型，用于生成共指链或单例叶节点。随后，研究人员将重点放在基于实体的特征上，以利用复杂的照应特征。Rahman和Ng [150] 提出提及排序聚类模型，以深入挖掘实体特性。Stoyanov和Eisner [151]开发凝聚聚类，通过实体特征合并最佳簇。

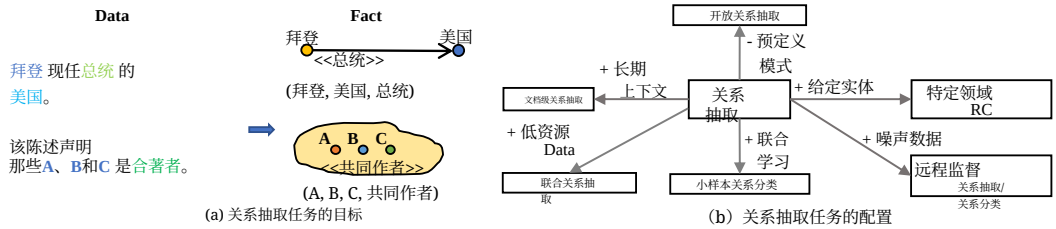


图 9. 关系抽取。关系抽取任务的目标是从数据中抽取事实三元组，并找到用于连接各节点的边。对于 n元关系，连接边是覆盖多个节点的超边。

基于树的模型和基于图的模型也是将共指消解转换为划分任务的常用设计。这些模型根据给定文档构建超图，其中每条边都可以连接两个以上的节点，以建模多个提及之间的共指。Cai和Strube [144]学习统计特征来为边赋权，并通过聚类算法获得共指划分。Sapena等人 [145] 进一步采用松弛标注来解释共指。研究人员还简化了用于共指消解的图，以适应基于树的方法。例如，Bean和Riloff [146] 引入决策树模型，结合上下文特征来区分回指提及。Fernandes等人 [147] 利用投票感知机算法检测提及对共指树。

4.2.2 基于深度学习的模型。

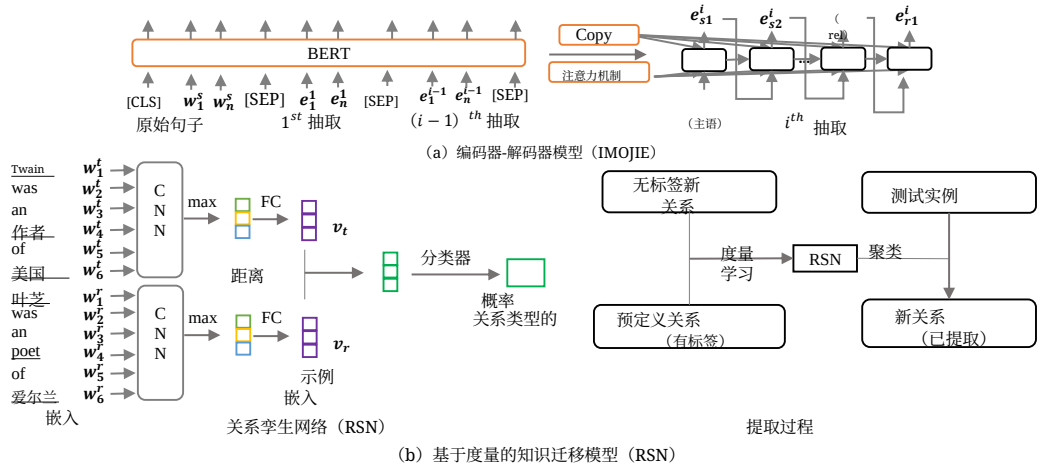
深度学习模型会自动将文档输入转换为词表示，以收集用于检测共指提及对的特征。

许多早期模型基于卷积神经网络。Xi 等人 [152] 通过将远距离特征与层次化提及对特征相结合来解决共指问题，并通过 Softmax 层对提及对进行评分。Wu 等人 [153]开发了一种共指消解模型，利用丰富的多尺度上下文有效处理共指和单例表达，并通过卷积和拼接，将先行词、提及及提及对特征的上下文特征组合纳入模型。

循环神经网络及其变体能够更好地提取词语提及对之间的全局特征。Wiseman 等人 [154] 提出了一种基于循环神经网络的共指消解模型。Lee 等人 [15] 开发了一种端到端的基于长短期记忆网络的模型，通过检测提及跨度内部的依存关系，理解这些跨度周围的全局上下文。Gu 等人 [155] 将簇修正算法应用于长短期记忆网络，以排除不相似对。图8描绘了一个标准的深度端到端共指消解模型。

注意力机制对共指消解任务中的语义交互进行建模。一个很好的例子是 [15] 中介绍的、采用词级注意力增强的双向长短期记忆网络结构。然而，为利用共指特征，人们开发了许多专用于共指消解的不同注意力机制，其中包括：用于共指消解任务的双仿射注意力模型 [156] ，它捕捉词跨度交互以检测关联表达；以及互注意力模型 [157]，它将句法特征与词跨度依存结构和先行词之间的交互特征相结合。此外，Clark 和 Manning 等人 [158] 采用基于强化学习的策略来增强其神经共指消解模型的鲁棒性；该模型使用启发式策略网络过滤错误的共指匹配操作。

研究人员还关注基于嵌入的分布模型在多种语义结构上的应用，以处理共指消解。Durrett 和 Klein [159] 利用先行词表示，通过分布特征实现共指推断。Martschat 和 Strube [160] 探索提及对和树模型上的分布语义，以增强共指表示，并直接选取稳健特征来优化共指消解任务。Chakrabarti 等人 [161] 进一步采用 MapReduce框架，通过查询上下文相似度覆盖回指实体名称。



（a）编码器-解码器模型（IMOJIE）

（b）基于度量的知识迁移模型（RSN）

图 10. 基于深度学习的开放关系抽取的两种范式。在此图中，(a) 展示了 IMOJIE 模型 [187]，它通过编码器—解码器设计抽取事实。(b) 展示了一个模型 [188]，该模型使用 RSN 比较关系模式，然后利用聚类收集关系。

4.3 关系抽取

关系抽取任务从非结构化或半结构化数据中抽取关系事实，以表明实体之间的交互和属性。关系抽取作为一项下游任务，通常也称为关系分类。二元关系抽取抽取实体对之间的关系三元组，而 n元关系抽取则获取多个实体之间的关系三元组，例如合著者。关系抽取赋予知识图谱语义链接。图 9 展示了关系抽取任务的概览。

4.3.1 从半结构化数据与非结构化数据中进行开放关系抽取。

开放关系抽取任务从非结构化数据中发现事实，而无需预定义关系类型。这些技术从自由文本中识别名词性词语（作为主语或宾语）和动词短语（作为谓词），从而构成（主语、谓词、宾语）形式的知识三元组。

统计方法也是开放关系抽取中日益流行的解决方案。在关系检测方面，基于概率图的模型是一种流行设计，它能让上下文信息在半结构化结构或非结构化自由文本中流动。Mulwad 等人[176] 提出了一种应用于半结构化表格的概率图，并采用语义消息传递来标记关系。Chen 和 Cafarella [177] 利用基于条件随机场结构并配有框架查找器的模块，为单元格标记位置标签（如左、中、右）。在此基础上构建分层树，通过父子结构即可恢复关系三元组。研究人员也将概率模型应用于文本的关系分类。例如，StatSnowball [178] 采用马尔可夫逻辑网络来识别关系。

聚焦于规则的方法，是针对不同数据结构类型执行关系抽取任务的最早尝试。这类方法会收集符合人工设计模板的字符串，例如，“$PEOPLE is born in $LOCATION.” 表示 ($PEOPLE, born-in, $LOCATION)。然而，这些无监督策略依赖复杂的语言学知识来标注数据。后来，研究人员开始专注于自动发现模式以挖掘三元组。半监督设计是一种富有启发性的策略，能够减少人工设计的特征和数据标注，并基于少量经过标注的样本发现更可靠的模式，例如 DIPRE [179] 通过种子迭代抽取模式；基于自举的 KnowItAll [7] 和 Snowball [180] 则为 DIPRE 增添了置信度评估。一些基于规则的模型会考虑

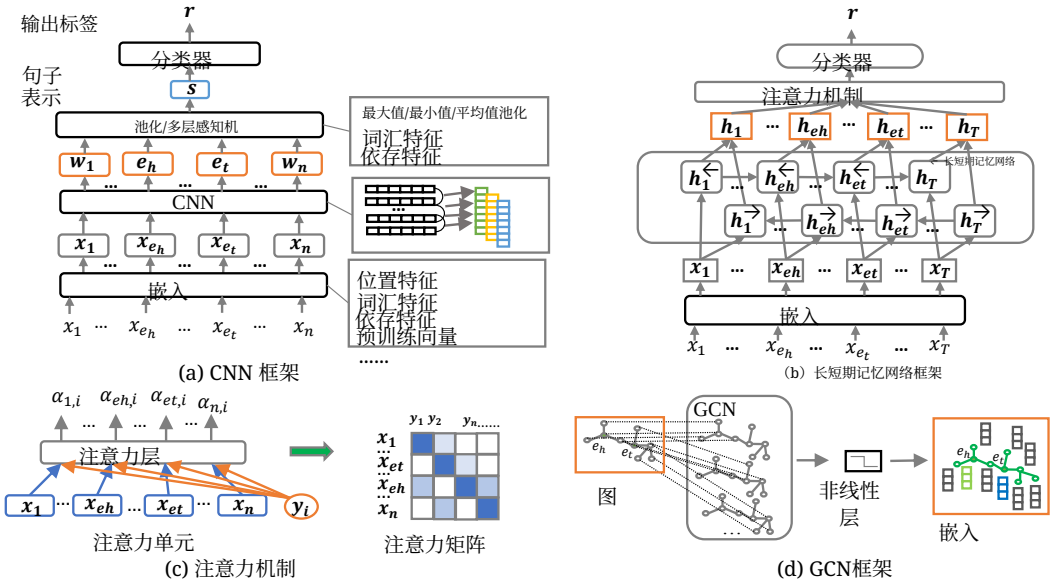


图 11. 经典关系分类模型的一些框架。在该图中，(a)、(b)、(d) 展示了卷积神经网络、长短期记忆网络和图卷积网络的架构，而 (c) 展示了广泛使用的注意力机制的结构。

更多词汇对象用于挖掘。OLLIE [61] 将词汇结构模式与文本中的关系依存路径结合起来。MetaPAD [181] 结合词汇切分与同义词聚类，生成信息充分、频繁且准确的关系三元组元模式。针对半结构化表格，研究人员设计了基于表格结构的规则，以获取排列在行、列和表头中的关系，例如 [182]。此外，一些采用远程监督的半结构化抽取系统能够容忍潜在错误，它们直接查询 DBpedia 和维基百科等外部数据库，以获取表格数据中已发现实体的关系，例如此前的方法 [122], [176],和 [183]。同样，Muñoz 等人[184] 查阅维基百科表格，以标注表格形式呈现的关系。Krause 等人 [185]也通过远程监督扩展了关系抽取的规则集。

深度学习模型也已被开发出来，用于处理开放关系抽取。一种常见框架是旨在获取事实模式的编码器-解码器模型。CopyAttention [186]包含一种机制，可通过神经自举策略将输入序列中的词复制到输出序列中。IMOJIE [187]在 BERT-LSTM 结构的基础上改进了 CopyAttention，同时采用无监督聚合方案进行迭代抽取。开放关系抽取的另一种思路是将监督知识迁移到模型中，使模型能够适应已知关系，从而获得无监督关系。沿着这一思路，Wu 等人 [188] 开发了一种基于度量学习的解决方案，将关系孪生网络（RSN）与聚类策略相结合，以发现新的事实。图10展示了两种深度学习范式。

4.3.2 基于非结构化句子级数据的特定领域关系分类。

对于包含概念（实体）提及的非结构化句子，特定领域关系分类任务会根据句子的上下文，使用预先定义的关系集合为给定提及标注关系标签。核方法和深度学习框架通常将关系分类作为多标签单类分类任务来处理。

基于核的支持向量机方法利用词语的特征向量训练分类器，以对非结构化文本执行监督关系分类任务。这些模型将特定的语义对象映射到

表 4. 经典及近期句子级关系分类模型设计的比较（各类别内按发表年份排列）。

类别	模型	架构	背景信息
局部上下文感知	卷积神经网络 [165]	卷积神经网络 + 最大池化	WordNet 上位词、位置特征
	多核卷积神经网络 [166]	多核卷积神经网络 + 最大池化	位置嵌入
	注意力卷积神经网络 [168]	卷积神经网络 + 词级注意力 + 多层感知机	预训练词向量、位置嵌入、WordNet、词性
	多层次注意力 [169]	词级输入注意力 + 卷积神经网络 + 基于注意力的池化	预训练词向量、位置嵌入
全局上下文感知模型	双向长短期记忆网络 + 注意力 [17]	双向长短期记忆网络 + 词级注意力	预训练词向量
	树结构长短期记忆网络 [167]	双向树结构长短期记忆网络 +复合标签嵌入	SPTree、WordNet、位置嵌入
图上下文感知模型	EPGNN [170]	BERT + 卷积神经网络（句子编码器）/图卷积网络（拓扑编码器）	预定义实体对图、预训练模型、位置嵌入
	AGGCN [171]	图卷积网络 + 多头注意力 + DC +FF	预训练词向量、位置特征、依存图
	RIFRE [172]	BERT + 异构图卷积网络	预训练模型
基于任务转换	问答 [173]	BERT + 跨度预测模型	转换后的问题/答案、预训练模型

用于分类的特征空间，映射过程通过核函数实现，例如基于词汇核的支持向量机（使用词性和实体标签）[162], 基于依存树核的支持向量机 [163], 或基于浅层解析树核的支持向量机 [164]。然而，高性能核函数可能很难设计。

基于深度学习的框架会自动收集实体相关的上下文信息，以完成关系分类任务。给定一个需要进行关系分类的句子 $\{x_1, x_2, x_{eh}, ..., x_{et}, x_n\}$，令 x_{eh} 和 x_{et} 分别表示头实体和尾实体。深度学习模型会为每个词生成一个表示：$\{w1, ..., e_h, ..., e_t, x_n\}$，然后特征提取器会得出一个向量 r，用以表示每种关系类型的概率。基于卷积的模型，例如结合了词汇特征和最大池化策略的特征型卷积神经网络，关注邻近词语的局部上下文。Nguyen 和 Grishman [166] 使用多尺度卷积窗口来增强局部特征聚合。一些研究利用能够捕获长距离依赖关系的长短期记忆网络框架，着重关注句子间的全局上下文感知。Zhou 等人 [17] 使用采用词间注意力的双向长短期记忆网络来捕获关系中的长距离依赖关系，而 Miwa 和 Bansal [167] 则将树结构融入长短期记忆网络框架。许多设计还通过注意力机制将全局上下文特征融入卷积神经网络结构，以建模显著交互，例如，注意力卷积神经网络 [168] 通过词级注意力选择实体相关上下文；多层卷积神经网络 [169] 则采用输入注意力机制和基于注意力的池化。一些较新的研究通过图卷积网络探索图级上下文，并利用预训练模型提取背景知识。这类方法的示例包括 EPGNN [170], , 其为图卷积网络引入实体对图（搭配预训练的 BERT 模型）；AGGCN [171], , 其将多头注意力机制整合到图卷积中；以及 RIFRE [172], , 其进一步采用异构图网络来融合高阶特征。Cohen 等人 [173] 将关系分类转换为问答任务，并引入 BERT 嵌入进行分类。图 11 展示了一些经典的关系分类模型框架，表 4 则比较了热门模型的关键设计方面。

有些任务要求模型处理多个实体之间的 n 元关系。为此，人们设计了语义角色标注方法，将 n 元关系分解为二元关系。示例包括 NNF [174] 和依存路径嵌入 [175]。

4.3.3 远程监督关系抽取/分类。

在大型数据集上进行全监督关系抽取/分类通常需要投入大量人力进行繁重的标签标注。为解决这一问题，Mintz 等人 [45] 提出了一种远程监督策略，利用外部知识库（原始研究中使用的是 Freebase）自动标注关系标签。该策略假设，出现在不同句子中的实体对反映了知识库中连接它们的相同关系。然而，远程监督并未充分考虑数据上下文，因此不可避免地会受到噪声影响。

数据噪声会妨碍传统模型中的关系分类任务。一些方法尝试通过改进特征提取器来克服这一问题，例如分段卷积（PCNN）[16]。这种方法根据

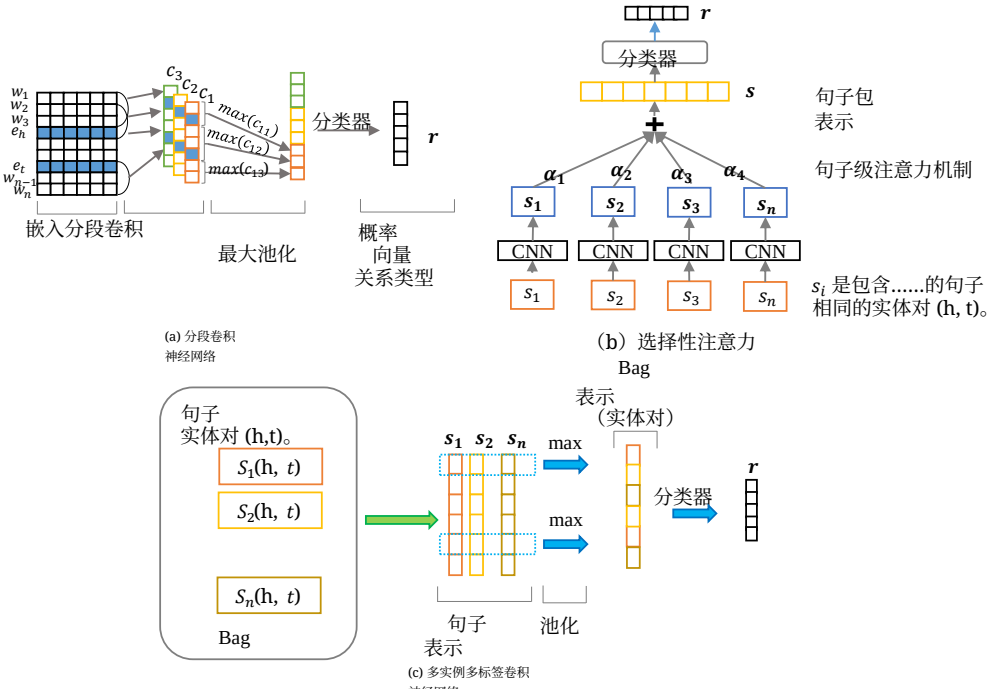


图12. 远程监督的里程碑模型。

实体，以保留关键的上下文特征。另请参见图 12(a)。层次注意力机制[189] 通过建模长尾标签来增强解码器特征。一些模型采用改进的学习策略，旨在提升对噪声的鲁棒性。转移矩阵结构 [190] 通过学习错误模式来抑制噪声。Huang 等人 [191] 采用协同学习来处理交互上下文，而 Qin 等人 [192] 利用强化学习去除错误标记的样本。此外，DSGAN [193] 通过对抗学习筛选可靠样本用于训练。

近来，研究工作主要致力于设计实例选择器结构，以比较样本包中各实例的可靠特征。例如，Riedel 等人 [23] 提出了多实例多标签学习（MIML）的“至少一个”假设。该假设认为，包含相同实体对的样本中，至少有一个会表达给定的远程监督关系（即该样本是正确的）。基于此，选择性注意力 [196] 采用了一种经典设计，将带有相同关系标签的句子分组。另请参见图12（b）。多实例多标签卷积神经网络 [194]使用卷积神经网络处理每个句子包，然后利用跨句池化操作，为多标签关系建模生成实体对表示。参见图12（c）。Ji 等人[195]结合实体描述来增强多实例多标签卷积神经网络。实现实例级特征提取的另一种方法是采用实例级注意力机制。随后，对带有相同关系标签的不同组中的各句子表示进行评分。最后，为每种关系类型生成注意力加权上下文表示。许多模型通过多实例多标签学习设计扩展了这一思路。包内/包间注意力 [197]就是其中一个例子。该方法通过复合注意力机制和跨关系注意力，捕捉内部关系的句子特征以及外部包-关系交互 [198]，其中利用贝叶斯定理获取不同关系类型的包之间的全局相似度。设计实例选择器的主要目的是突出正确样本中的指导性特征，同时抑制错误标注数据中的虚假特征。

表5. 远程监督关系抽取/关系分类常用模型比较。

类别	模型	架构	背景信息
基于增强型特征提取器	分段卷积神经网络 [16]	分段卷积神经网络 + 池化	位置嵌入
	TM [190]	分段卷积神经网络 + 转移矩阵 +包嵌入	位置嵌入
	HAtt [189]	CNN/PCNN + 层次注意力	位置嵌入、关系层次结构
基于增强学习策略	基于强化学习 [192]	基于强化学习的数据重新分配器 + CNN/PCNN +结果驱动奖励	位置嵌入
	DSGAN [193]	生成对抗网络 + CNN/PCNN + 注意力机制	位置嵌入
	CCL-CT [191]	CNN/PCNN + [网络注意力 + 自注意力] + CCL-CT	位置嵌入
基于实例特征	Lin 等人[196]	CNN/PCNN + 选择性注意力 + 最大池化	位置嵌入
	多实例多标签卷积神经网络 [194]	CNN（句子）+跨句最大池化	位置嵌入
	Ye 等人 [197]	CNN/PCNN +包内注意力	预训练模型
	Yuan 等人 [198]	分段卷积神经网络（句子）+ 跨关系跨包选择性注意力	位置嵌入
背景信息增强	多实例多标签卷积神经网络 +ED [195]	描述嵌入 + 多实例多标签卷积神经网络 +	实体描述、位置嵌入
	RESIDE [200]	双向门控循环单元（句子）+句法图卷积神经网络	依存图、外部知识库
	Zhang 等人 [199]	CNN/PCNN（句子）+图卷积网络 + 知识感知注意力	外部知识图谱、位置嵌入

深度学习方法还会考虑外部知识，以改进远程监督关系抽取，例如将实体的知识图谱嵌入引入模型 [199]。RESIDE [200] 进一步利用带有辅助信息的句法图来生成基于图卷积网络的表示。我们在表5中比较了这些具有代表性的成果。

4.3.4 小样本关系分类。

低资源场景，尤其是小样本和零样本关系分类，需要深度学习模型从少量示例中进行学习。少样本/零样本学习，也称为元学习，只需提供少量样本即可驱动深度学习模型，具体包括小样本学习和零样本学习。小样本学习以 N-way K-shot 的形式提供支持数据集，其中每种关系类型有 K 个实例，共涵盖 N 种关系类型（总计 N*K 个样本），并根据给定的支持集预测查询集中的数据标签。进一步而言，零样本学习遵循上述形式，但查询集包含支持集中未出现过的未见样本标签。值得注意的是，在大数据环境中，知识库存在长尾现象，即大多数知识类型仅由少量样本表示。 [201] 元学习配置常见于知识获取和知识精炼的各种子任务中。一般而言，研究人员尝试通过三种方法增强这些低资源配置的可用特征：度量学习、元学习和领域自适应。

元学习通过从有限监督中保留可传递的元信息来增强优化器。模型无关元学习（MAML） [208] 通过两阶段的多次梯度下降改进批量学习。在这种方法中，模型先针对每种关系类型学习估计其梯度，然后所有样本类型再利用这些估计进行常规优化。各关系类型的任务敏感元信息被保存在部分梯度值中。其他模型也采用通过单独反向传播进行梯度估计的方法。MetaNet[209] 利用快慢机制，获取样本与特定任务的高阶隐式关系元特征。元学习器和基学习器都包含一组用于优化的慢权重和快权重。另一个关键问题是灾难性遗忘。为解决这一问题，Wu 等人 [210] 提出了一种课程式元学习策略，按顺序回顾样本，并通过记忆机制保留已学到的特征。

度量学习旨在寻找用于比较不同关系类型的度量空间。为了确定查询样本中的关系类型，ProtoNet [202], 例如，将支持集中的每种关系类型的嵌入进行平均，得到一个原型支持向量。LM-ProtoNet[203]利用关系上下文的细粒度特征构建支持向量，将短语嵌入与通过卷积神经网络诱导的句子嵌入进行拼接。低资源样本中的噪声会吸引无关的元特征，从而降低深度学习模型的鲁棒性。Gao等人 [204]将特征级注意力与实例级注意力相结合，以突出可靠的原型特征。此外，MatchingNetwork [205] 提出了一种基于注意力的嵌入策略，用于

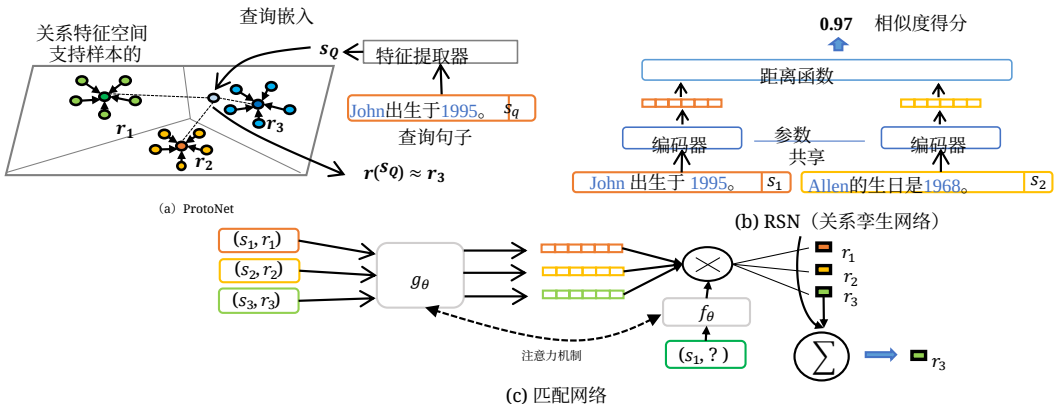


图13. 基于度量的小样本关系分类模型。图中，(a) 展示了 ProtoNet [202]，它比较查询样本与不同关系类型的支持向量之间的距离。(b) 展示了 RSN [206]，它计算样本对之间的相似度。(c) 展示了匹配网络 [205]，通过将查询样本与具有不同标签的支持样本进行匹配，使用注意力机制为查询样本标注关系标签。

分类，通过向量乘法计算查询样本相对于每个支持样本的注意力分数。RSN [206] 比较样本嵌入的相似性。另一种称为多层匹配与聚合网络（MLMAN）的方法 [207], 通过将支持向量与查询向量进行聚合来聚合局部特征和实例特征，从而为查询样本匹配正确的长尾类别标签。图13 概述了经典的基于度量的范式。

小样本关系分类的设计还会考虑特征增强策略，通过新颖的模型设计和背景知识来缓解数据匮乏。与 [173], Levy等人 [211] 类似，他们借助模板转换器将零样本关系分类转化为阅读理解问题，以理解未见标签。Soares等人 [212] 利用实体对的 BERT 上下文化嵌入及对应句子，为每个句子构建复合关系表示。图卷积网络还为小样本学习提供额外的图级特征。Satorras和Estrach [213] 提出了一种新颖的 GCN框架，通过计算节点间的相似度来确定查询样本的关系标签。此外，Qu等人 [214] 将后验分布用于原型向量。一些设计还可基于度量学习进行半监督数据增强。此前的 Neural Snowball [206] （基于 RSN）使用孪生网络标注查询集，同时从外部远程监督样本集中选取相似的样本候选项，以扩充支持集。

小样本领域适应将未见标签映射到分类任务中。BERT-PAIR 与领域适应策略相结合，用于处理未见的“以上皆非”类型。Gao等人 [215] 将小样本关系分类中的领域自适应视为一种搜索领域不变特征的博弈过程。他们通过对抗训练实现领域自适应。更多关于小样本关系分类的领域适应策略可见 [24]。

4.3.5 联合关系抽取模型。

传统的基于流水线的关系抽取（关系分类）模型会在各个阶段遭遇误差传播，同时削弱任务间交互。早期研究者致力于探索有趣的基于统计的特征，以实现快速的端到端联合关系抽取，例如基于整数线性规划（ILP）的算法 [216]通过条件概率模型求解实体和关系、半马尔可夫链模型 [217] 联合解码全局级关系特征，以及马尔可夫逻辑网络（MLN） [218] 对实体标签和关系的联合逻辑规则进行建模。早期尝试为实体-关系交互提供了原型。然而，对于复杂语境，统计模式并不明确。研究者因此转而

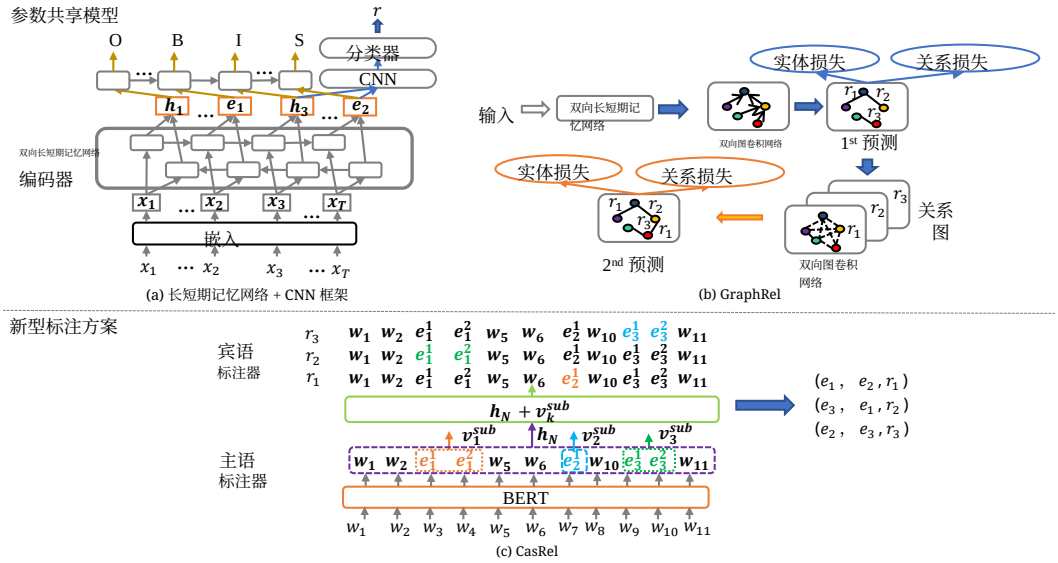


图 14。 联合抽取模型范式。(a) 描绘了 [219], (b) 描绘了 [220] 而 (c) 描绘了 [222] .

采用联合抽取模型，其中主流的高性能设计侧重于参数共享策略和新颖的标注模式。

参数共享策略将不同任务类型的神经网络架构整合起来。它们共享权重，并使用不同的输出层来提取存在关系的实体。Zheng 等人 [219]将用于命名实体识别和关系分类任务的双向长短期记忆网络层合并以共享参数，然后分别使用卷积神经网络和 LSTM网络标注关系和实体。Miwa 和 Bansal [167] 也将与命名实体识别和关系分类相关的依存特征，与双向长短期记忆网络层和双向树结构长短期记忆网络层相结合。一些模型专注于精细的策略，以分配跨任务特征。例如，GraphRel 模型 [220], 直观地利用两阶段监督，通过两个相应的 BiGCN 层实现跨任务交互。GCN框架将依存图与关系-实体图结合，以挖掘深层特征。

为处理重叠标签，新型标注方案通过复合标注，为输出层设定联合解码目标。Zheng 等人 [221] 将 BIES标签扩展为词语的关系类型和角色（例如，句子的主语或宾语），从而构建一个结合命名实体识别与关系分类的序列标注任务。Wei 等人 [222] 直观地针对每种关系类型，通过级联映射函数标注主体实体的所有候选客体，以涵盖重叠提及。此外，Wang 等人 [223] 提出了一种握手方案，以缓解重叠实体中的曝光偏差。Bekoulis 等人 [224] 设计了一种多头选择机制，以探索所有实体/关系组合。Li 等人 [225] 将实体-关系标注转化为多轮问答问题，利用机器阅读理解（MRC）模型捕捉实体间的长程语义。与以往方案不同，KGGen [226]基于预训练模型并结合对抗学习，通过编码器-解码器/生成器结构直接生成三元组，从而克服对实体共现信息的特征依赖。图 14 展示了一些具有开创性的联合抽取模型。

此外，还提出了新颖的基于分布嵌入的模型，用于对跨任务分布进行建模，以弥合命名实体识别与关系分类之间的语义鸿沟。Ren 等人 [227] 提出了一种知识增强的分布式 CoType 模型，用于联合抽取任务。在该模型中，首先将实体对映射到知识库中的对应提及，然后为其标注实体类型以及所有关系

表6. 文档级关系抽取模型设计比较（各类别内按发表年份排列）。

类别	模型	词编码器	长上下文编码器	推理	跨句特征
统计图方法	图LSTM[229]	嵌入	图长短期记忆网络	Softmax	根链接跨句依存树
	图状态LSTM [230]	FFN	图状态LSTM	Softmax	根链接跨句依存树
	AGGCN [171]	LSTM	图卷积网络 + 多头注意力	DC + Softmax	根链接跨句依存树
	Sahu等人 [231]	嵌入	GCN	基于多实例学习	共指/相邻句子边
动态基于图的模型	GP-GNN [233]	双向长短期记忆网络	GCNs	多层感知机 + Softmax	带生成参数的节点间图
	EoG [232]	双向长短期记忆网络	GCNN	节点特征聚合 + Softmax	句子-提及-实体对图
	GraphRel [220]	双向长短期记忆网络	BiGCN	基于阈值的	关系加权图
	DyGIE [238]	ELMO + 双向长短期记忆网络	图卷积网络 + 跨度枚举	FFN	动态跨度图
	LSR [234]	双向长短期记忆网络/双向编码器表示器	GCN	前馈神经网络 + 图卷积网络 + DC	加权依存图
	GAIN [236]	LSTM	GCN	前馈神经网络 + 注意力机制	hMG + EG
	Xu 等人 [235]	双向长短期记忆网络	AGGCN [171]	长短期记忆网络 +Softmax	重构的异构 S-M-E 图 [232]
	DRN [239]	双向长短期记忆网络	GAIN [236]	聚合 +多层感知机	异构文档级元路径
	RARE [237]	BERT	关系图卷积网络	多层感知机 + Softmax	推理依据图、预训练模型
其他	ATLOP [242]	BERT	局部上下文池化	群双线性	预训练模型
	U-Net [241]	BERT	二维卷积	基于矩阵	特征可视化

候选项由知识库提供。该模型利用上下文化的词汇和句法特征学习关系提及的嵌入，同时在训练实体提及嵌入时纳入其类型；随后，通过翻译嵌入（TranE），根据头实体和尾实体的嵌入推导出上下文化的关系提及 [228] 模型。CoType 模型假设实体与其关系标签之间存在交互式共现，并利用外部领域的知识和额外的类型特征来弥补分布差异。值得注意的是，该模型还能有效抑制远程监督数据集中的噪声。不过，该模型还需要特征工程和额外的知识库。

4.3.6 文档级关系抽取。

文档中的实体可以通过复杂的跨句上下文表达关系，这使大多数传统的句子级上下文编码器难以应对。因此，人们提出了新型架构来捕获文档级上下文。

句内语义段落对于文档级抽取至关重要。因此，研究人员最初开发了多种将图结构融入长短期记忆网络的变体，以处理长期依赖关系，例如图LSTM [229] 和图状态LSTM [230]。然而，近年来，研究人员开始关注基于GCN的模型，以利用新型的跨句图结构探索多样的语言学特征。许多方法使用静态文档图处理句间语义上下文。例如，对于 *n*元关系抽取，AGGCN [171] 通过注意力引导的GCN层连接相邻句子依存树的根节点，同时克服了对语义角色标注的依赖。Sahu等人 [231] 引入共指边和相邻词边，形成同质文档图。Christopoulou等人 [232]采用提及/句子/实体（M、S、E）节点来创建异质语义图，以区分各种语言学角色，并利用上述中间节点结构，通过EoG推理层进行推理。

随后，研究人员开发了用于高阶推理的动态文档图模型。许多模型利用动态边。GP-GNNs [233] 在全连接图中利用动态边权重进行推理，从而推断隐藏语义逻辑。LSR [234] 将图结构视为潜变量，通过迭代优化链接和权重，从上下文中构建逻辑特征。Xu 等人 [235] 考虑重构依存路径，以重新赋予关系实体对权重。GraphRel[220] 通过两阶段过程联合抽取实体和关系，将静态依赖与动态关系加权图相结合，以增强多跳推理。一些模型还考虑利用多个图进行特征提取。例如，Zeng 等人 [236] 设计了一个异构提及级图，并配合一个实体级图进行多跳推断。

文档级关系抽取的另一种方向是基于证据进行推理。Zhang 等人[237] 构建了一种推理依据图，并利用外部标注的共现证据特征来捕捉长期关系依赖。人们还研究了具有可变节点的动态图

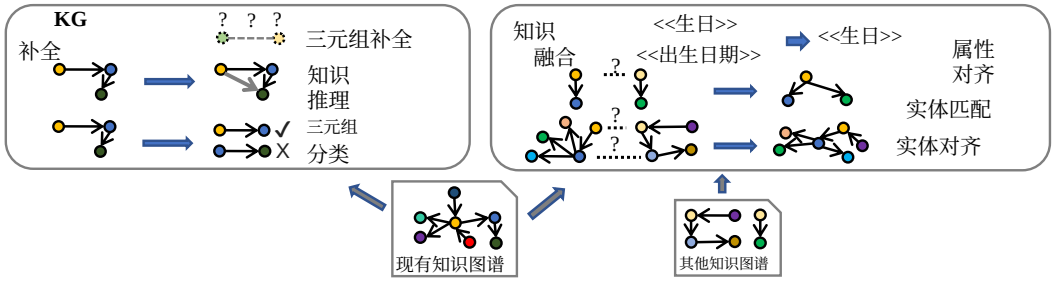


图15. 知识精炼示意图。

复杂推理领域中的相关研究。DyGIE [238] 通过门控机制剪除置信度低的实体跨度节点，以实现文档级特征利用。其他方法则试图理解不同上下文中隐含的常识规则。判别式推理网络（DRN） [239] 利用异构图表示特征，在进行句内推理的同时识别常识关系。该方法基于这样一种假设：包含句法结构的多尺度上下文中蕴含着可区分的常识特征。还可以从COMET [240]等预训练模型中获取背景常识特征。然而，理解常识如何在上下文中表达，以及它如何在人类逻辑中展开，仍然颇具挑战。

U-Net [241] 通过多层卷积实现 U 形分割，以进行文档级推理。此外，它将文档视为视觉语义信息。ATLOP [242] 引入局部上下文池化，以提取 BERT 的实体相关特征，同时使用自适应阈值解码可靠的关系。这种技术不依赖图结构。我们在表格 6中比较了典型里程碑模型的设计。

5 基于结构化数据的知识图谱精炼

从非结构化或半结构化数据构建的原始知识图谱可能较为稀疏，知识三元组也可能不完整或遭到破坏。知识图谱精炼通过背景语义修复这些问题，或利用额外的知识图谱（结构化数据）补充知识三元组。知识图谱精炼的子任务包括知识图谱补全和知识融合。总体流程如图15所示。

5.1 知识图谱补全

知识图谱补全填补不完整的三元组，并从已补全的三元组中推导出新三元组。对于已补全的三元组，知识图谱补全通过三元组分类评估每个三元组的可信度。这里的可信度指三元组的正确性。

5.1.1 基于嵌入的三元组补全。

基于嵌入的链接预测模型利用分布式表示，搜索能够填补以 (h, ?, t) 或 (?, r, t)（实体预测）以及 (h, ?, t)（关系预测）形式表示的缺失部分的元素。例如，基于 TransE 的模型 [112] 搜索头实体 h、尾实体 t 和关系 r，其表示逐渐接近 $\mathbf{h}+\mathbf{r}=\mathbf{t}$，从而补全三元组。后来，研究人员发现，先前对称的 TransE 模型并未考虑一对多关系。因此，研究重点转向引入基于距离的平移模型中的超空间结构来进行链接预测，例如 TransR [243], TransH [244], 和 TranSparse [245]。一些模型，如 RESCAL [246], TuckerER [247], DistMult [248] 和 NTN [249], ，则通过将实体对表示匹配到潜在关系语义空间中，在大型图上进行预测。

近年来，研究人员的关注重点转向语义知识结构。例如，HAKE [250],利用极坐标系对知识图谱中的语义层级进行建模，例如上位词、下位词以及实体本体关联的同位关系，并通过模长和角度约束区分不同层级的实体向量。CAKE [251] 利用常识规则增强负采样。许多模型，如 SimKGC [252] 和 HaLE [253],通过对比学习优化低维嵌入的负采样。CAFE [254]引入邻域子图特征集，以增强相关链接信息。此外，研究人员也开始借助图卷积网络的语义扩散机制，利用知识图谱的子结构分解知识表示的语义组成部分。DisenKGAT[255] 通过将表示组件解耦为子图结构中隐含的不同语义，辨别知识图谱中高阶邻居节点特征。这些模型背后的假设是，大型知识图谱应包含足够多的子集，可将其缩减为 k 个组件，以推理相连的实体节点。

5.1.2 关系路径推理。

关系路径推理通过完整的三元组序列推导新事实，并将其作为支持证据，例如“(B, 居住于, 西雅图)←(A, 就职于, 微软), (微软, 位于, 西雅图)”。

早期研究尝试开发用于关系路径推理的随机游走模型，在潜变量逻辑图模型中推断关系逻辑路径。路径排序算法 (PRA) [256] 生成特征矩阵，以采样潜在的关系路径。然而，图中的特征稀疏性阻碍了随机游走方法的应用。为缓解这一瓶颈，研究者提出了语义增强策略，例如利用向量空间相似度 [257] 并对关联关系进行聚类 [258]。

后来，研究者将关系路径推理任务建模为马尔可夫决策过程，以识别知识环境中的逻辑约束。深度强化学习通过学习一个策略智能体来实现这一思路，该智能体评估每一步选择并扩展推理路径。DeepPath [259] 将状态空间建模为实体及其诱导关系的（预训练）基于平移的表示。随后执行的动作通过实体对的特征空间寻找最匹配的关系标签。动作的奖励通过二元函数计算。然而，二元奖励函数给出的低质量评估意味着基于强化学习的模型无法很好地泛化到不完整知识结构的处理上 [260]。为此，Lin 等人 [260]设计了一个基于关系和实体向量空间的软奖励塑形函数，而 Li 等人 [261] 则采用多智能体来选择实体对和关系。M-Walk[262] 利用循环神经网络捕捉路径决策之间的时间序列状态依赖关系。

更多设计采用能够捕获全局特征的神经网络来寻找合理路径。Path-RNN [263] 递归聚合关系路径特征，以进行多跳推理。chains-of-reason 模型 [264] 在 Path-RNN 中加入注意力机制，以突出实体类型信息所涉及的多路径依赖关系。Chen 等人 [265] 通过变分编码统一了路径推理和路径查找任务。

一些方法进一步关注注意力机制，以增强强化学习中的特征。ADRL [266] 利用自注意力机制，突出邻域实体—关系交互特征。类似地，Wang 等人 [267] 引入图注意力机制，以增强知识特征。近期的研究兴趣转向融入能够处理复杂语义特征的神经结构，例如 Zheng 等人的分层策略网络[268] 和 DAPath [269], , 后者采用距离感知机制，通过路径长度特征提供奖励。MemoryPath [270] 是一种基于注意力的记忆组件，可为强化学习保留知识特征，并减轻模型对预训练嵌入的依赖。

许多研究也致力于自动挖掘逻辑规则，以开辟推理路径。已有一些规则发现方法，例如 AMIE [271], RLvLR [272] 和 RuleN [273]。规则挖掘方法并不搜索接近知识符号本质的潜在关系路径模式，而是从合理的知识图谱结构中提取并筛选逻辑规则，然后利用收集到的规则模板进行链接预测。然而，在不完整图中，逻辑规则难以推导出未见过的知识路径。

另一个研究方向是将逻辑规则融入神经模型，以增强路径推理。KALE [274]将一阶逻辑规则与知识嵌入联合嵌入，从而增强关系推断。RUGE [275] 通过学习到的软规则迭代修正知识图谱嵌入，然后执行关系路径推理。逻辑规则也被用作神经模型的辅助语义信息。NeuralLP [276] 提出了一种神经框架，利用注意力机制将逻辑规则结构编码为向量化嵌入。pLogicNet [277] 引入马尔可夫逻辑网络，对用于推理的不确定规则进行建模。ExpressGNN [278] 进一步采用图卷积神经网络，结合逻辑规则处理邻域图语义。这些基于规则的神经模型也被视为可微学习的应用，可用于对逻辑编程采用基于梯度的优化算法。

5.1.3 可解释的关系推理。

可解释性旨在让人类用户能够理解机器学习模型 [279]，这对于评估模型的可靠性以及其应对不同数据环境的能力至关重要。解释模型包括自解释的事前模型和可检查的事后模型。

由透明决策过程构成的事前推理模型，可以通过内省其内部结构来实现自我解释。AMIE [271]和 RLvLR [272] 等逻辑规则挖掘方法能够反馈逻辑规则，向用户解释链接决策。一些模型只包含部分可供人类理解的组件（例如，学到的规则）。在使用基于神经模型的规则发现方法（如 NeuralLP [276], 、pLogicNet [277] 和 ExpressGNN [278]）进行推理时，用户可以将学到的规则作为辅助信息来查看。然而，这些神经网络仍然是黑箱。主流的部分事前模型还包括基于随机游走（潜在路径的概率值）、强化学习（每个动作的奖励值）和注意力机制（用于显著相关性的注意力分数）的模型。

事后解释方法会构建代理模型，以探查黑箱模型中的隐式特征，例如矩阵和神经网络框架。一些代理模型会提取规则，或学习概率分布来复现模型。Carmona 等人 [280] 将贝叶斯网络与一阶逻辑结合训练，从嵌入模型中提取规则。OXKBC [281] 通过关系和实体之间的相似性生成合理的解释路径。模型简化无法拆解相互交织的非线性神经模型特征。一种解决方法是进行敏感性分析，以挖掘深层特征。分析过程中会对模型施加微小扰动，以观察输出如何变化。这些变化揭示了影响特征。GNNExplainer [282] 探索会影响单实例和多实例预测的子图结构。CRIAGE [283] 生成虚假事实，以评估模型性能，并定位每种关系中具有干扰性的事实三元组。

5.1.4 三元组分类。

三元组分类旨在从知识图谱中区分确定的三元组与异常（不真实）的三元组。许多语义模型专为此任务而设计，用于判断知识图谱中可疑的三元组，而该图谱会不断更新，纳入新的关系类型和事实。

负三元组样本赋予知识表示模型判断无序三元组的表达能力。例如，CKRL [284], 包含一个用于确定可靠三元组的索引系统，其中包括

局部三元组置信度，用于比较三元组与负样本之间的距离；全局路径置信度，用于检验构成三元组的推理路径的全局资源；以及自适应路径置信度，用于评估推导出三元组的局部推理路径。

然而，由于负采样不足，许多可能合理的三元组未能被涵盖，具体而言，涉及一对多关系的三元组 [35]。因此，研究人员利用更复杂的语义结构来缓解这一问题。沿着这一思路，Dong 等人 [285] 将实体嵌入扩展为 n 球结构，并利用这些结构纳入细粒度类型链，以此对三元组进行分类。Amador-Domínguez 等人 [286]则加入本体信息，以增强模型无关的表达能力。一些模型专注于先进的神经网络嵌入，以检测可信的三元组。例如，R-MeN [287] 采用多头注意力机制生成基于记忆的嵌入，从而捕捉三元组之间的潜在依赖关系。

5.2 知识融合

现实世界中的知识通常是开放且可更新的。在大多数场景下，用户应能够添加外部知识，以丰富现有的外部知识图谱。为此，知识融合旨在合并语义等价的元素，例如“特朗普”和“唐纳德·特朗普”，从而将新知识整合到新的概念或事实中。知识融合的子任务包括属性对齐、小规模输入三元组的实体匹配，以及与完整知识图谱的实体对齐。

5.2.1 属性对齐。

属性三元组表示概念的某项属性，其描述值可以是颜色、日期、数字或字符串。用户可能会使用不同术语指代同一属性，例如“生日”和“出生日期”，而同义词可能导致语义稀疏。因此，属性对齐旨在统一属性记法。

许多方法专注于对齐属性的语义嵌入，其前提是，如果两个属性名称的嵌入彼此接近，那么它们就应当相同。一些模型利用属性名称字符串之间的相似性来生成分布式嵌入，例如 [288] 和 [289]。Yang 等人 [290] 利用词袋模型学习属性的上下文嵌入。类似地，JAPE [291] 利用 Skip-gram 模型进行属性嵌入，以建模经常一起使用来描述实体的共现属性，例如描述位置的“纬度”和“经度”。

属性嵌入也为实体对齐任务提供辅助信息，例如定义和描述。然而，属性也可能包含信息量不大的数据，例如电话号码，这会给生成知识级表示带来挑战。因此，一些模型会考虑使用神经网络，根据上下文值生成嵌入。例如，AttrE [292] 使用长短期记忆网络框架为属性值中的每个字符生成嵌入，以组合成属性嵌入，从而预测单语表达中的潜在短语。该方法采用属性名称谓词对齐策略来处理未见过的属性。

5.2.2 小规模知识图谱中的实体匹配。

在其输入包含一对小型三元组集合（通常定义为 1 到 10,000 个三元组）的场景下，实体匹配旨在确定这两个知识图谱中的哪些实体指代现实世界中同一个对象。这一任务在文献中也被称为实体解析、记录链接、实体识别或实体同一性问题。

早期的一些研究将实体匹配视为一个二元分类问题，即判断一对实体是否匹配。这些方法通常依赖于手工特征工程，以及传统的机器学习分类器，如决策树、支持向量机或朴素贝叶斯。例如，Nikolov 等人 [293] 提出了一种基于主动学习的实体匹配方法，该方法通过迭代地向人工标注者请求最具信息量的样本来减少所需的训练数据量。Isele 和 Bizer [294] 则学习由多个相似度函数组成的线性组合，以确定两个实体是否匹配。

近年来，研究重点已转向利用深度学习技术来提升匹配性能。这些方法通常将实体匹配视为一个序列分类问题，即将一对实体的描述作为输入，并预测它们是否匹配。例如，DeepMatcher [295] 是一个

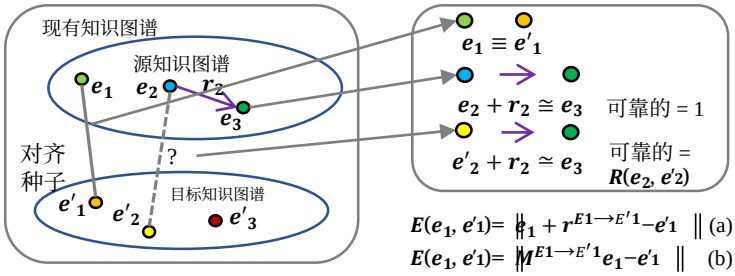


图16. IPtranE [308]. IPtranE 通过（a）平移模型和（b）线性变换模型为实体对评分，并通过硬对齐或软对齐合并相同的实体对。

深度学习系统，它采用结合注意力机制的循环神经网络结构来表示实体匹配中的属性词。与传统模型相比，基于深度学习的模型更擅长处理文本中的噪声，尤其是结合 WordNet 的概念丰富任务 [296] 。

早期研究也致力于利用实体的独特属性进行实体匹配。许多模型采用基于距离的方法，处理实体描述或定义的分布式表示。VCU [297] 提出一阶和二阶向量模型，将实体对的描述词嵌入其中，以全面衡量概念距离。TALN [298]利用基于 BabelNet 得出的基于词义的嵌入，结合词语的定义性描述：首先结合词性标注器、句法特征和 BabelNet，为每个筛选出的定义词生成嵌入，然后取其平均值，得到词义质心，以找出最佳匹配候选项。可用于实体匹配的基于字符串相似度的模型还包括 TF-IDF [299],I-Sub [300]。

基于图的方法在小规模知识图谱的实体匹配任务上取得了可行的性能，这类知识图谱可以由层次化图结构构成。ETF [301]通过语义特征和基于图的特征学习概念表示，这些特征包括 Katz相似度、随机游走介数中心性和信息传播分数。ParGenFS [302] 利用基于图的模糊聚类算法将新实体概念化。该方法模拟主题分布，以获得具有区分性的概念簇，从而在目标知识图谱中搜索实体更新的对应位置。

在权衡性能与计算成本时，也可以采用基于文本相似度的模型，通过检测实体表层相似性来处理实体匹配任务。Rdf-ai [303] 提出了一种系统化模型，用于匹配两个实体节点图；该模型利用字符串匹配和词汇特征相似度比较算法对齐可用属性，然后计算实体相似度以实现对齐。类似地，Lime [304] 进一步利用度量空间检测对齐的实体对：首先生成实体样本以筛选可对齐的候选项，然后计算相似度以进行实体融合。与小规模知识图谱不同，形态各异的大型知识图谱包含有意义的关系路径和丰富的概念分类体系。HolisticEM[305] 利用逆文档频率分数计算实体名称的表层相似度，以生成种子；并利用个性化PageRank（PPR）分别遍历各自的邻居节点，衡量实体图之间的距离。

自治社区可能会向知识图谱系统输入独特的信息，例如昵称、电话号码和其他个性化数据。这些知识只有用户才知道。因此，需要采用策略检测缺失的独特信息，并请用户补充。通过向用户提问来判断信息或解决冲突的主动学习方法 [306] 是最可靠的解决方案，在这些场景中不可或缺。

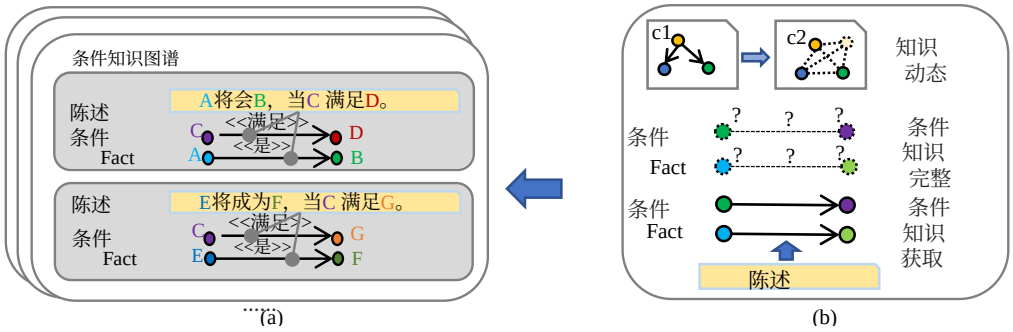


图 17. 知识演化。图 (b) 所示的演化分析任务将数据整理成知识图谱组（条件知识图谱或事实知识图谱），如图 (a) 所示，以呈现动态条件下的知识。

5.2.3 与大规模知识图谱进行实体对齐。

大规模知识图谱通常包含丰富的属性信息和图结构，这些内容能够形成包含概念实体和关系链接的知识感知结构。实体对齐任务旨在将结构化数据与构建完善的大规模知识图谱相整合，这些知识图谱在知识层面包含语义结构。

基于嵌入的模型通过拥有三元组知识嵌入的种子实体，学习用于实体对齐任务的图间实体映射。Sun 等人 [307] 指出，用于链接预测的常规负样本会削弱区分同一类型不同实体的能力。因此，他们利用特征空间中靠近相应目标实体的实体来生成负样本。IPTransE [308] 是一种用于知识表示和实体映射学习的迭代式联合嵌入策略。它采用路径平移嵌入方法，对连接同一实体对的不同关系路径进行嵌入，并将其视为具有相同效果的链接。随后使用软对齐策略来减少匹配错误。参见图 16。MultiEA [309] 考虑实体图属性、链接和邻居节点的多视图特征。BootEA [307] 采用一种自举式“可能对齐”标注算法，通过迭代添加可靠的种子来实现对齐。在跨语言场景中，MtransE [310] 生成坐标轴校准和平移向量，以建模不同语言中特征空间的不变性。此外，一些模型会采用自监督策略来利用种子信息，例如 SS-AGA [311] 和 SelfKG [312]。

实体对齐面临的一项关键挑战是，许多实体不具备表面特征或结构分布特征。因此，许多实体对齐模型还会使用属性表示来增强特征。例如，KDCoE [313], 利用带有描述属性的协同训练策略。JarKA [314]对稀疏多语言知识图谱中的属性之间的交互进行建模，以推断等价实体。一些模型利用基于深度学习的神经网络来生成属性上下文嵌入。例如，AttrE [292] 利用长短期记忆网络提取属性值的依存特征。不同于以往方法，JAPE [291] 将属性嵌入与叠加关系图结构相结合，以捕捉跨语言差异。

另一个具有挑战性的问题是用于对齐的语义图结构。GCN-Align [315] 首次提出了一个基于GCN的框架，用于实体对齐任务。此后，近期研究开始利用基于GCN的模型，聚焦于复杂的图语义。例如，RNM [316]匹配邻域节点特征，以比较实体对。RDGCN [317] 利用双关系图，解决三角实体图结构中的矛盾表示。

大规模知识图谱通常包含鲜明的语义子图结构，可用于对齐。这里，图匹配神经网络（GMNN） [318]构建了一个主题实体图，

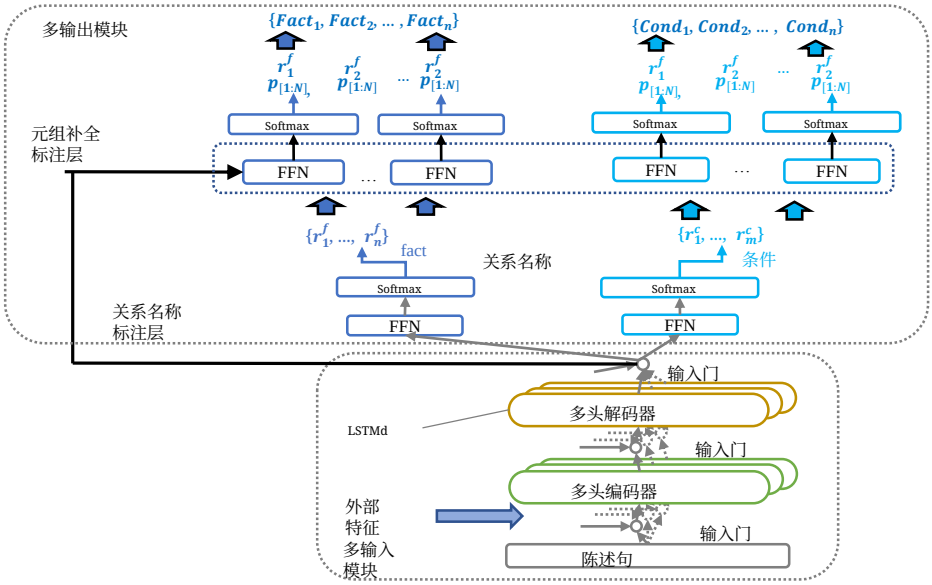


图 18. MIMO模型的架构 [26] 用于从文本中提取带条件的事实 s.

将相邻节点链接起来，以合并相同的实体。AttrGNN [319] 根据属性三元组类型划分知识图谱，以理解异构实体信息。

近期的研究方向也致力于对跨图交互进行建模。例如，MuGNN [320], 提出了一种跨知识图谱注意力机制，并采用多通道图神经网络编码器，能够一致地建模图间结构特征。类似地，GTEA [321] 采用联合图注意力机制来融合跨图关系信息。

6 知识演化

近期，研究人员关注知识如何在环境条件下演化。条件知识图谱通过反映在特定条件下确立的事实来实现这一目标。条件元组表示为 (h, r, t, γ)，其中 γ可以是某个事实的前置三元组。许多研究人员在一种简化情形下对此进行了研究，即时序知识图谱，其中 γ 表示某种时间信息（如时间戳），例如 (拜登, 职务, 副总统, 2009年-2017年)、(拜登, 职务, 总统, 2020年-）。图 17 展示了知识演化示意图。

6.1 条件知识获取

许多科学事实都是在特定条件下确立的，尤其是在生物医学领域。早期研究并未从系统性的角度全面考虑这一情况。因此，Jiang等人 [322] 开发了一种新的标注方案，用“B/I-XYZ”格式表示条件元组，其中“BI”表示位置信息（起始/中间），“X”表示逻辑角色（事实/条件），“Y”表示元组角色（主语/宾语），“Z”表示组成部分类型（概念/属性/谓词）。条件知识抽取旨在实现三个目标：抽取事实元组、收集条件元组，并关联事实与条件。Jiang等人 [26] 指出，传统抽取系统会将条件信息合并到实体中，形成事实三元组，这会损害实体链接效果。此外，在非结构化陈述中，同一词元可能在不同元组中既充当主语，也充当宾语。因此，他们提出了一种基于多输入多输出序列标注（MIMO）的联合抽取方法来解决这一问题。

他们的MIMO模型利用关系名称标注层，通过事实和条件标注子层分别为每个词元标注关系标签。随后，使用元组补全标注层来区分每个词元在不同关系名称下的逻辑角色。然而，Zheng等人 [323] 指出，MIMO标注模式无法有效处理重叠三元组。因此，他们利用层次解析，将MIMO模型中的多输出模式简化为单输出模式。图18 展示了MIMO模型。

条件知识抽取中的另一种热门趋势是时间知识抽取，它将条件三元组简化为时间。许多早期模型利用循环神经网络结构捕捉时间依赖关系，从而识别句子中的时间关系，例如 [324] 和 [325]。在抽取细粒度时间知识方面，Vashishtha等人[326] 通过多个堆叠注意力层对事件、状态和持续时间进行建模，以匹配其时间线。近期研究改进了文档级时间知识抽取的解决方案。例如，TIMERS [327] 是一种修辞感知图，可供GCN模型通过文档中的表达来解读复杂且连贯的“基本话语单元”。基本话语单元是参与时间活动的最小语义单位。

6.2 条件知识图谱补全

条件知识图谱补全任务旨在补全条件知识图谱中的不完整三元组，例如 (h, ?, t, γ)、(h, r, ?, γ) 和 (h, r, t, ?)。请注意，本节主要关注时序知识图谱补全方法。

研究人员可以通过时间信息嵌入模型预测不完整的时间元组。TTransE [328] 在TransE 的基础上扩展了时间嵌入向量。HyTE [329] 将时间戳视为超平面，以匹配实体嵌入和关系嵌入。另一个前景广阔的方向是时间感知嵌入。在这一方向中，基于LSTM的模型 [330] 对时间编码序列进行解释，而基于CNN的模型 [331] 则捕捉上下文的时间一致性。

时序知识图谱表示可以视为沿时间维度展开的张量结构，这意味着可以使用张量分解来补全时序知识图谱。张量分解的主要方法包括典范多重分解和Tucker分解。典范多重分解使用若干一阶张量之和来逼近目标张量。许多时序知识图谱补全模型都采用典范多重分解，例如 [332] 和 [333]。Tucker分解将目标张量分解为核张量与沿目标张量各维度展开的多个张量的乘积。Shao等人 [334] 开发了一种基于Tucker分解的模型，用于解释时间语义关联，从而提高了包含时间戳的表示的灵活性。SpliMe [335] 通过静态模型获得时间视角实体嵌入。

时序知识图谱补全的另一个关键主题是时间知识推理。近期研究主要聚焦于基于GCN的方法。在此方面，Han 等人 [25]通过基于边注意力分数扩展查询依赖干扰子图，利用历史上下文。Jung等人 [336] 通过基于边的注意力传播实现多跳时间推理，而 Liu 等人 [337] 则通过基于强化学习的模型增强时序知识图谱推理。此外，时间线中的事实不能忽视时间依赖关系，例如“出生于”应先于“就职于”。Jiang等人 [338] 定义了一种包含非对称矩阵的评分函数，以保留用于推理的时间顺序约束。

填补不完整的一般条件元组仍有待进一步探索。元组可能包含多个条件，例如仅在特定温度范围内发生的化学反应。应针对这些复杂场景提出系统性解决方案。我们还建议读者考虑因果关系发现方法 [339]。

Text	Mary和Henry 计划去旅行。他们上周去了英格兰。Mary 认为埃文河是最浪漫的景致，位于英格兰，而Henry 在四处漫步时，为自己是莎士比亚的同乡而感到荣幸家乡。他们确信女儿Lily也会喜欢这个地方。	
推理	Fact	详情
共指推理	(Mary，实体-目的地，英格兰) (Henry, 实体-目的地, 英格兰)	Mary和Henry 计划去旅行。他们游览了英格兰.....
模式识别	(埃文河, 部分属于, 英格兰)	$X 是.....中的景点
逻辑推理	(Henry, 国籍, 英格兰)	(Henry，莎士比亚的同胞，莎士比亚) (莎士比亚，国籍，英格兰) ➡ (Henry，国籍，英格兰)
常识推理	(Mary, 伴侣, Henry)	(Mary，女儿，Lily) (Henry，女儿，Lily) ➡ (Mary，伴侣，Henry)

图 19. 文档中长上下文关系推断的示例。

6.3 知识动态

许多研究人员为知识动态领域的文献作出了贡献。其中相当一部分研究采用RNN结构来理解历时依赖关系，从而预测状态变化。例如，Know-Evolve [340] 涉及一种多变量时间点过程，并结合增强型RNN结构来学习时间演化表示函数。RE-NET [341]引入邻域聚合器，以捕捉实体节点之间的并发交互。研究人员还设计了包含演化表示的模型，例如MGraph [342] 和DyERNIE[343]。Gracious 等人 [344] 系统地构建了一种神经潜在空间模型，将异构知识图谱的演化信息结合起来。Yan 等人 [345] 提升了GCN模型捕捉拓扑不变特征的能力。其思想是对齐不同时间知识图谱快照中的节点，并构建概念的动态概况。

当不同类型的条件发生变化时，知识如何演变仍然是一个挑战——以结束COVID-19疫情所需的条件为例。我们建议读者参考因果特征选择方法 [339] ，并结合专家意见和多源证据。

7 知识图谱存储

在本节　中，我们简要概述适用于不同数据环境的知识图谱存储工具　.

早期的图存储工作采用关系模型来持久化构建的知识图谱。传统关系数据库管理系统为表格形式的数据库提供可靠且快速的CRUD操作。开发人员还采用了深度优先遍历和最短路径搜索等图算法来增强关系数据库 [32]。此类算法的代表性示例包括PostgreSQL [346] 和filament [10]。然而，关系数据库处理稀疏知识图谱或执行用于分布式存储的数据分区时，成本可能非常高。

键值数据库是用于在大型知识图谱中保存簇的轻量级解决方案。此外，它们支持采用简化且灵活的数据格式进行分布式存储。Trinity [347]提供高性能的内存键值存储系统，用于管理包含数十亿个节点的大型知识图谱，例如 Probase。CouchDB [348] 使用复制机制来维护动态知识图谱。MapReduce 技术会自动将数据组转换为键值映射。Hadoop [11] 支持面向知识

[10]https://filament.sourceforge.net [11] http://hadoop.apache.org

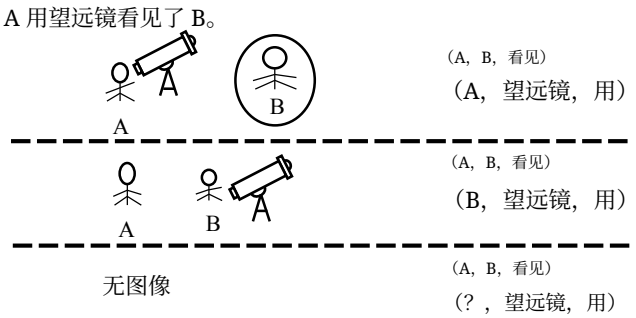


图20. 通过跨模态依赖解决信息不完整问题的示例。在此案例中，文本表述“某人用望远镜看见了某人。”含义不明确。提取器只能通过解读图像数据中的辅助信息，确定人物与“望远镜”之间的关系。

通过 MapReduce 实现图存储。Pregel [349] 开发了一种超步机制，用于在顶点之间共享消息，以实现并行计算。

另一个有前景的方向是设计适用于知识三元组结构的图数据库。Neo4j[350] 是一种基于 NoSQL 的轻量级图数据库，支持嵌入式动态知识图谱存储。SOnes [12] 为知识图谱数据库提供面向对象查询。人们还开发了用于知识存储的新型语言，例如资源描述框架（RDF）和网络本体语言（OWL） [13]。一些基于 RDF 的图数据库对图结构的存储进行了优化。例如，gStore [349] 通过子图匹配算法改进了基于 RDF 结构的知识图谱数据库。

8 关于知识图谱构建的讨论

研究人员针对知识图谱构建的不同方面提出了多种解决方案。然而，一些具有挑战性的问题和研究方向仍有待进一步探讨。

8.1 知识图谱构建中的长上下文与复杂语境

复杂的跨句或跨段上下文会妨碍知识图谱构建的各项子任务在实际应用中的表现，尤其是关系抽取任务。值得提醒读者的是，复杂上下文并非仅与长距离依赖有关。Yao 等人 [22] 指出，模式识别、共指推理、逻辑推理和常识推理这四类推理对于涵盖高阶上下文语义至关重要。图19展示了一个具体示例。

能够处理复杂长上下文的模型，应在针对多个语言对象进行推理时，重点关注复杂的跨句模式。除第4.3.6节中的文档级抽取模型外，第4.1.2 节中的一些工作也通过异构模型对文档级上下文进行建模，以判定实体类型。值得注意的是，用户生成的文本中可能出现歧义表达，而模型通常无法在缺乏外部信息的情况下正确解读这些表达。推理面临的另一个挑战是多跳推理。还应探索更多语言结构，以理解曲折复杂的表达。

需要背景知识才能处理的脱离上下文表达，是知识图谱构建的瓶颈。障碍主要有两方面：1）自发知识；2）证据支持。自发生成的常识知识常用于推导新事实，

[12]http://github.com/sones/sones [13]RDF 和 OWL 都是 W3C 标准，另请参阅 http://www.w3.org/RDF

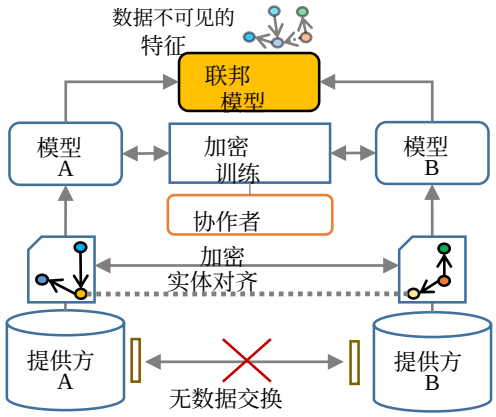


图21. 在保护隐私的同时，从不同知识提供方构建联邦模型的示意图。在此过程中，在利用多源数据部分分别训练模型之前，会先执行加密实体对齐；随后，协作者计算并聚合每个模型的加密梯度，以防止信息泄露。联邦模型仅保留数据不可见的众包知识特征。

例如，有孩子的男人和女人应该是夫妻／伴侣，尽管这类判断有时并不准确。如何获取常识规则并将其应用于合适的场景，是一个重要的研究方向。与此同时，许多文档级数据集不包含用于确定正确逻辑路径的证据信息。诸如 [357] 等工作已探究关系提及的文档级证据结构。然而，我们无法预见模型能否学会正确组织线索，从而在所有场景中解决事实问题（例如，验证哲学书中的结论）。我们认为，长上下文不仅仅是一个自然语言处理问题，模型 [358] 对语言表达的理解也将是一个关键方向。此外，为了严谨地理解上下文，还应考虑所提供数据源中的时间和地理信息等条件。

8.2 多模态知识图谱构建与补全

多模态知识图谱能够完整地表达和存储异构信息，以供展示。多模态知识还可用于检测虚假或低质量内容，例如图像不匹配的文本（如一份标注为巴黎、配图却是伦敦的文档）。多模态知识图谱[353] 是一种用于补全多模态元组的模型，可基于图像信息进行推理，而 Dost 等人 [354] 则探索了文本与图像的跨模态实体链接。多模态知识图谱面临的另一项问题，是通过跨模态依赖解决单模态表达中的语义不完整性。我们用图20中的一个具体示例来说明这一挑战。

8.3 联邦学习

联邦学习为满足隐私保护这一基本需求提供了一个富有启发性的方向。针对知识图谱的一种联邦设置，是从多个来源训练模型集成，这也是一种流行策略。联邦知识嵌入已取得显著进展，例如 FKGE [355] 和 FedE [356], , 它们在训练过程中融入跨模态特征，同时禁止数据交换。然而，实体对齐是一个矛盾的瓶颈，阻碍了联邦学习：模型学习之前必须共享多源知识图谱，这会在知识融合过程中交换敏感信息。如何在联邦特征的同时，为加密实体对齐构建隐私保护超特征空间，仍有待进一步探索。

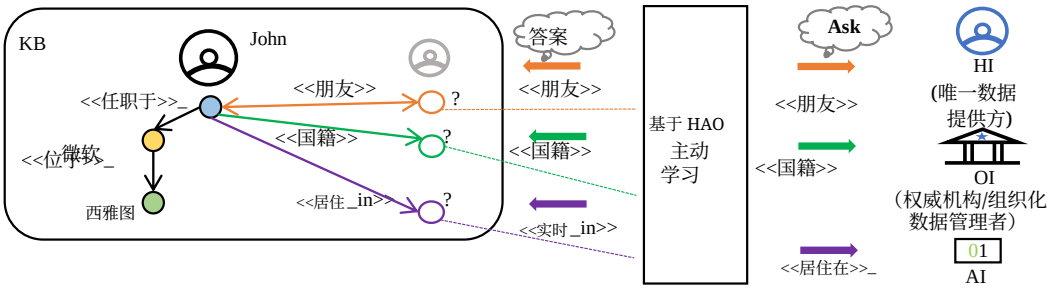


图 22. 一个基于 HAO 的主动学习知识图谱构建案例。基于 HAO 的主动学习模型会选择具有适当角色的用户，为不确定样本添加标签。在此案例中（如果隐私政策允许），John 的国籍将由权威机构（OI）进行标注，而他的朋友则通过询问其社交网络中的用户（HI）来确定。随后，人工智能将根据已知事实推导出他的居住地。

针对敏感数据场景，设计更具隐私保护性的知识图谱构建模型至关重要。我们在图 21 中展示了开发联邦模型的流程。

8.4 知识图谱构建任务中的高级语义与动态

近期研究已拓展至高级语义评估任务，例如检测歧义[351] 以及运用常识验证事实 [352]，以处理复杂的语言现象。解读文学表达（如明喻和隐喻）是智能知识图谱构建的未来方向。例如，“Tom went to heaven in 2008.”的意思是“Tom，死于，2008年”。开发或微调具备高级语义能力的预训练模型，将成为实现高生产力的起点。

此外，已有许多研究探讨时序知识图谱的动态特性。然而，知识语义如何随着一般相关条件的变化而演变，仍是一个尚未探索的领域。深入探讨启发式问题，例如“医务人员的职业社交网络如何随着大流行病不同阶段的变化而改变？”，或许有助于我们发现隐含因素，从而促进公共卫生领域的政策制定。捕捉相关条件如何影响相关事实的动态变化，是模拟一般人类知识的最终方向。

8.5 知识图谱构建中的人机协同

邀请合适的用户补充和修正知识图谱，是在开放世界中获取未知事实的终极解决方案。为此，Wu 等人 [8] 设计了 HAO 模型，通过人机协作解决不同的构建问题。基于 HAO 的主动学习模型能够自动识别不同角色（例如领域专家（HI）、组织权威（OI）、计算系统（AI）等），并将待确定的数据分配给合适的用户进行标注。将这一模型融入知识图谱构建框架，以赋予其智慧，将是一个有前景的方向。我们在图22中展示了这一重要构想。

8.6 跨语言知识图谱

构建跨语言知识图谱是一项长期目标，旨在整合分布于不同语言中的不均衡资源。Xlore [42] 提供了一个启发性的示例，展示了如何通过深度学习方法对齐跨语言实体。然而，机器翻译仍是跨语言任务面临的一大瓶颈。首先，翻译过程中产生的错误和冲突会影响知识精炼工作。其次，以少数语言表达的数据资源可能不足以支持机器学习。如何准确地执行自动化低资源知识翻译，同时解决跨语言冲突，是一个颇具前景的方向。

8.7 端到端统一构建框架

端到端抽取方法（如基于GCN的框架）将知识获取的各个子任务整合为一个统一的抽取任务，优于流水线式设计。然而，将知识获取与知识精炼任务结合起来，构建一个集成的联合模型，仍是一大瓶颈。探索能够统一知识图谱抽取与精炼的端到端框架，或许是一个值得深入研究的未来方向。提供高质量、开箱即用的解决方案，可以避免手动调整各个组件的需要；而能够考虑跨任务语义的方案，也值得付诸实践。此外，训练一个能够统一知识图谱构建通用流程的框架，将是一个值得解决的多任务学习难题。

9 结论

本文围绕知识图谱构建这一主题进行了全面综述。具体而言，我们回顾了在不同场景下，利用各种数据类型构建、完善和集成知识图谱所涉及的任务、方法、挑战及相关资源。为了探讨大数据环境中的核心主题，我们系统介绍了用于获取细粒度概念（实体类型判定）、处理低资源知识（少样本场景下的抽取任务）、理解大型语言对象（文档级关系抽取）、复杂推理（逻辑与可解释推理）以及处理知识图谱中条件结构（时间条件与一般条件）的代表性模型。此外，我们还简要介绍了实用的知识图谱工具包和项目。总之，知识图谱构建已成为人工智能应用实现人类智能的关键主题。未来，研究界必将继续探索更多范式，使知识图谱能够在海量、异构、自主、复杂且不断演化的数据环境中发挥智慧，同时加强知识社区之间的协作。

参考文献

[1] J. Liu, J. Ren, W. Zheng, L. Chi, I. Lee, 和 F. Xia, 《学者网络：学者知识图谱》, 载于 *SIGIR*, *2020*年, 第 2153–2156 页, 2020年。[2] X. Wang, X. He, Y. Cao, M. Liu, 和 T. Chua, 《KGAT：用于推荐的知识图谱注意力网络》, 载于 *KDD*, *2019*年, 第 950–958 页, 2019年。[3] J. Bao, N. Duan, Z. Yan, M. Zhou, 和 T. Zhao, 《基于知识图谱的约束问答》, 载于 *COLING* 论文集, *2016*年, 第 2503–2514 页, 2016年。[4] K. D. Bollacker, R. P. Cook, 和 P. Tufts, 《Freebase：一个共享的结构化通用人类知识数据库》, 载于 *AAAI-07* 论文集, *2007*年, 第 1962–1963 页, 2007年。[5] D. Vrandecic, 《Wikidata：协作式数据收集的新平台》, 载于 *WWW* 论文集, *2012*年, *2012*年（伴随卷）, 第 1063–1064 页, 2012年。[6] A. Yates, M. Banko, M. Broadhead, M. J. Cafarella, O. Etzioni, 和 S. Soderland, 《TextRunner：网页上的开放信息抽取》, 载于 *HLT-NAACL* 论文集, *2007*年, 第 25–26 页, 2007年。[7] O. Etzioni, M. J. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, 和 A. Yates, 《KnowItAll 中的网页规模信息抽取：（初步结果）》, 载于 *WWW* 论文集, *2004*年, 第 100–110 页, 2004年。[8] M. Wu 和 X. Wu, 《论大智慧》, 知识与信息系统, 第 58 卷, 第 1 期, 第 1–8 页, 2019年。[9] Z. Huang, W. Xu, 和 K. Yu, 《用于序列标注的双向LSTM-CRF模型》, *CoRR*, 第 abs/1508.01991 卷, 2015年。[10] X. Ma 和 E. H. Hovy, 《通过双向LSTM-CNN-CRF实现端到端序列标注》, 载于 *ACL* 论文集, *2016*年, 第 *1* 卷：长篇论文, 2016年。[11] P. Xu 和 D. Barbosa, 《具有层级感知损失的神经细粒度实体类型分类》, 载于 *NAACL-HLT* 论文集, *2018*年, 第 *1* 卷（长篇论文）, 第 16–25 页, 2018年。[12] X. Ren, W. He, M. Qu, C. R. Voss, H. Ji, 和 J. Han, 《通过异构部分标签嵌入降低实体类型判定中的标签噪声》, 载于 *KDD*, *2016*年, 第 1825–1834 页, 2016年。[13] O. Ganea 和 T. Hofmann, 《利用局部神经注意力进行深度联合实体消歧》, 载于 *EMNLP* 论文集, *2017*年, 第 2619–2629 页, 2017年。

[14] P. Le 和 I. Titov，《通过对提及之间的潜在关系建模来改进实体链接》，载于 ACL 论文集，2018年，第 1 卷：长篇论文，第 1595–1604 页，2018年。[15] K. Lee, L. He, M. Lewis, 和 L. Zettlemoyer，《端到端神经共指消解》，载于 EMNLP，2017年，第 188–197 页，2017年。[16] D. Zeng, K. Liu, Y. Chen, 和 J. Zhao，《通过分段卷积神经网络进行关系抽取的远程监督》，载于 EMNLP，2015年，第 1753–1762 页，2015年。[17] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, 和 B. Xu，《用于关系分类的基于注意力的双向长短期记忆网络》，载于 ACL 论文集，2016年，第 2 卷：短篇论文，2016年。[18] H. Zhang, D. Khashabi, Y. Song, 和 D. Roth，《TransOMCS：从语言图到常识知识》，载于 IJCAI 论文集，2020年，第 4004–4010 页，2020年。[19] H. Zhang, X. Liu, H. Pan, H. Ke, J. Ou, T. Fang, 和 Y. Song，《ASER：通过事件上的高阶选择偏好实现大规模常识知识获取》，CoRR，第 abs/2104.02137 卷，2021年。[20] W. Xin-Dong, S. Shao-Jing, J. Ting-Ting, B. Chen-Yang, 和 W. Ming-Hui，《Huapu-cp：从知识图谱到数据中央平台》，《自动化学报》，第 46 卷，第 10 期，第 2045–2059 页，2020年。[21] J. Devlin, M. Chang, K. Lee, 和 K. Toutanova，《BERT：用于语言理解的深度双向变换器预训练》，载于 NAACL-HLT，2019年，第 1 卷（长篇与短篇论文），第 4171–4186 页，2019年。[22] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, 和 M. Sun，《DocRED：大规模文档级关系抽取数据集》，载于 ACL 论文集，2019年，第 1 卷：长篇论文，第 764–777 页，2019年。[23] S. Riedel, L. Yao, 和 A. McCallum，《在没有标注文本的情况下对关系及其提及进行建模》，载于 ECML PKDD，2010年，论文集，第 III 部分，第 6323 卷，第 148–163 页，2010年。[24] X. Wang, X. Han, Y. Lin, Z. Liu, 和 M. Sun，《对抗式多语言神经关系抽取》，载于 COLING 论文集，2018年，第 1156–1166 页，2018年。[25] Z. Han, P. Chen, Y. Ma, 和 V. Tresp，《用于时序知识图谱预测的可解释子图推理》，载于 ICLR，2021年。[26] T. Jiang, T. Zhao, B. Qin, T. Liu, N. V. Chawla, 和 M. Jiang，《从科学文本中联合抽取事实和条件元组的多输入多输出序列标注》，载于 EMNLP-IJCNLP 论文集，2019年，第 302–312 页，2019年。[27] A. Pradhan, K. K. Todi, A. Selvarasu, 和 A. Sanyal，《基于深度主动学习的知识图谱生成》，载于 IJCNN，2020年，第 1–8 页，2020年。[28] X. Wu, X. Zhu, G. Wu, 和 W. Ding，《大数据数据挖掘》，IEEE知识与数据工程汇刊，第 26 卷，第 1 期，第 97–107 页，2014年。[29] A. Hogan, E. Blomqvist, M. Cochez, C. d'A mato, G. de Melo, C. Gutiérrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, 和 A. Zimmermann，《知识图谱》。《数据、语义与知识综合讲座》，Morgan & Claypool 出版社，2021年。[30] H. Paulheim，《知识图谱精炼：方法与评估技术综述》，《语义网》，第 8 卷，第 3 期，第 489–508 页，2017年。[31] X. Wu, J. Wu, X. Fu, J. Li, P. Zhou, 和 X. Jiang，《自动知识图谱构建：2019年 ICDM/ICBK竞赛报告》，载于 ICDM，2019年，第 1540–1545 页，2019年。[32] J. Yan, C. Wang, W. Cheng, M. Gao, 和 A. Zhou，《知识图谱回顾》，《计算机科学前沿》，第 12 卷，第 1 期，第 55–74 页，2018年。[33] T. Nayak, N. Majumder, P. Goyal, 和 S. Poria，《关系三元组抽取的深度神经方法：全面综述》，《认知计算》，第 13 卷，第 5 期，第 1215–1232 页，2021年。[34] S. Pawar, P. Bhattacharyya, 和 G. K. Palshikar，《联合抽取实体和关系的技术：综述》，CoRR，第 abs/2103.06118 卷，2021年。[35] S. Ji, S. Pan, E. Cambria, P. Marttinen, 和 P. S. Yu，《知识图谱综述：表示、获取与应用》，《IEEE神经网络与学习系统汇刊》，第 33 卷，第 2 期，第 494–514 页，2022年。[36] S. Arora，《用于知识图谱补全的图神经网络综述》，CoRR，第 abs/2007.12374 卷，2020年。[37] B. Cai, Y. Xiang, L. Gao, H. Zhang, Y. Li, 和 J. Li，《时序知识图谱补全：综述》，CoRR，第 abs/2201.08236 卷，2022年。[38] Q. Wang, Z. Mao, B. Wang, 和 L. Guo，《知识图谱嵌入：方法与应用综述》，IEEE知识与数据工程汇刊，第 29 卷，第 12 期，第 2724–2743 页，2017年。[39] L. Ehrlinger 和 W. Wöß，《迈向知识图谱的定义》，载于 SEMANTiCS，SuCCESS'16，2016年，第 1695 卷 CEUR 研讨会论文集，2016年。[40] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, 和 Z. G. Ives，《DBpedia：开放数据网络的核心》，载于 ISWC + ASWC，2007年，第 4825 卷，第 722–735 页，2007年。[41] W. Wu, H. Li, H. Wang, 和 K. Q. Zhu，《Probase：用于文本理解的概率分类体系》，载于 SIGMOD，2012年，第 481–492 页，2012年。

[42] Z. Wang、J. Li、Z. Wang、S. Li、M. Li、D. Zhang、Y. Shi、Y. Liu、P. Zhang 和 J. Tang，“Xlore：一个大规模英汉双语知识图谱”，载于 ISWC 论文集，2013年，CEUR 研讨会论文集第1035卷（海报与演示专题），第121–124页，2013年。[43] B. Xu、Y. Xu、J. Liang、C. Xie、B. Liang、W. Cui 和 Y. Xiao，“CN-DBpedia：一个永不停歇的中文知识抽取系统”，载于 IEA/AIE，2017年，论文集，第二部分，第10351卷，第428–438页，2017年。[44] F. M. Suchanek、G. Kasneci 和 G. Weikum，“YAGO：语义知识的核心”，载于 WWW 论文集，2007年，706，2007年697页。[45] M. Mintz、S. Bills、R. Snow 和 D. Jurafsky，“无需标注数据的关系抽取远程监督”，载于 ACL 论文集，2009年，第1003–1011页，2009年。[46] J. Chen、A. Wang、J. Chen、Y. Xiao、Z. Chu、J. Liu、J. Liang 和 W. Wang，“CN-Probase：一种大规模中文分类体系构建的数据驱动方法”，载于 ICDE 论文集，2019年，第1706–1709页，2019年。[47] G. A. Miller，“WordNet：英语词汇数据库”，《ACM通讯》，第38卷，第11期，第39–41页，1995。[48] R. Navigli 和 S. P. Ponzetto，“BabelNet：构建超大型多语言语义网络”，载于 ACL 论文集，2010年，第216–225页，2010年。[49] H. Liu 和 P. Singh，“ConceptNet——一种实用的常识推理工具包”，《BT技术期刊》，第22卷，第4期，第211–226页，2004年。[50] Z. Dong 和 Q. Dong，“HowNet——一种混合型语言与知识资源”，载于 ICNLP，2003年，824，2003年820页，论文集。[51] S. Han、Y. Zhang、Y. Ma、C. Tu、Z. Guo、Z. Liu 和 M. Sun，“THUOCL：清华开放中文词库”，清华大学，2016年。[52] C. Matuszek、J. Cabral、M. J. Witbrock 和 J. DeOliveira，“Cyc 的语法与内容简介”，载于 2006年 AAAI 春季研讨会论文集，技术报告 SS-06-05，2006年，第44–49页，2006年。[53] T. Steiner、R. Verborgh、R. Troncy、J. Gabarró 和 R. V. de Walle，“为谷歌知识图谱添加实时覆盖”，载于 ISWC 论文集，2012年，CEUR 研讨会论文集第914卷，2012年。[54] R. J. Roberts，“PubMed Central：已发表文献的 GenBank”，《美国国家科学院院刊》，第98卷，第2期，第381页，2001年。[55] D. S. Wishart、C. Knox、A. Guo、S. Shrivastava、M. Hassanali、P. Stothard、Z. Chang 和 J. Woolsey，“DrugBank：用于计算机药物发现与探索的综合资源”，《核酸研究》，第34卷，数据库专刊，第668–672页，2006年。[56] J. Tang、D. Zhang 和 L. Yao，“学术研究人员社交网络抽取”，载于 ICDM 论文集，2007年，301，2007年292页。[57] F. Zhang、X. Liu、J. Tang、Y. Dong、P. Yao、J. Zhang、X. Gu、Y. Wang、B. Shao、R. Li 和 K. Wang，“OAG：迈向大规模异构实体图的链接”，载于 KDD，2019年，第2585–2595页，2019年。[58] H. Chen、N. Hu、G. Qi、H. Wang、Z. Bi、J. Li 和 F. Yang，“OpenKG Chain：开放知识图谱的区块链基础设施”，《数据智能》，第3卷，第2期，第205–227页，2021年。[59] W. Gong-Qing、H. Jun、L. Li、X. Zhe-Hao、L. Peng-Cheng、H. Xue-Gang 和 W. Xin-Dong，“基于标签路径特征融合的在线网页新闻抽取。软件学报”，《软件学报》，第27卷，第3期，第714–735页，2016年。[60] A. Fader、S. Soderland 和 O. Etzioni，“识别开放信息抽取中的关系”，载于 EMNLP 论文集，2011年。[61] Mausam、M. Schmitz、S. Soderland、R. Bart 和 O. Etzioni，“面向信息抽取的开放式语言学习”，载于 EMNLP-CoNLL 论文集，ACL 2012，第523–534页，2012年。[62] S. Bird，“NLTK：自然语言工具包”，载于 ACL，2006年（N. Calzolari、C. Cardie 和 P. Isabelle 编），2006年。[63] C. D. Manning、M. Surdeanu、J. Bauer、J. R. Finkel、S. Bethard 和 D. McClosky，“Stanford CoreNLP 自然语言处理工具包”，载于 ACL 论文集，2014年，第55–60页，2014年。[64] Z. Zhang，“使用 TableMiner+ 进行有效高效的语义表格解释”，《语义网》，第8卷，第6期，第921–957页，2017年。[65] M. Cremaschi、A. Rula、A. Siano 和 F. D. Paoli，“MantisTable：用于创建表格数据语义标注的工具”，载于 ESWC 2019 卫星活动，2019年，修订精选论文，第11762卷，第18–23页，2019年。[66] M. Honnibal 和 I. Montani，“spaCy 2：借助 Bloom 嵌入、卷积神经网络和增量解析实现自然语言理解。”即将发表，2017年。[67] X. Han、T. Gao、Y. Yao、D. Ye、Z. Liu 和 M. Sun，“OpenNRE：用于神经关系抽取的开放式可扩展工具包”，载于 EMNLP-IJCNLP 论文集：系统演示，第169–174页，2019年。[68] X. Han、S. Cao、X. Lv、Y. Lin、Z. Liu、M. Sun 和 J. Li，“OpenKE：知识嵌入开放工具包”，载于 EMNLP 论文集，2018年，第139–144页，2018年。[69] Z. Sun、Q. Zhang、W. Hu、C. Wang、M. Chen、F. Akrami 和 C. Li，“基于嵌入的知识图谱实体对齐基准研究”，《VLDB 论文集》，第13卷，第11期，第2326–2340页，2020年。[70] A. Grover 和 J. Leskovec，“node2 864，2016 855 vec：面向网络的可扩展特征学习”，载于 KDD 论文集，第–页。

[71] J. Tang、M. Qu、M. Wang、M. Zhang、J. Yan 和 Q. Mei，“LINE：大规模信息网络嵌入”，载于 WWW 论文集，第1067–1077页，2015年。[72] W. Hu 和 Y. Qu，“Falcon-AO：一种实用的本体匹配系统”，《网络语义学杂志》，第6卷，第3期，第237–239页，2008年。[73] C. Chang、C. Hsu 和 S. Lui，“通过模式发现从半结构化网页中自动抽取信息”，《决策支持系统》，第35卷，第1期，第129–147页，2003年。[74] C. Hsu 和 M. Dung，“为从网页中抽取半结构化数据生成有限状态转换器”，《信息系统》，第23卷，第8期，第521–538页，1998年。[75] P. B. Golgher、A. S. da Silva、A. H. F. Laender 和 B. A. Ribeiro-Neto，“基于示例的数据抽取自举”，载于 CIKM 论文集，2001年，第371–378页，2001年。[76] A. Carlson 和 C. Schafer，“从半结构化网页中自举信息抽取”，载于 ECML/PKDD，2008年，论文集，第一部分，第5211卷，第195–210页，2008年。[77] B. Adelberg，“NoDoSE——一种从文本文件中半自动抽取半结构化数据的工具”，载于 SIGMOD，1998年（L. M. Haas 和 A. Tiwary 编），第283–294页，1998年。[78] A. H. F. Laender、B. A. Ribeiro-Neto 和 A. S. da Silva，“DEByE——基于示例的数据抽取”，《数据与知识工程》，第40卷，第2期，第121–154页，2002年。[79] A. Finn、N. Kushmerick 和 B. Smyth，“事实还是虚构：数字图书馆的内容分类”，载于 DELOS，2001年，ERCIM 研讨会论文集第01/W03卷，2001年。[80] T. Weninger、W. H. Hsu 和 J. Han，“CETR：基于标签比率的内容抽取”，载于 WWW 论文集，2010年，第971–980页，2010年。[81] F. Sun、D. Song 和 L. Liao，“基于 DOM 的内容抽取：利用文本密度”，载于 SIGIR，2011年，第245–254页，2011年。[82] G. Wu、L. Li、X. Hu 和 X. Wu，“基于路径比率的网络新闻抽取”，载于 CIKM，2013年，第2059–2068页，2013年。[83] D. Cai、S. Yu、J. Wen 和 W. Ma，“基于视觉表示抽取网页内容结构”，载于 APWeb，2003年，论文集，第2642卷，第406–417页，2003年。[84] Y. Wang 和 J. Hu，“一种基于机器学习的网页表格检测方法”，载于 WWW，2002年，250，2002年242页。[85] M. J. Cafarella、A. Y. Halevy、Y. Zhang、D. Z. Wang 和 E. Wu，“揭示关系型网络”，载于第11届万维网与数据库国际研讨会（WebDB），2008年。[86] J. Eberius、K. Braunschweig、M. Hentsch、M. Thiele、A. Ahmadov 和 W. Lehner，“构建德累斯顿网页表格语料库：一种分类方法”，载于 BDC，2015年，第41–50页，2015年。[87] M. J. Cafarella、A. Y. Halevy 和 N. Khoussainova，“关系型网络的数据集成”，《VLDB 论文集》，第2卷，第1期，第1090–1101页，2009年。[88] S. Zhang 和 K. Balog，“网页表格抽取、检索与增强：综述”，《ACM智能系统与技术汇刊》，第11卷，第2期，第13:1–13:35页，2020年。[89] N. Kushmerick，“包装器归纳：效率与表达能力”，《人工智能》，第118卷，68，2000年1 2，第15期，–页。[90] D. Buttler、L. Liu 和 C. Pu，“用于万维网的全自动对象抽取系统”，载于 ICDCS 论文集，2001年，第361–370页，2001年。[91] B. M. Sundheim，“消息理解会议”，载于 TIPSTER文本计划第二阶段：1996年在维也纳举行的研讨会论文集，第35–37页，1996年。[92] S. Thenmalar、B. Jagan 和 T. V. Geetha，“命名实体识别的半监督自举方法”，CoRR，第abs/1511.06833卷，2015年。[93] G. Zhou 和 J. Su，“使用基于隐马尔可夫模型的分块标注器进行命名实体识别”，载于 ACL 论文集，2002年，480，2002年473页。[94] J. R. Finkel、T. Grenager 和 C. D. Manning，“通过吉布斯采样将非局部信息纳入信息抽取系统”，载于 ACL 论文集，2005年，第363–370页，2005年。[95] J. Zhu、Z. Nie、J. Wen、B. Zhang 和 W. Ma，“用于网页信息抽取的二维条件随机场”，载于国际机器学习大会论文集，2005年，第119卷，第1044–1051页，2005年。[96] C. Sutton、K. Rohanimanesh 和 A. McCallum，“动态条件随机场：用于序列数据标注与分割的因子化概率模型”，载于国际机器学习大会论文集，2004年，第69卷，2004年。[97] J. Zhu、Z. Nie、J. Wen、B. Zhang 和 W. Ma，“网页数据抽取中的记录检测与属性标注同步进行”，载于 KDD，2006年，第494–503页，2006年。[98] A. Finn 和 N. Kushmerick，“用于信息抽取的多级边界分类”，载于 ECML 论文集，2004年，第3201卷，第111–122页，2004年。[99] J. Li、A. Sun、J. Han 和 C. Li，“深度学习用于命名实体识别的综述”，《IEEE知识与数据工程汇刊》，第34卷，第1期，第50–70页，2022年。[100] R. Collobert、J. Weston、L. Bottou、M. Karlen、K. Kavukcuoglu 和 P. P. Kuksa，“（几乎）从零开始的自然语言处理”，《机器学习研究杂志》，第12卷，第2493–2537页，2011年。

[101] E. Strubell, P. Verga, D. Belanger 和 A. McCallum，“利用迭代扩张卷积实现快速、准确的实体识别”，载于 EMNLP 论文集，2017年，第2670–2680页，2017年。[102] L. Li, L. Jin, Z. Jiang, D. Song 和 D. Huang，“基于扩展循环神经网络的生物医学命名实体识别”，载于 BIBM，2015年，第649–652页，2015年。[103] T. H. Nguyen, A. Sil, G. Dinu 和 R. Florian，“利用循环神经网络提升提及检测的鲁棒性”，CoRR，第 abs/1602.07749 卷，2016年。[104] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin 和 J. Wang，“一种基于注意力的 BiLSTM-CRF 文档级化学命名实体识别方法”，Bioinform.，第34卷，第8期，第1381–1388页，2018年。[105] A. Z. Gregoric, Y. Bachrach, P. Minkovsky, S. Coope 和 B. Maksak，“使用自注意力机制进行神经命名实体识别”，载于 ICTAI，2017年，第652–656页，2017年。[106] A. Cetoli, S. Bragaglia, A. D. O’ Harney 和 M. Sloan，“用于命名实体识别的图卷积网络”，载于 TLT 论文集，2018年，第37–45页，2018年。[107] C. Dogan, A. Dutra, A. Gara, A. Gemma, L. Shi, M. Sigamani 和 E. Walters，“使用 ELMO 和 Wikidata 进行细粒度命名实体识别”，CoRR，第 abs/1904.10503 卷，2019年。[108] M. Liu, Z. Tu, T. Zhang, T. Su, X. Xu 和 Z. Wang，“Ltp：一种用于基于 CRF 的命名实体识别的新型主动学习策略”，Neural Processing Letters，第1–22页，2022年。[109] I. Yamada, A. Asai, H. Shindo, H. Takeda 和 Y. Matsumoto，“LUKE：通过实体感知自注意力实现深度上下文化的实体表示”，载于 EMNLP，2020年，第6442–6454页，2020年。[110] S. Shimaoka, P. Stenetorp, K. Inui 和 S. Riedel，“用于细粒度实体类型分类的神经网络架构”，载于 EACL 论文集，2017年，第1卷：长篇论文，第1271–1280页，2017年。[111] S. Zhang, K. Duh 和 B. V. Durme，“通过扩展话语上下文和自适应分类阈值实现细粒度实体类型判定”，载于 *SEM@NAACL-HLT 论文集，2018年，第173–179页，2018年。[112] C. Moon, P. Jones 和 N. F. Samatova，“学习用于知识图谱补全的实体类型嵌入”，载于 CIKM 论文集，2017年，第2215–2218页，2017年。[113] J. Hao, M. Chen, W. Yu, Y. Sun 和 W. Wang，“通过联合嵌入实例和本体概念实现知识库的通用表示学习”，载于 KDD，2019年，第1709–1719页，2019年。[114] Y. Zhao, A. Zhang, R. Xie, K. Liu 和 X. Wang，“连接用于知识图谱实体类型判定的嵌入”，载于 ACL 论文集，2020年，第6419–6428页，2020年。[115] H. Jin, L. Hou 和 J. Li，“利用类型层级增强异构网络嵌入，以实现知识库中的细粒度实体类型判定”，载于 NLP-NABD，2018年，论文集，第11221卷，第170–182页，2018年。[116] H. Jin, L. Hou, J. Li 和 T. Dong，“用于知识库中细粒度实体类型判定的属性化与预测性实体嵌入”，载于 COLING 论文集，2018年，第282–292页，2018年。[117] Y. Cao, L. Hou, J. Li, Z. Liu, C. Li, X. Chen 和 T. Dong，“通过注意力式远程监督联合表示学习跨语言词语和实体”，载于 EMNLP 论文集，2018年，第227–237页，2018年。[118] G. Limaye, S. Sarawagi 和 S. Chakrabarti，“利用实体、类型和关系标注及搜索网页表格”，VLDB Endow. 论文集，第3卷，第1期，第1338–1347页，2010年。[119] C. S. Bhagavatula, T. Noraset 和 D. Downey，“TabEL：网页表格中的实体链接”，载于 ISWC，2015年，论文集，第一部分，第9366卷，第425–441页，2015年。[120] T. Wu, S. Yan, Z. Piao, L. Xu, R. Wang 和 G. Qi，“利用多个关联知识库在网页表格中进行实体链接”，载于 JIST，2016年，修订精选论文，第10055卷，第239–253页，2016年。[121] V. Efthymiou, O. Hassanzadeh, M. Rodriguez-Muro 和 V. Christophides，“将网页表格与知识库实体匹配：从实体查找到实体嵌入”，载于 ISWC 论文集，2017年，第一部分，第10587卷，第260–277页，2017年。[122] V. Mulwad, T. Finin, Z. Syed 和 A. Joshi，“使用关联数据解释表格”，载于第一届关联数据消费国际研讨会论文集，2010年，CEUR 研讨会论文集第665卷，2010年。[123] Y. Guo, W. Che, T. Liu 和 S. Li，“一种基于图的实体链接方法”，载于 IJCNLP，2011年，第1010–1018页，2011年。[124] X. Han, L. Sun 和 J. Zhao，“网页文本中的集体实体链接：一种基于图的方法”，载于 SIGIR，2011年，774，2011年 765页 – 。[125] W. Shen, J. Wang, P. Luo 和 M. Wang，“LIEGE：利用知识库链接网页列表中的实体”，载于 KDD，2012年，第1424–1432页，2012年。[126] A. Bagga 和 B. Baldwin，“使用向量空间模型进行基于实体的跨文档共指”，载于 COLING-ACL，1998年，会议论文集，第79–85页，1998年。[127] I. Lasek 和 P. Vojtás，“用于命名实体消歧的多种文本表示方法”，载于 IIWAS，2012年，第256–262页，2012年。[128] H. Huang, L. P. Heck 和 H. Ji，“利用深度神经网络和知识图谱进行实体消歧”，CoRR，第 abs/1504.07678 卷，2015年。[129] W. Fang, J. Zhang, D. Wang, Z. Chen 和 M. Li，“通过联合嵌入知识与文本进行实体消歧”，载于 SIGNLL、CoNLL、ACL，2016年，第260–269页，2016年。

[130] S. J. Pan, Z. Toh 和 J. Su，“用于跨领域命名实体识别的迁移联合嵌入”，ACM信息系统汇刊，第31卷，第2期，第7页，2013年。[131] B. Y. Lin 和 W. Lu，“用于跨领域命名实体识别的神经适配层”，载于 EMNLP 论文集，2018年，第2012–2022页，2018年。[132] K. Narasimhan, A. Yala 和 R. Barzilay，“通过强化学习获取外部证据以改进信息抽取”，载于 EMNLP 论文集，2016年，第2355–2365页，2016年。[133] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg 和 D. Hassabis，“通过深度强化学习实现人类水平的控制”，Nat.，第518卷，第7540期，第529–533页，2015年。[134] Y. Yang, W. Chen, Z. Li, Z. He 和 M. Zhang，“结合部分标注学习和强化学习的远程监督命名实体识别”，载于 COLING 论文集，2018年，第2159–2169页，2018年。[135] J. T. Zhou, H. Zhang, D. Jin, H. Zhu, M. Fang, R. S. M. Goh 和 K. Kwok，“用于低资源命名实体识别的双重对抗神经迁移”，载于 ACL 论文集，2019年，第1卷：长篇论文，第3461–3471页，2019年。[136] P. Cao, Y. Chen, K. Liu, J. Zhao 和 S. Liu，“利用自注意力机制对中文命名实体识别进行对抗迁移学习”，载于 EMNLP 论文集，2018年，第182–192页，2018年。[137] J. Li, D. Ye 和 S. Shang，“利用指针网络对命名实体边界检测进行对抗迁移”，载于 IJCAI 论文集，2019年，第5053–5059页，2019年。[138] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod 和 A. Anandkumar，“用于命名实体识别的深度主动学习”，载于 ICLR，2018年，会议论文集，2018年。[139] Y. Ma, E. Cambria 和 S. Gao，“用于零样本细粒度命名实体类型识别的标签嵌入”，载于 COLING，2016年，会议论文集：技术论文，第171–180页，2016年。[140] T. Zhang, C. Xia, C. Lu 和 P. S. Yu，“MZET：记忆增强的零样本细粒度命名实体类型识别”，载于 COLING 论文集，2020年，第77–87页，2020年。[141] A. Sil 和 A. Yates，“用于联合命名实体识别与链接的重排序”，载于 CIKM，2013年，第2369–2374页，2013年。[142] X. Liu, M. Zhou, X. Zhou, Z. Fu 和 F. Wei，“对推文中的命名实体识别与规范化进行联合推断”，载于 ACL 论文集，2012年，第1卷：长篇论文，第526–535页，2012年。[143] M. C. Phan, A. Sun, Y. Tay, J. Han 和 C. Li，“用于集体实体消歧的成对链接：两个可能胜过全部”，IEEE知识与数据工程汇刊，第31卷，第7期，第1383–1396页，2019年。[144] J. Cai 和 M. Strube，“通过超图划分实现端到端共指消解”，载于 COLING 论文集，2010年，第143–151页，2010年。[145] E. Sapena, L. Padró 和 J. Turmo，“一种基于约束的超图划分共指消解方法”，Comput. Linguistics，第39卷，第4期，第847–884页，2013年。[146] D. L. Bean 和 E. Riloff，“无监督学习用于共指消解的上下文角色知识”，载于 HLT-NAACL，2004年，第297–304页，2004年。[147] E. R. Fernandes, C. N. dos Santos 和 R. L. Milidiú，“采用特征归纳的潜在结构感知机，用于无约束共指消解”，载于 EMNLP-CoNLL、ACL，2012年，第41–48页，2012年。[148] W. M. Soon, H. T. Ng 和 C. Y. Lim，“一种针对名词短语共指消解的机器学习方法”，Comput. Linguistics，第27卷，第4期，第521–544页，2001年。[149] M. Recasens, M. de Marneffe 和 C. Potts，“话语实体的生与死：识别单例提及”，载于 NAACL-HLT，论文集，2013年，第627–633页，2013年。[150] M. A. ur Rahman 和 V. Ng，“用于共指消解的监督模型”，载于 EMNLP 论文集，2009年，ACL 特别兴趣小组 SIGDAT 会议，第968–977页，2009年。[151] V. Stoyanov 和 J. Eisner，“易优先共指消解”，载于 COLING 论文集，2012年，第2519–2534页，2012年。[152] X. Xi, G. Zhou, F. Hu 和 B. Fu，“通过建模层次特征实现共指消解的卷积深度神经网络”，载于 IScIDE，2015年，修订精选论文，论文集，第二部分，第9243卷，第361–372页，2015年。[153] J. Wu 和 W. Ma，“一种基于卷积神经网络的共指消解深度学习框架”，载于 ICSC，2017年，第61–64页，2017年。[154] S. Wiseman, A. M. Rush 和 S. M. Shieber，“学习用于共指消解的全局特征”，载于 NAACL-HLT，2016年，第994–1004页，2016年。[155] J. Gu, Z. Ling 和 N. Indurkhya，“关于改进端到端神经共指消解的研究”，载于 NLP-NABD，2018年，论文集，第11221卷，第159–169页，2018年。[156] R. Zhang, C. N. dos Santos, M. Yasunaga, B. Xiang 和 D. R. Radev，“通过联合提及检测和提及聚类，采用深度双仿射注意力进行神经共指消解”，载于 ACL 论文集，2018年，107，2018年 2 102 卷：短篇论文，第 – 页。[157] J. Ma, J. Liu, Y. Li, X. Hu, Y. Pan, S. Sun 和 Q. Lin，“通过互注意力联合优化神经共指消解”，载于 WSDM，2020年，第402–410页，2020年。

[158] K. Clark 和 C. D. Manning，《用于提及排序共指模型的深度强化学习》，载于 EMNLP 论文集，2016年，第 2256–2262 页，2016年。[159] G. Durrett 和 D. Klein，《共指消解中的轻松胜利与艰难挑战》，载于 EMNLP 论文集，2013年，ACL 特别兴趣小组 SIGDAT 会议，第 1971–1982 页，2013年。[160] S. Martschat 和 M. Strube，《用于共指消解的潜在结构》，《计算语言学协会汇刊》，第 3 卷，第 405–418 页，2015 年。[161] K. Chakrabarti、S. Chaudhuri、T. Cheng 和 D. Xin，《稳健发现实体同义词的框架》，载于 KDD，2012年，第 1384–1392 页，2012年。[162] R. C. Bunescu 和 R. J. Mooney，《利用最小监督学习从网页中抽取关系》，载于 ACL 论文集，2007年。[163] F. Reichartz、H. Korte 和 G. Paass，《基于带类型依存树上核函数的语义关系抽取》，载于 KDD，2010年，第 773–782 页，2010年。[164] D. Zelenko、C. Aone 和 A. Richardella，《关系抽取的核方法》，载于 EMNLP 论文集，2002年，第 78 页，2002年，第 71 页。[165] D. Zeng、K. Liu、S. Lai、G. Zhou 和 J. Zhao，《通过卷积深度神经网络进行关系分类》，载于 COLING 论文集，2014年，第 2335–2344 页，2014年。[166] T. H. Nguyen 和 R. Grishman，《关系抽取：来自卷积神经网络的视角》，载于 VS@NAACL-HLT 论文集，2015年，第 39–48 页，2015年。[167] M. Miwa 和 M. Bansal，《利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 ACL 论文集，2016年，第 1 卷：长篇论文，2016年。[168] Y. Shen 和 X. Huang，《用于语义关系抽取的基于注意力的卷积神经网络》，载于 COLING 论文集，2016年，第 2526–2536 页，2016年。[169] L. Wang、Z. Cao、G. de Melo 和 Z. Liu，《通过多层次注意力卷积神经网络进行关系分类》，载于 ACL 2016 论文集，第 1 卷：长篇论文，2016年。[170] Y. Zhao、H. Wan、J. Gao 和 Y. Lin，《通过实体对图改进关系分类》，载于 ACML，2019年，第 101 卷，第 1156–1171 页，2019年。[171] Z. Guo、Y. Zhang 和 W. Lu，《用于关系抽取的注意力引导图卷积网络》，载于 ACL 论文集，2019年，第 1 卷：长篇论文，第 241–251 页，2019年。[172] K. Zhao、H. Xu、Y. Cheng、X. Li 和 K. Gao，《基于异构图神经网络的实体与关系联合抽取表示迭代融合》，《知识库系统》，第 219 卷，第 106888 页，2021年。[173] A. D. N. Cohen、S. Rosenman 和 Y. Goldberg，《将关系抽取视为双向跨度预测》，CoRR，卷 abs/2010.04829，2020年。[174] N. FitzGerald、O. Täckström、K. Ganchev 和 D. Das，《利用神经网络因子进行语义角色标注》，载于 EMNLP 论文集，2015年，第 960–970 页，2015年。[175] M. Roth 和 M. Lapata，《基于依存路径嵌入的神经语义角色标注》，载于 ACL 2016 论文集，第 1 卷：长篇论文，2016年。[176] V. Mulwad、T. Finin 和 A. Joshi，《通过语义消息传递从表格生成关联数据》，载于 ISWC，2013年，论文集，第 I 部分，第 8218 卷，第 363–378 页，2013 年。[177] Z. Chen 和 M. J. Cafarella，《自动抽取网页电子表格数据》，载于第三届国际网络语义搜索研讨会（SSW ’13），2013年，第 1:1–1:8 页，2013年。[178] J. Zhu、Z. Nie、X. Liu、B. Zhang 和 J. Wen，《StatSnowball：一种抽取实体关系的统计方法》，载于 WWW 论文集，2009年，第 101–110 页，2009年。[179] S. Brin，《从万维网中抽取模式与关系》，载于《万维网与数据库》，国际研讨会 WebDB’98，1998年，精选论文，第 1590 卷，第 172–183 页，1998年。[180] E. Agichtein 和 L. Gravano，《Snowball：从大型纯文本集合中抽取关系》，载于 ACM 论文集，2000年，第 85–94 页，2000年。[181] M. Jiang、J. Shang、T. Cassidy、X. Ren、L. M. Kaplan、T. P. Hanratty 和 J. Han，《MetaPAD：从海量文本语料库中发现元模式》，载于 KDD，2017年，第 877–886 页，2017年。[182] Y. Ahmad、T. Antoniu、S. Goldwater 和 S. Krishnamurthi，《用于静态检测电子表格错误的类型系统》，载于 ASE，2003年，第 174–183 页，2003年。[183] Y. A. Sekhavat、F. D. Paolo、D. Barbosa 和 P. Merialdo，《利用表格数据增强知识库》，载于 WWW 论文集，CEUR 研讨会论文集第 1184 卷，2014年。[184] E. Muñoz、A. Hogan 和 A. Mileo，《利用关联数据从维基百科表格中挖掘 RDF》，载于 WSDM，2014年，第 533–542 页，2014年。[185] S. Krause、H. Li、H. Uszkoreit 和 F. Xu，《利用来自网页的远程监督进行关系抽取规则的大规模学习》，载于 ISWC，2012年，论文集，第 I 部分，第 7649 卷，第 263–278 页，2012年。

[186] L. Cui、F. Wei 和 M. Zhou，《神经开放信息抽取》，载于 ACL 论文集，2018年，第 2 卷：短篇论文，第 407–413 页，2018年。[187] K. Kolluru、S. Aggarwal、V. Rathore、Mausam 和 S. Chakrabarti，《IMOJIE：基于迭代记忆的联合开放信息抽取》，载于 ACL 论文集，2020年，第 5871–5886 页，2020年。[188] R. Wu、Y. Yao、X. Han、R. Xie、Z. Liu、F. Lin、L. Lin 和 M. Sun，《开放关系抽取：将关系知识从有监督数据迁移到无监督数据》，载于 EMNLP-IJCNLP 论文集，2019年，第 219–228 页，2019年。[189] X. Han、P. Yu、Z. Liu、M. Sun 和 P. Li，《利用由粗到细的注意力进行层次化关系抽取》，载于 EMNLP 论文集，2018年，第 2236–2245 页，2018年。[190] B. Luo、Y. Feng、Z. Wang、Z. Zhu、S. Huang、R. Yan 和 D. Zhao，《噪声学习：利用动态转移矩阵增强远程监督关系抽取》，载于 ACL，2017年，第 1 卷：长篇论文，第 430–439 页，2017年。[191] Y. Huang 和 J. Du，《自注意力增强卷积神经网络与用于远程监督关系抽取的协同课程学习》，载于 EMNLP-IJCNLP 论文集，2019年，第 389–398 页，2019年。[192] P. Qin、W. Xu 和 W. Y. Wang，《通过深度强化学习实现稳健的远程监督关系抽取》，载于 ACL 论文集，2018年，第 1 卷：长篇论文，第 2137–2147 页，2018年。[193] P. Qin、W. Xu 和 W. Y. Wang，《DSGAN：用于远程监督关系抽取的生成对抗训练》，载于 ACL 论文集，2018年，第 1 卷：长篇论文，第 496–505 页，2018年。[194] X. Jiang、Q. Wang、P. Li 和 B. Wang，《利用多实例多标签卷积神经网络进行关系抽取》，载于 COLING 会议论文集：技术论文，2016年，第 1471–1480 页，2016年。[195] G. Ji、K. Liu、S. He 和 J. Zhao，《利用句子级注意力和实体描述进行关系抽取的远程监督》，载于 AAAI-17 论文集，2017年，第 3060–3066 页，2017年。[196] Y. Lin、S. Shen、Z. Liu、H. Luan 和 M. Sun，《通过对实例进行选择性注意力的神经关系抽取》，载于 ACL 论文集，2016年，第 1 卷：长篇论文，2016年。[197] Z. Ye 和 Z. Ling，《利用包内与包间注意力进行远程监督关系抽取》，载于 NAACL-HLT 论文集，2019年，第 1 卷（长篇与短篇论文），第 2810–2819 页，2019年。[198] Y. Yuan、L. Liu、S. Tang、Z. Zhang、Y. Zhuang、S. Pu、F. Wu 和 X. Ren，《用于远程监督关系抽取的跨关系跨包注意力》，载于 AAAI、IAAI、EAAI，2019年，第 419–426 页，2019年。[199] N. Zhang、S. Deng、Z. Sun、G. Wang、X. Chen、W. Zhang 和 H. Chen，《通过知识图谱嵌入和图卷积网络进行长尾关系抽取》，载于 NAACL-HLT 论文集，2019年，第 1 卷（长篇与短篇论文），第 3016–3025 页，2019年。[200] S. Vashishth、R. Joshi、S. S. Prayaga、C. Bhattacharyya 和 P. P. Talukdar，《RESIDE：利用辅助信息改进远程监督神经关系抽取》，载于 EMNLP 论文集，2018 年，第 1257–1266 页，2018 年。[201] W. Xiong 、 M. Yu 、 S. Chang 、 X. Guo 和 W. Y. Wang，《面向知识图谱的单样本关系学习》，载于 EMNLP 论文集，2018年，第 1980–1990 页，2018年。[202] J. Snell、K. Swersky 和 R. S. Zemel，《面向小样本学习的原型网络》，载于 NeurIPS，2017年，第 4087 页，2017年，第 4077 页。[203] M. Fan、Y. Bai、M. Sun 和 P. Li，《利用细粒度特征进行小样本关系分类的大间隔原型网络》，载于 CIKM 论文集，2019年，第 2353–2356 页，2019年。[204] T. Gao、X. Han、Z. Liu 和 M. Sun，《用于噪声少样本关系分类的混合注意力原型网络》，载于 AAAI、IAAI、EAAI，2019年，第 6407–6414 页，2019年。[205] O. Vinyals、C. Blundell、T. Lillicrap、K. Kavukcuoglu 和 D. Wierstra，《用于单样本学习的匹配网络》，载于 NeurIPS，2016年，第 3630–3638 页，2016年。[206] T. Gao、X. Han、R. Xie、Z. Liu、F. Lin、L. Lin 和 M. Sun，《用于少样本关系学习的 Neural Snowball》，载于 AAAI、IAAI、EAAI，2020年，第 7772–7779 页，2020年。[207] Z. Ye 和 Z. Ling，《用于少样本关系分类的多层匹配与聚合网络》，载于 ACL 论文集，2019年，第 1 卷：长篇论文，第 2872–2881 页，2019年。[208] C. Finn、P. Abbeel 和 S. Levine，《用于深度网络快速适应的模型无关元学习》，载于国际机器学习大会论文集，2017年，第 70 卷，第 1126–1135 页，2017年。[209] T. Munkhdalai 和 H. Yu，《元网络》，载于国际机器学习大会，2017年，第 70 卷，第 2554–2563 页，2017年。[210] T. Wu、X. Li、Y. Li、G. Haffari、G. Qi、Y. Zhu 和 G. Xu，《面向顺序鲁棒持续关系抽取的课程元学习》，载于 AAAI、IAAI、EAAI，2021年，第 10363–10369 页，2021年。[211] O. Levy、M. Seo、E. Choi 和 L. Zettlemoyer，《通过阅读理解进行零样本关系抽取》，载于计算自然语言学习会议、ACL 论文集，2017年，第 333–342 页，2017年。[212] L. B. Soares、N. FitzGerald、J. Ling 和 T. Kwiatkowski，《匹配空白：关系学习的分布相似度》，载于 ACL 论文集，2019年，第 1 卷：长篇论文，第 2895–2905 页，2019年。[213] V. G. Satorras 和 J. B. Estrach，《利用图神经网络进行小样本学习》，载于 ICLR，2018年，会议论文集，2018年。

[214] M. Qu、T. Gao、L. A. C. Xhonneux 和 J. Tang，《基于关系图贝叶斯元学习的小样本关系抽取》，载于 ICML 论文集，2020 年，第 119 卷，第 7867–7876 页，2020 年。[215] T. Gao、X. Han、H. Zhu、Z. Liu、P. Li、M. Sun 和 J. Zhou，《FewRel 2.0：迈向更具挑战性的小样本关系分类》，载于 EMNLP-IJCNLP 论文集，2019 年，第 6249–6254 页，2019 年。[216] D. Roth 和 W. Yih，《自然语言任务中全局推理的线性规划形式化》，载于 HLT-NAACL，2004 年，第 1–8 页，2004 年。[217] Q. Li 和 H. Ji，《实体提及与关系的增量式联合抽取》，载于 ACL，2014 年，第 1 卷：长篇论文，第 402–412 页，2014 年。[218] S. Pawar、P. Bhattacharyya 和 G. K. Palshikar，《使用神经网络和马尔可夫逻辑网络进行端到端关系抽取》，载于计算语言学协会欧洲分会第 15 届会议论文集，EACL 2017，西班牙瓦伦西亚，2017 年4月3日至7日，第 1 卷：长篇论文（M. Lapata、P. Blunsom 和 A. Koller 编），第 818–827 页，计算语言学协会，2017 年。[219] S. Zheng、Y. Hao、D. Lu、H. Bao、J. Xu、H. Hao 和 B. Xu，《基于混合神经网络的实体与关系联合抽取》，《神经计算》，第 257 卷，第 59–66 页，2017 年。[220] T. Fu、P. Li 和 W. Ma，《GraphRel：将文本建模为关系图以进行实体与关系联合抽取》，载于 ACL 论文集，2019 年，第 1 卷：长篇论文，第 1409–1418 页，2019 年。[221] S. Zheng、F. Wang、H. Bao、Y. Hao、P. Zhou 和 B. Xu，《基于新型标注方案的实体和关系联合抽取》，载于 ACL 2017 论文集，第 1 卷：长篇论文，第 1227–1236 页，2017 年。[222] Z. Wei、J. Su、Y. Wang、Y. Tian 和 Y. Chang，《用于关系三元组抽取的新型级联二元标注框架》，载于 ACL 论文集，2020 年，第 1476–1488 页，2020 年。[223] Y. Wang、B. Yu、Y. Zhang、T. Liu、H. Zhu 和 L. Sun，《TPLinker：通过词元对链接进行单阶段实体与关系联合抽取》，载于 COLING 论文集，2020 年，第 1572–1582 页，2020 年。[224] G. Bekoulis、J. Deleu、T. Demeester 和 C. Develder，《将实体识别与关系抽取视为多头选择问题》，《专家系统及其应用》，第 114 卷，第 34–45 页，2018 年。[225] X. Li、F. Yin、Z. Sun、X. Li、A. Yuan、D. Chai、M. Zhou 和 J. Li，《将实体-关系抽取视为多轮问答》，载于 ACL 论文集，2019 年，第 1 卷：长篇论文，第 1340–1350 页，2019 年。[226] H. Chen、C. Zhang、J. Li、P. S. Yu 和 N. Jing，《KGGen：一种用于初始知识图谱填充的生成式方法》，《IEEE知识与数据工程汇刊》，第 34 卷，第 5 期，第 2254–2267 页，2022 年。[227] X. Ren、Z. Wu、W. He、M. Qu、C. R. Voss、H. Ji、T. F. Abdelzaher 和 J. Han，《CoType：结合知识库的类型化实体与关系联合抽取》，载于 WWW 论文集，2017 年，第 1015–1024 页，2017 年。[228] A. Bordes、N. Usunier、A. García-Durán、J. Weston 和 O. Yakhnenko，《用于多关系数据建模的嵌入转换》，载于神经信息处理系统大会论文集，2013 年，第 2787–2795 页，2013 年。[229] N. Peng、H. Poon、C. Quirk、K. Toutanova 和 W. Yih，《使用图长短期记忆网络进行跨句 n 元关系抽取》，《计算语言学协会汇刊》，第 5 卷，第 101–115 页，2017 年。[230] L. Song、Y. Zhang、Z. Wang 和 D. Gildea，《使用图状态LSTM进行 n 元关系抽取》，载于 EMNLP 论文集，2018 年，第 2226–2235 页，2018 年。[231] S. K. Sahu、F. Christopoulou、M. Miwa 和 S. Ananiadou，《使用文档级图卷积神经网络进行句间关系抽取》，载于 ACL 论文集，2019 年，第 1 卷：长篇论文，第 4309–4316 页，2019 年。[232] F. Christopoulou、M. Miwa 和 S. Ananiadou，《连接点滴：使用边导向图进行文档级神经关系抽取》，载于 EMNLP-IJCNLP 论文集，2019 年，第 4924–4935 页，2019 年。[233] H. Zhu、Y. Lin、Z. Liu、J. Fu、T. Chua 和 M. Sun，《使用生成参数的图神经网络进行关系抽取》，载于 ACL 论文集，2019 年，第 1 卷：长篇论文，第 1331–1339 页，2019 年。[234] G. Nan、Z. Guo、I. Sekulic 和 W. Lu，《通过潜在结构精炼进行文档级关系抽取推理》，载于 ACL 论文集，2020 年，第 1546–1557 页，2020 年。[235] W. Xu、K. Chen 和 T. Zhao，《通过重构进行文档级关系抽取》，载于 AAAI、IAAI、EAAI，2021 年，第 14167–14175 页，2021 年。[236] S. Zeng、R. Xu、B. Chang 和 L. Li，《用于文档级关系抽取的基于双图的推理》，载于 EMNLP 论文集，2020 年，第 1630–1640 页，2020 年。[237] Z. Zhang、B. Yu、X. Shu、M. Xue、T. Liu 和 L. Guo，《从 “ 是什么 ” 到 “ 为什么 ”：利用推理依据图改进关系抽取》，载于《计算语言学协会研究成果：ACL/IJCNLP》，2021 年，第 95，2021 年 2021 年，第 86 卷。ACL/IJCNLP，第 – 页。[238] Y. Luan、D. Wadden、L. He、A. Shah、M. Ostendorf 和 H. Hajishirzi，《使用动态跨度图进行信息抽取的通用框架》，载于 NAACL-HLT 论文集，2019 年，第 1 卷（长篇和短篇论文），第 3036–3046 页，2019 年。[239] W. Xu、K. Chen 和 T. Zhao，《用于文档级关系抽取的判别式推理》，载于 ACL/IJCNLP，2021 年，第 1653–1663 页，2021 年。[240] A. Bosselut、H. Rashkin、M. Sap、C. Malaviya、A. Celikyilmaz 和 Y. Choi，《COMET：用于自动知识图谱构建的常识变换器》，载于 ACL 论文集，2019 年，第 1 卷：长篇论文，第 4762–4779 页，2019 年。

[241] N. Zhang、X. Chen、X. Xie、S. Deng、C. Tan、M. Chen、F. Huang、L. Si 和 H. Chen，《将文档级关系抽取视为语义分割》，载于 IJCAI，2021 年，第 3999–4006 页，2021 年。[242] W. Zhou、K. Huang、T. Ma 和 J. Huang，《结合自适应阈值化与局部上下文池化的文档级关系抽取》，载于 AAAI 、 IAAI 、 EAAI，2021 年，第 14612–14620 页，2021 年。[243] Y. Lin、Z. Liu、M. Sun、Y. Liu 和 X. Zhu，《学习用于知识图谱补全的实体和关系嵌入》，载于 AAAI-15 论文集，2015 年，第 2181–2187 页，2015 年。[244] Z. Wang、J. Zhang、J. Feng 和 Z. Chen，《通过在超平面上进行平移实现知识图谱嵌入》，载于 AAAI-14 论文集，2014 年，第 1112–1119 页，2014 年。[245] G. Ji、K. Liu、S. He 和 J. Zhao，《使用自适应稀疏迁移矩阵进行知识图谱补全》，载于 AAAI-16 论文集，2016 年，第 985–991 页，2016 年。[246] M. Nickel、V. Tresp 和 H. Kriegel，《用于多关系数据集体学习的三向模型》，载于 ICML 论文集，2011 年，第 809–816 页，2011 年。[247] I. Balazevic、C. Allen 和 T. M. Hospedales，《Tucker：用于知识图谱补全的张量分解》，载于 EMNLP-IJCNLP 论文集，2019 年，第 5184–5193 页，2019 年。[248] B. Yang、W. Yih、X. He、J. Gao 和 L. Deng，《用于知识库学习与推理的实体和关系嵌入》，载于 ICLR，2015 年，会议论文集，2015 年。[249] R. Socher、D. Chen、C. D. Manning 和 A. Y. Ng，《使用神经张量网络进行知识库补全推理》，载于神经信息处理系统大会论文集，2013 年，第 926–934 页，2013 年。[250] Z. Zhang、J. Cai、Y. Zhang 和 J. Wang，《学习层次感知知识图谱嵌入以进行链接预测》，载于 AAAI、IAAI、EAAI，2020 年，第 3065–3072 页，2020 年。[251] G. Niu、B. Li、Y. Zhang 和 S. Pu，《CAKE：用于多视图知识图谱补全的可扩展常识感知框架》，载于 ACL 论文集，2022 年。[252] L. Wang、W. Zhao、Z. Wei 和 J. Liu，《SimKGC：结合预训练语言模型的简单对比知识图谱补全》，载于 ACL 论文集，2022 年。[253] K. Wang、Y. Liu 和 Q. Z. Sheng，《快速而可靠：用于低维知识图谱嵌入的难度感知对比学习》，载于 WWW '22：2022年ACM Web会议，虚拟会议，法国里昂，2022年4月25日至29日（F. Laforest、R. Troncy、E. Simperl、D. Agarwal、A. Gionis、I. Herman 和 L. Médini 编），第 838–849 页，ACM，2022 年。[254] A. Borrego、D. Ayala、I. Hernández、C. R. Rivero 和 D. Ruiz，《使用邻域感知特征进行知识图谱补全》，《工程应用人工智能》，第 103 卷，第 104302 页，2021 年。[255] J. Wu、W. Shi、X. Cao、J. Chen、W. Lei、F. Zhang、W. Wu 和 X. He，《DisenKGAT：使用解耦图注意力网络进行知识图谱嵌入》，载于 CIKM，2021 年，第 2140–2149 页，2021 年。[256] N. Lao 和 W. W. Cohen，《结合路径约束随机游走的关系检索》，《机器学习》，第 81 卷，第 1 期，第 53–67 页，2010 年。[257] M. Gardner、P. P. Talukdar、J. Krishnamurthy 和 T. M. Mitchell，《将向量空间相似度融入知识库上的随机游走推理》，载于 EMNLP 论文集，2014 年，SIGDAT 会议，即 ACL特别兴趣小组会议，第 397–406 页，2014 年。[258] Q. Wang、J. Liu、Y. Luo、B. Wang 和 C. Lin，《通过耦合路径排序进行知识库补全》，载于 ACL 论文集，2016 年，第 1 卷：长篇论文，2016 年。[259] W. Xiong、T. Hoang 和 W. Y. Wang，《DeepPath：一种用于知识图谱推理的强化学习方法》，载于 EMNLP 论文集，2017 年，第 564–573 页，2017 年。[260] X. V. Lin、R. Socher 和 C. Xiong，《通过奖励塑形进行多跳知识图谱推理》，载于 EMNLP 论文集，2018 年，第 3243–3253 页，2018 年。[261] Z. Li、X. Jin、S. Guan、Y. Wang 和 X. Cheng，《知识图谱上的路径推理：一种基于多智能体和强化学习的方法》，载于 ICDM Workshops，2018 年，第 929–936 页，2018 年。[262] Y. Shen、J. Chen、P. Huang、Y. Guo 和 J. Gao，《M-Walk：使用蒙特卡罗树搜索学习在图上游走》，载于神经信息处理系统大会，2018 年，第 6787–6798 页，2018 年。[263] A. Neelakantan、B. Roth 和 A. McCallum，《用于知识库补全的组合向量空间模型》，载于 ACL，2015 年，第 1 卷：长篇论文，第 156–166 页，2015 年。[264] R. Das、A. Neelakantan、D. Belanger 和 A. McCallum，《使用循环神经网络对实体、关系和文本进行推理链建模》，载于 EACL 论文集，2017 年，第 1 卷：长篇论文，第 132–141 页，2017 年。[265] Y. Zhang、H. Dai、Z. Kozareva、A. J. Smola 和 L. Song，《使用知识图谱进行问答的变分推理》，载于 AAAI、IAAI、EAAI，2018 年，第 6069–6076 页，2018 年。[266] S. Kardani-Moghaddam、R. Buyya 和 K. Ramamohanarao，《ADRL：基于异常感知深度强化学习的云环境混合资源扩展方法》，《IEEE并行与分布式系统汇刊》，第 32 卷，第 3 期，第 514–526 页，2021 年。[267] H. Wang、S. Li、R. Pan 和 M. Mao，《将图注意力机制融入基于深度强化学习的知识图谱推理》，载于 EMNLP-IJCNLP 论文集，2019 年，第 2623–2631 页，2019 年。

[268] M. Zheng、Y. Zhou 和 Q. Cui，“基于强化学习的知识图谱推理多智能体分层策略网络”，载于 KSEM，2021，论文集，第 I 部分，第 12815 卷，第 445–457 页，2021年。[269] P. Tiwari、H. Zhu 和 H. M. Pandey，“DAPath：基于深度强化学习的距离感知知识图谱推理”，《神经网络》，第 135 卷，第 1–12 页，2021年。[270] S. Li、H. Wang、R. Pan 和 M. Mao，“MemoryPath：一种将记忆组件纳入知识图谱推理的深度强化学习框架”，《神经计算》，第 419 卷，第 273–286 页，2021年。[271] L. A. Galárraga、C. Teflioudi、K. Hose 和 F. M. Suchanek，“AMIE：本体知识库中不完整证据下的关联规则挖掘”，载于 WWW，2013，第 413–422 页，2013年。[272] P. G. Omran、K. Wang 和 Z. Wang，“通过学习表示实现可扩展规则学习”，载于 IJCAI论文集，2018，第 2149–2155 页，2018年。[273] C. Meilicke、M. Fink、Y. Wang、D. Ruffinelli、R. Gemulla 和 H. Stuckenschmidt，“对基于规则和基于嵌入的知识图谱补全系统进行细粒度评估”，载于 ISWC，2018，20，2018 11136，论文集，第 I 部分，第 卷 页 – 。[274] S. Guo、Q. Wang、L. Wang、B. Wang 和 L. Guo，“联合嵌入知识图谱与逻辑规则”，载于 EMNLP论文集，2016，第 192–202 页，2016年。[275] S. Guo、Q. Wang、L. Wang、B. Wang 和 L. Guo，“通过软规则的迭代引导进行知识图谱嵌入”，载于 AAAI、IAAI、EAAI，2018，第 4816–4823 页，2018年。[276] F. Yang、Z. Yang 和 W. W. Cohen，“用于知识库推理的逻辑规则可微学习”，载于神经信息处理系统大会，2017，第 2319–2328 页，2017年。[277] M. Qu 和 J. Tang，“用于推理的概率逻辑神经网络”，载于神经信息处理系统大会，2019，第 7710–7720 页，2019年。[278] Y. Zhang、X. Chen、Y. Yang、A. Ramamurthy、B. Li、Y. Qi 和 L. Song，“利用图神经网络进行高效概率逻辑推理”，载于 ICLR，2020。[279] T. Miller，“人工智能中的解释：来自社会科学的启示”，Artificial Intelligence，第 267 卷，第 1–38 页，2019年。[280] V. I. S. Carmona、T. Rocktäschel、S. Riedel 和 S. Singh，“迈向提取潜变量模型忠实且具描述性的表示”，载于 AAAI春季研讨会，2015。[281] Y. Nandwani、A. Gupta、A. Agrawal、M. S. Chauhan、P. Singla 和 Mausam，“OXKBC：基于因子分解的知识库补全的结果解释”，载于 AKBC，2020。[282] Z. Ying、D. Bourgeois、J. You、M. Zitnik 和 J. Leskovec，“GNNExplainer：为图神经网络生成解释”，载于神经信息处理系统大会，2019，第 9240–9251 页，2019年。[283] P. Pezeshkpour、Y. Tian 和 S. Singh，“通过对抗性修改研究链接预测的鲁棒性与可解释性”，载于 NAACL-HLT论文集，2019，第 1 卷（长篇与短篇论文），第 3336–3347 页，2019年。[284] R. Xie、Z. Liu、F. Lin 和 L. Lin，“《威廉·莎士比亚真的写了《哈姆雷特》吗？基于置信度的知识表示学习”，载于 AAAI、IAAI、EAAI，2018，第 4954–4961 页，2018年。[285] T. Dong、Z. Wang、J. Li、C. Bauckhage 和 A. B. Cremers，“使用区域和细粒度实体类型判定进行三元组分类”，载于 AAAI、IAAI、EAAI，2019，第 77–85 页，2019年。[286] E. Amador-Domínguez、E. Serrano、D. Manrique、P. Hohenecker 和 T. Lukasiewicz，“一种基于本体的深度学习方法，用于对知识库外实体进行三元组分类”，信息科学，第 564 卷，第 85–102 页，2021年。[287] D. Q. Nguyen、T. Nguyen 和 D. Phung，“一种基于关系记忆的嵌入模型，用于三元组分类和搜索个性化”，载于 ACL论文集，2020，第 3429–3435 页，2020年。[288] T. Sun、J. Zhai 和 Q. Wang，“Novea：一种使用属性三元组和关系三元组进行实体对齐的新模型”，载于 KSEM，2020，论文集，第 I 部分，第 12274 卷，第 161–173 页，2020年。[289] F. He、Z. Li、Q. Yang、A. Liu、G. Liu、P. Zhao、L. Zhao、M. Zhang 和 Z. Chen，“使用属性三元组和关系三元组进行无监督实体对齐”，载于 DASFAA，2019，论文集，第 I 部分，第 11446 卷，第 367–382 页，2019年。[290] H. Yang、Y. Zou、P. Shi、W. Lu、J. Lin 和 X. Sun，“利用多方面信息对齐跨语言实体”，载于 EMNLP-IJCNLP论文集，2019，第 4430–4440 页，2019年。[291] Z. Sun、W. Hu 和 C. Li，“通过联合属性保持嵌入实现跨语言实体对齐”，载于 ISWC，2017，论文集，第 I 部分，第 10587 卷，第 628–644 页，2017年。[292] B. D. Trisedya、J. Qi 和 R. Zhang，“使用属性嵌入对齐知识图谱中的实体”，载于 AAAI、IAAI、EAAI，2019，第 297–304 页，2019年。[293] W. Tan，“技术视角：使用 Magellan 进行实体匹配”，《ACM通讯》，第 63 卷，第 8 期，第 82 页，2020年。[294] M. S. Schlichtkrull 和 H. M. Alonso，“MSEJRKU 在 SemEval-2016任务14中的表现：通过证据排序扩充分类体系”，载于 SemEval@NAACL-HLT论文集，2016，第 1337–1341 页，2016年。[295] S. Mudgal、H. Li、T. Rekatsinas、A. Doan、Y. Park、G. Krishnan、R. Deep、E. Arcaute 和 V. Raghavendra，“用于实体匹配的深度学习：设计空间探索”，载于 SIGMOD论文集，2018，第 19–34 页，2018年。[296] D. Jurgens 和 M. T. Pilehvar，“SemEval-2016任务14：语义分类体系扩充”，载于 SemEval@NAACL-HLT论文集，2016，第 1092–1102 页，2016年。

[297] B. T. McInnes，“VCU 在 SemEval-2016任务14中的表现：评估基于定义的语义分类体系扩充相似度度量”，载于 SemEval@NAACL-HLT论文集，2016，第 1351–1355 页，2016年。[298] L. E. Anke、F. Ronzano 和 H. Saggion，“TALN 在 SemEval-2016任务14中的表现：通过基于词义的嵌入扩充语义分类体系”，载于 SemEval@NAACL-HLT论文集，2016，第 1332–1336 页，2016年。[299] K. S. Jones，“术语特异性的统计解释及其在检索中的应用”，《文献学杂志》，第 60 卷，第 5 期，第 493–502 页，2004年。[300] G. Stoilos、G. B. Stamou 和 S. D. Kollias，“一种用于本体对齐的字符串度量”，载于 ISWC，2005，论文集，第 3729 卷，第 624–637 页，2005年。[301] N. Vedula、P. K. Nicholson、D. Ajwani、S. Dutta、A. Sala 和 S. Part hasarathy，“利用功能领域知识扩充分类体系”，载于 SIGIR，2018，第 745–754 页，2018年。[302] D. Frolov、S. Nascimento、T. I. Fenner 和 B. G. Mirkin，“使用分类树泛化模糊主题簇”，载于 FUZZ-IEEE，2019，第 1–6 页，2019年。[303] F. Scharffe、Y. Liu 和 C. Zhou，“RDF-AI：一种用于 RDF 数据集匹配、融合和互联的架构”，载于 IJCAI 关于 IR-KR 的研讨会论文集，2019，第 23 页，2009年。[304] A. N. Ngomo 和 S. Auer，“LIMES——一种在数据网上进行大规模链接发现的高效方法”，载于 IJCAI论文集，2011，第 2312–2317 页，2011年。[305] M. Pershina、M. Yakout 和 K. Chakrabarti，“跨知识图谱的整体式实体匹配”，载于 IEEE BigData，2015，第 1585–1590 页，2015 年。[306] M. Berrendorf 、 E. Faerman 和 V. Tresp，“ 用于实体对齐的主动学习 ”，载于 ECIR，2021，论文集，第 I 部分，第 12656 卷，第 48–62 页，2021年。[307] Z. Sun、W. Hu、Q. Zhang 和 Y. Qu，“利用知识图谱嵌入自举实体对齐”，载于 IJCAI论文集，2018，第 4396–4402 页，2018年。[308] H. Zhu、R. Xie、Z. Liu 和 M. Sun，“通过联合知识嵌入实现迭代式实体对齐”，载于 IJCAI论文集，2017，第 4258–4264 页，2017年。[309] Q. Zhang、Z. Sun、W. Hu、M. Chen、L. Guo 和 Y. Qu，“用于实体对齐的多视图知识图谱嵌入”，载于 IJCAI论文集，2019，第 5429–5435 页，2019年。[310] M. Chen、Y. Tian、M. Yang 和 C. Zaniolo，“用于跨语言知识对齐的多语言知识图谱嵌入”，载于 IJCAI论文集，2017，第 1511–1517 页，2017年。[311] Z. Huang、Z. Li、H. Jiang、T. Cao、H. Lu、B. Yin、K. Subbian、Y. Sun 和 W. Wang，“通过自监督自适应图对齐实现多语言知识图谱补全”，载于 ACL论文集，2022年。[312] X. Liu、H. Hong、X. Wang、Z. Chen、E. Kharlamov、Y. Dong 和 J. Tang，“SelfKG：知识图谱中的自监督实体对齐”，载于 WWW ’22：2022年ACM Web会议，线上活动，法国里昂，2022年4月25日至29日（F. Laforest、R. Troncy、E. Simperl、D. Agarwal、A. Gionis、I. Herman 和 L. Médini 编），第 860–870 页，ACM，2022年。[313] M. Chen、Y. Tian、K. Chang、S. Skiena 和 C. Zaniolo，“协同训练知识图谱嵌入与实体描述，以实现跨语言实体对齐”，载于 IJCAI 论文集，2018，第 3998–4004 页，2018 年。[314] B. Chen 、 J. Zhang 、 X. Tang 、 H. Chen 和 C. Li，“JARKA：对跨语言知识对齐中的属性交互进行建模”，载于 PAKDD，2020，论文集，第 I 部分，第 12084 卷，第 845–856 页，2020年。[315] Z. Wang、Q. Lv、X. Lan 和 Y. Zhang，“通过图卷积网络实现跨语言知识图谱对齐”，载于 EMNLP论文集，2018，第 349–357 页，2018年。[316] Y. Zhu、H. Liu、Z. Wu 和 Y. Du，“用于实体对齐的关系感知邻域匹配模型”，载于 AAAI、IAAI、EAAI，2021，第 4749–4756 页，2021年。[317] Y. Wu、X. Liu、Y. Feng、Z. Wang、R. Yan 和 D. Zhao，“异构知识图谱的关系感知实体对齐”，载于 IJCAI论文集，2019，第 5278–5284 页，2019年。[318] K. Xu、L. Wang、M. Yu、Y. Feng、Y. Song、Z. Wang 和 D. Yu，“通过图匹配神经网络实现跨语言知识图谱对齐”，载于 ACL论文集，2019，第 1 卷：长篇论文，第 3156–3161 页，2019年。[319] Z. Liu、Y. Cao、L. Pan、J. Li 和 T. Chua，“探索和评估实体对齐中的属性、值与结构”，载于 EMNLP论文集，2020，第 6355–6364 页，2020年。[320] Y. Cao、Z. Liu、C. Li、Z. Liu、J. Li 和 T. Chua，“用于实体对齐的多通道图神经网络”，载于 ACL论文集，2019，第 1 卷：长篇论文，第 1452–1461 页，2019年。[321] S. Jiang、T. Nie、D. Shen、Y. Kou 和 G. Yu，“通过联合图注意力和平移表示实现知识图谱实体对齐”，载于 WISA，2021，论文集，第 12999 卷，第 347–358 页，2021年。[322] T. Jiang、T. Zhao、B. Qin、T. Liu、N. V. Chawla 和 M. Jiang，“ ¨条件¨的作用：一种新颖的科学知识图谱表示与构建模型”，载于 KDD，2019，第 1634–1642 页，2019年。[323] T. Zheng、Z. Xu、Y. Li、Y. Zhao、B. Wang 和 X. Yang，“一种新颖的条件知识图谱表示与构建方法”，载于 CICAI，2021，论文集，第 II 部分，第 13070 卷，第 383–394 页，2021年。[324] F. Cheng 和 Y. Miyao，“利用依存路径上的双向长短期记忆网络对时间关系进行分类”，载于 ACL论文集，2017，第 2 卷：短篇论文，第 1–6 页，2017年。

[325] Y. Meng、A. Rumshisky 和 A. Romanov，“在基于 LSTM 的架构中利用句法依存关系进行问答的时间信息抽取”，载于 EMNLP 论文集，2017年，第 887–896 页，2017年。[326] S. Vashishtha、B. V. Durme 和 A. S. White，“细粒度时间关系抽取”，载于 ACL 论文集，2019年，第 1 卷：长篇论文，第 2906–2919 页，2019年。[327] P. Mathur、R. Jain、F. Dernoncourt、V. I. Morariu、Q. H. Tran 和 D. Manocha，“TIMERS：文档级时间关系抽取”，载于 ACL/IJCNLP，2021年，第 2 卷：短篇论文，第 524–533 页，2021年。[328] J. Leblay 和 M. W. Chekol，“推导知识图谱中的有效时间”，载于 WWW，2018年，第 1771–1776 页，2018年。[329] S. S. Dasgupta、S. N. Ray 和 P. P. Talukdar，“HyTE：基于超平面的时间感知知识图谱嵌入”，载于 EMNLP 论文集，2018年，第 2001–2011 页，2018年。[330] A. García-Durán、S. Dumancic 和 M. Niepert，“学习用于时序知识图谱补全的序列编码器”，载于 EMNLP 论文集，2018年，第 4816–4821 页，2018年。[331] Y. Liu、W. Hua、K. Xin 和 X. Zhou，“上下文感知的时间知识图谱嵌入”，载于 WISE，2019年，论文集，第 11881 卷，第 583–598 页，2019年。[332] L. Lin 和 K. She，“基于张量分解的时间知识图谱嵌入”，载于 ICTAI，2020年，975，2020年 969 页。[333] T. Lacroix、G. Obozinski 和 N. Usunier，“用于时间知识库补全的张量分解”，载于 ICLR，2020 年。[334] P. Shao 、 D. Zhang 、 G. Yang 、 J. Tao 、 F. Che 和 T. Liu，“ 基于 Tucker 分解的时间知识图谱补全”，《知识型系统》，第 238 卷，页码 107841，2022年。[335] W. Radstok、M. Chekol 和 Y. Velegrakis，“利用静态模型进行时序知识图谱链接预测”，载于 ICTAI，2021年，第 1034–1041 页，2021年。[336] J. Jung、J. Jung 和 U. Kang，“学习穿越时间进行可解释的时序知识图谱补全”，载于 KDD，2021年，第 786–795 页，2021年。[337] H. Liu、S. Zhou、C. Chen、T. Gao、J. Xu 和 M. Shu，“基于深度强化学习的动态知识图谱推理”，《知识型系统》，第 241 卷，页码 108235，2022年。[338] T. Jiang、T. Liu、T. Ge、L. Sha、S. Li、B. Chang 和 Z. Sui，“为时间感知链接预测编码时间信息”，载于 EMNLP 论文集，2016年，第 2350–2354 页，2016年。[339] K. Yu、X. Guo、L. Liu、J. Li、H. Wang、Z. Ling 和 X. Wu，“基于因果关系的特征选择：方法与评估”，《ACM计算调查》，第 53 卷，第 5 期，第 111:1–111:36 页，2020年。[340] R. Trivedi、H. Dai、Y. Wang 和 L. Song，“Know-Evolve：动态知识图谱的深度时间推理”，载于 2017年国际机器学习大会，第 70 卷，第 3462–3471 页，2017年。[341] W. Jin、C. Zhang、P. A. Szekely 和 X. Ren，“用于时序知识图谱推理的循环事件网络”，CoRR，卷 abs/1904.05530，2019年。[342] X. Li、L. Liu、X. Wang、Y. Li、Q. Wu 和 T. Qian，“大数据背景下演化知识表示的探索”，《电子图书馆》，第 39 卷，第 3 期，第 392–410 页，2021年。[343] Z. Han、P. Chen、Y. Ma 和 V. Tresp，“DyERNIE：用于时序知识图谱补全的黎曼流形嵌入动态演化”，载于 EMNLP，2020年，第 7301–7316 页，2020年。[344] T. Gracious、S. Gupta、A. Kanthali、R. M. Castro 和 A. Dukkipati，“用于动态网络和时序知识图谱的神经潜在空间模型”，载于 AAAI、IAAI、EAAI，2021年，第 4054–4062 页，2021 年。[345] Y. Yan 、 L. Liu 、 Y. Ban 、 B. Jing 和 H. Tong，“ 动态知识图谱对齐 ”，载于 AAAI、IAAI、EAAI，2021年，第 4564–4572 页，2021年。[346] B. Momjian，《PostgreSQL：简介与概念》，第 192 卷。Addison-Wesley，纽约，2001年。[347] B. Shao、H. Wang 和 Y. Li，“Trinity：内存云上的分布式图引擎”，载于 SIGMOD，2013年，516，2013年 505 页。[348] J. C. Anderson、J. Lehnardt 和 N. Slater，“CouchDB：权威指南——是时候放松了”，2010年。[349] L. Zou、J. Mo、L. Chen、M. T. Özsu 和 D. Zhao，“gStore：通过子图匹配回答 SPARQL 查询”，《VLDB 学报》，第 4 卷，第 8 期，第 482–493 页，2011年。[350] J. Webber，“Neo4j 编程入门”，载于 SPLASH，论文集，2012年，第 217–218 页，2012年。[351] T. Miller、C. Hempelmann 和 I. Gurevych，“SemEval-2017 任务 7：英语双关语的检测与解读”，载于 SemEval@ACL 论文集，2017年，第 58–68 页，2017年。[352] C. Wang、S. Liang、Y. Jin、Y. Wang、X. Zhu 和 Y. Zhang，“SemEval-2020 任务 4：常识验证与解释”，载于 SemEval@COLING 论文集，2020年，第 307–321 页，2020年。[353] Y. Liu、H. Li、A. García-Durán、M. Niepert、D. Oñoro-Rubio 和 D. S. Rosenblum，“多模态知识图谱”，载于 ESWC，2019年，论文集，第 11503 卷，第 459–474 页，2019年。[354] S. Dost、L. Serafini、M. Rospocher、L. Ballan 和 A. Sperduti，“对齐并链接图像、文本和知识库中的实体提及”，《数据与知识工程》，第 138 卷，页码 101975，2022年。

[355] H. Peng、H. Li、Y. Song、V. W. Zheng 和 J. Li，“差分隐私联邦知识图谱嵌入”，载于 *CIKM，2021*，第1416–1425 页，2021年。 [356] M. Chen、W. Zhang、Z. Yuan、Y. Jia 和 H. Chen，“FedE：在联邦设置中嵌入知识图谱”，载于 *IJCKG，2021*，第 80–88 页，2021年。 [357] Y. Xie、J. Shen、S. Li、Y. Mao 和 J. Han，“Eider：证据增强的文档级关系抽取”，*CoRR*，卷 abs/2106.08657，2021年。[358] H. Weld、X. Huang、S. Long、J. Poon 和 S. C. Han，“自然语言理解中的联合意图检测与槽填充模型综述”，*CoRR*，卷 abs/2101.08091，2021年。