

面向知识图谱的信息抽取技术综述^①



姜磊, 刘琦, 赵肄江, 袁鹏, 李媛, 邹子维

(湖南科技大学 计算机科学与工程学院, 湘潭 411100)

通信作者: 姜磊, E-mail: jleihn@126.com

摘要: 互联网时代, 数据呈爆发式的增长, 怎样从这些数据中抽取有用的信息, 已是人工智能研究中的一个核心问题. 知识图谱作为解决这一问题的重要方法, 已成为人工智能技术发展的核心推动力. 信息抽取是知识图谱构建过程中的首要环节, 它实现了从海量的数据中抽取结构化实体以及实体之间的关系. 本文探讨知识图谱中信息抽取的发展趋势, 对实体抽取、关系抽取和事件抽取及其关键技术进行了综述, 分析和讨论了当前存在的问题、挑战以及未来发展的方向.

关键词: 知识图谱; 信息抽取; 实体抽取; 关系抽取; 事件抽取

引用格式: 姜磊, 刘琦, 赵肄江, 袁鹏, 李媛, 邹子维. 面向知识图谱的信息抽取技术综述. 计算机系统应用, 2022, 31(7): 46-54. <http://www.c-s-a.org.cn/1003-3254/8590.html>

Review on Information Extraction Techniques for Knowledge Graph

JIANG Lei, LIU Qi, ZHAO Yi-Jiang, YUAN Peng, LI Yuan, ZOU Zi-Wei

(School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411100, China)

Abstract: How to extract useful information from surging data has become a critical issue confronting artificial intelligence in the Internet age. As an important method, knowledge graph has become the main driving force to promote the development of artificial intelligence technology. Information extraction realizes the extraction of structured entities and their relationships from massive data, which is the primary step in constructing a knowledge graph. This study discusses the development trend of information extraction in knowledge graphs, as well as entity extraction, relationship extraction, event extraction, and key technologies. Finally, it analyzes and discusses the current problems, challenges, and future development.

Key words: knowledge graph (KG); information extraction; entity extraction; relationship extraction; event extraction

随着信息时代的到来, 数据呈爆发式的增长, 如何从这些数据中通过智能技术自动提取出真正有价值的信息, 尤为重要. 知识图谱^[1]是一类知识表示, 由实体、关系以及属性构成^[2]. 实体又称为类或实例, 是不依附于其他东西而存在的, 比如人, 机构等. 关系表示实体, 实体属性之间的关系. 属性是用来描述实体的某种特征, 比如身高, 体重等. 知识图谱技术在数据分析、智能搜索、决策支持以及医疗健康等领域越来越发挥出主要的作用.

知识图谱的构建过程: 首先从数据源中提取出碎片化事实^[3], 然后对碎片化事实进行知识的融合, 再经过知识加工后通过迭代更新建立基于知识的体系^[4]. 显然, 知识图谱的构建过程包括: 信息抽取 (information extraction), 知识融合, 知识加工^[5]等. 信息抽取作为知识图谱的主要组成, 从数据源中抽取实体和实体之间的关系等结构化信息^[6]. 包括对实体、关系以及事件等方面的抽取^[7]. 信息抽取的正确性对知识图谱的后续构建质量和效率产生影响.

① 基金项目: 国家自然科学基金面上项目 (41871320); 教育部人文社科规划项目 (17YJAZH032); 湖南省教育厅创新平台开放基金 (20K050)

收稿时间: 2021-10-18; 修改时间: 2021-11-17; 采用时间: 2021-11-30; csa 在线出版时间: 2022-05-31

信息抽取通常采用基于 NLP 和文本挖掘的方法。在信息抽取研究的成果中我们发现当前的研究在提升信息抽取准确度的基础上主要围绕如何减少人工标注语料、人工提取特征以及人工构建模式展开。这使得知识图谱中的信息抽取面对着 3 个挑战: 第一, 如何从需要构建知识图谱的领域语料中利用启发式的信息来发现语料中的隐含知识, 从而在较少的人工标注中获得较高的准确度。第二, 如何解决已有知识图谱中部分不完整的实体、关系与事件信息所带来的噪声与语义漂移问题。第三, 如何在开放领域中利用现有的标注或加上较少的标注实现知识图谱在新信息中的完善与更新。

本文综述了知识图谱中信息抽取技术, 详细地描述了近年来在实体、关系和事件抽取中的各项技术, 探讨这些技术在解决上述 3 大挑战的进展。以便研究者能对信息抽取技术有一个全局认识, 进而厘清技术的发展趋势与方向。期望研究者能从中汲取技术的精华和理念, 进一步推动信息抽取技术的发展。

1 实体抽取

命名实体识别 (named entity recognition, NER) 是实体抽取的别称。命名实体能在具有相似属性的一组事物中清楚的标识出一个事物。它可以理解为有文本标识的实体, 而实体 (entity) 是不依附于其他东西而存在的。在现实世界中, 实体通常分为 3 大类 7 小类。NER 是从文本中抽取实体信息元素。NER 主要有基于规则和字典的方法^[8,9], 基于监督学习的方法^[10-13]和基于深度学习的方法。近年来, 越来越多的研究者开始关注基于深度学习的 NER 方法。

1.1 基于深度学习的方法

近年来, 基于深度学习的 NER 模型逐渐占主导地位, 与传统的机器学习相比, 深度学习有助于自动发现隐藏的特征^[14], 进行特征抽取, 使得泛化能力得到了提升。Hammerton 等人^[15]是最早使用 LSTM 来进行 NER, 该模型在序列建模上具有良好的表现。

深度学习在实体抽取领域取得较好效果后, 研究人员开始在词汇级别上对其神经网络结构进行改进研究。Lample 等人^[16]通过加入 CRF 模块以优化标签序列输出, 提出了 BiLSTM-CRF 模型, 在语料库上取得了比较高的 $F1$ 值。Ma 等人^[17]在双向 LSTM-CNNs 结构上, 添加了 CRF 模块, 提出 BiLSTM-CNNs-CRF 模型, 模型能同时利用词和字符级表示。Luo 等人^[18]

提出了一种 Att-BiLSTM-CRF 模型, 该模型用于文档级实体识别, 在数据集上取得的 $F1$ 值为 91.14%。

上述模型偏重于词或字符的特征提取, 无法动态的表征上下文语境中的一词多义。为了改善这问题, Devlin 等人^[19]提出了 BERT 模型, 该模型可以使上下文语境或语义中的词得到充分的表征。Souza 等人^[20]将 BERT-CRF 模型应用于葡萄牙 NER 任务上, 获得了新的最佳 $F1$ 值。谢腾等人^[21]提出一种 BERT-BiLSTM-CRF 模型, 该模型在两个语料库上进行实验, 得到的 $F1$ 值分别是 94.65% 和 95.67%。在文献 [19] 的基础上, 百度推出了 ERNIE 模型^[22], 该模型是通过加强 BERT 的 masking 来获取知识。实验表明, 在 5 项 NER 任务上, ERNIE 刷新了榜单。微软提出了一种多任务的训练方式的 MT-DNN 模型^[23], 该模型比 BERT 更加稳定, 泛化能力更好。由卡内基梅隆大学提出的 XLNet 模型^[24]是一种通用的自回归预训练模型, 该模型解决了 BERT 在预训练时加入 [MASK] 的 token, 从而导致 pretrain 和 finetune 在训练数据上的差异。Liu 等人^[25]对 BERT 的预训练进行了仔细的评估, 提出了一种能更好地训练 BERT 的方法, 称为 RoBERTa 模型, 该模型比 BERT 之后的所有 psot-BERT 模型的效果好。Joshi 等人^[26]提出一种 SpanBERT 模型, 模型旨在更清晰的预测和表达文本跨度, 不再通过随机标记而是通过屏蔽连续的随机跨度来使得 BERT 得到扩展。谷歌提出了 ALBERT 模型^[27], 该模型在 BERT 模型的基础上, 减小了两种参数量, 通过两个参数稍减技术克服了扩展预训练模型面临的主要障碍, 使得训练更加稳定。

近几年, 在基于深度学习的方法上加入注意力机制^[28], 迁移学习^[29], 对抗学习^[30], 远程监督^[9]等热门研究技术也是 NER 中的一个研究热潮。

2 关系抽取

关系抽取 (relation extraction, RE) 旨根据实体之间的上下文语境来确定语义关系, 它为许多下游任务提供了基础支持, 比如文本理解中, 为了理解复杂的语句, 识别语句中的实体对之间的关系是至关重要的。在问答系统中, 关系抽取所得到的实体间的关系实例可以作为背景知识支撑问题的问答。在 NLP 领域中, 关系抽取最重要的应用是构建知识图谱。

2.1 基于深度学习的方法

传统的关系分类模型需要耗费大量的人力去设计

特征,而且很多隐性特征也难以定义.因此传统方法在大规模关系抽取任务中效果不佳.基于深度学习的关系抽取能够自动学习有效特征.有监督的关系抽取方法是深度学习方法中的一个主要方法,在解决人工特征选择和特征提取误差传播等问题上有不错的效果.流水线学习和联合学习是有监督的关系抽取方法主要的两种类别.基于深度学习的关系抽取的另一个主要方法是远程监督的方法,其利用已知知识库信息以减少人工处理.

2.1.1 流水线学习

流水线学习方法中的关系抽取是在实体抽取完成的基础上进行的,因此关系抽取结果的好坏与实体抽取的结果有直接关联.主要采用的方法是 CNNs 和 RNNs.其中 CNNs 有利于识别目标的结构特征.RNNs 有利于识别序列^[31].

Wang 等人^[32]提出了一种新的结构块驱动卷积神经网络新型轻量级关系提取方法.在两个数据集上进行实验,验证了该方法的有效性.Lin 等人^[33]将注意力机制引入句子级中,提出了一种纯文本的实体关系抽取方法,该方法动态地降低噪声对句子的影响,有效地提高了跨语言的一致性和互补性.深度学习模型在受到有限的标记实例的限制时,可以借助于合适的网络结构来获得良好的性能.如 Lin 等人^[34]提出了一个自训练的框架内具有多个语义异构嵌入的循环神经网络.

随着不断的改进和完善 CNNs 和 RNNs,使它们产生了许多变体,如双向长短期记忆网络 (Bi-LSTM).Xiao 等人^[35]提出了一种能从原始句子中提取信息进行关系分类的递归神经网络模型.Xu 等人^[36]提出了神经网络 SDP-LSTM 模型,该模型对句子中两个实体之间的关系进行分类.

随着 GCN 在 NLP 领域的应用,GCN 也被应用到关系抽取的研究中.Schlichtkrull 等人^[37]提出的 R-GCNs 模型,是一种关系图卷积网络.Zhang 等人^[38]提出了一种图卷积网络的扩展方法来对实体关系进行抽取.在数据集上取得的最佳结果 $F1$ 为 68.2%.优于现有的基于序列和依赖关系的神经模型.Zhu 等人^[39]提出的 GP-GNNS 模型,是一种新的带生成参数的图神经网络,该模型可以通过多跳关系推理来发现更精确的关系.在跨句子的 n 元关系中检测出 n 个实体之间的关系.典型的方法是将输入制定为文档图,集成各种句内和句间依赖关系,但是这种模型可能会使重要的

信息在分割过程中丢失,因此, Song 等人^[40]提出了一种 Graph-State LSTM 模型,来改进这个问题.为了有效地利用相关信息和忽略不相关信息, Guo 等人^[41]提出了注意力机制图卷积网络 (AGGCNs),一种直接以完全依赖树为输入模型.

流水线方法使得关系抽取能得到实体抽取的有用信息,从而提升了关系抽取的效果.但该方法也会产生错误传播,使得没有关系的两个实体之间出现关系.

2.1.2 联合学习

为了避免流水线学习中存在的问题,联合抽取将实体和关系放在同一模型中共同抽取.联合学习主要有两种类别:参数共享和标注策略.

参数共享是指模型通过共享编码层产生的共享参数来彼此依赖,最后通过训练得到全局参数^[32].Zheng 等人^[42]提出了一种用 BiLSTM-ED 模块对实体进行提取和用于关系分类的 CNN 模块组成的一种混合型神经网络模型,在 BiLSTM-ED 模块中获得的实体的上下文信息进一步传递到 CNN 模块以改进关系分类.Miwa 等人^[43]提出的模型同样是通过参数共享来联合学习, $F1$ 达到了 84.4%.上述模型实际还是分别提取实体和关系,通过参数共享机制相关联.这会出现没有关系的实体对信息.针对这个问题,Zheng 等人^[44]提出了将联合抽取任务转换为标注问题.直接提取实体及其关系,无需分别识别实体和关系.取得了不错的效果.对于之前的模型没有考虑实体关系重叠问题,Bekoulis 等人^[45]将联合抽取问题看作一个 multi-head selection (多头选择)问题,以此来解决重叠问题.Bekoulis 等人^[46]将对抗学习加入到文献 [45] 的模型中,使得模型中的词嵌入的质量更好,性能得到显著提高.

基于神经网络的联合学习除了共享参数和标注策略之外,Nayak 等人^[47]通过编解码架构的设计来实现联合抽取实体和关系.Li 等人^[48]将实体关系联合抽取的任务当作一个多轮问答问题来处理.Wei 等人^[49]设计了一种层次二进制标记框架.Sun 等人^[50]提出一种首先识别实体跨度,然后对实体类型和关系类型执行联合推理.Fu 等人^[51]提出了一个端到端的关系提取模型 GraphRel,它使用图卷积网络 (GCNs)来联合学习命名实体和关系.这些方法都取得了较好的结果.

2.1.3 远程监督的方法

在文本中,如果实体之间存在某种关联,那么就会以某种形式表现出这种关联.在这种前提下,基于远程

监督的方法,首先从文本中抽取出存在关系的实体对句子,然后将句子作为训练数据放入模型中进行关系抽取。

采用知识图谱和文本对齐方式来自动提取训练数据,减少了人工标注。但是,这些数据中会引入大量的噪声,从而引起语义漂移现象。为了减少语义漂移现象的出现, Ji 等人提出了 APCNNs 模型^[52],它在句子级别引入 attention mechanism。APCNNs 模型有可能会包含同一实体对的所有样例句子都含有大量噪声的情况。针对这一问题, Feng 等人提出了基于强化学习的关系分类模型 CNN-RL^[53],该模型能有效地处理数据中的噪声,并在句子层次上取得了较好的关系分类性能。远程监督能够自动生成大量用于关系提取的训练样本。然而,会带来两个主要的问题:不平衡的训练数据和训练数据中出现噪声,使得获取到的数据集准确率较低,影响整个关系抽取模型的性能。因此,有较大的提升空间^[54]。

2.2 基于开放领域的方法

基于开放领域的关系抽取方法,在大规模非限定类型的语料库中结合语形和语义特征自动进行关系抽取,减少了人工标注成本。TextRunner 开放信息抽取原型系统是一个面向开放领域的信息抽取框架(OIE),实体关系能够自动进行抽取,但 $F1$ 的值不太理想。在 OIE 的基础上, Wu 等人提出了 WOE 系统^[55], $F1$ 的值比 TextRunner 的 $F1$ 提高了 18%–34%,但是该系统在速度方面出现了不足。Nakashole 等人提出了 PATTY 系统^[56],用于表示实体之间二元关系的文本模式,这些模式在语义上被分类并构建成一个分类体系。该模型可以处理 Web 规模的语料库中的关系抽取。Mausam 等人^[57]提出了一种系统,该系统解决了 OIE 系统仅以动词为主的关系抽取和忽略了上下文这两个限制,有效地改善了 $F1$ 的值。TextRunner、WOE、PATTY、OLLIE 系统都属于二元的开放式关系抽取。Kraken^[58]是由 Akbik 等人提出的一个多元关系抽取系统。该系统是一种高精度 OIE 框架,比现有的 OIE 更能完整地捕获每个句子中的多元关系,但容易受到噪声和不合法的文本的影响。

基于开放域的关系抽取在二元关系抽取上的准确率和正确率有待于提高,在挖掘隐藏信息方面的提升,有助于关系的抽取。面向开放域的关系抽取方法在性能上存在不足,这给研究者留下了研究空间。

3 事件抽取

事件抽取(event extraction, EE)被定义为从文本中提取出对人类有用的信息事件,并以结构化的形式表示出来。例如从“李华 1922 年出生于湖南长沙”文本中抽取事件{类型: 出生, 人物: 李华, 时间: 1922 年, 出生地: 湖南长沙}。事件抽取主要的任务包括从文本中发现触发词和从文本中识别出元素扮演的角色。如图 1 和表 1 所示。

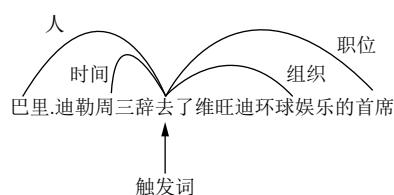


图 1 事件抽取结构分析

表 1 事件抽取任务

触发词	元素
辞去	角色 = 人 (巴里·迪勒)
	角色 = 组织 (维旺迪环球娱乐)
	角色 = 职位 (首席)
	角色 = 时间 (星期三)

事件抽取中,基于模式匹配的方法通过模式匹配算法进行事件抽取,主要的模型有 ExDisco, GenPAM 等。模式匹配方法在特定领域能取得很好的性能,但移植性差,在跨领域进行事件抽取时,需要重新构建。

在机器学习方法中,事件抽取问题转换成了分类问题。常见的分类算法有 SVM, ME 等。基于机器学习的事件抽取方法移植性能好,但是需要依赖大规模的知识库,否则可能会出现数据稀疏问题。另外,特征选取也是一个重要因素。怎样解决这两个因素,成为了机器学习方法在事件抽取研究中的重要方向。

基于深度学习的事件抽取模型主要有动态多池卷积神经网络(DMCNN)^[59],该模型能够从单词的连续及其广义表示中自动学习隐藏特征表示,解决了人工设计特征、可扩展性差以及依赖复杂 NLP 工具等问题。Nguyen 等人^[60]提出了一种双向循环神经网络的联合框架(JRNN),该模型与 DMCNN 相比, JRNN 避免了误差累计传播导致模型性能下降的问题,可以同时抽取出所有的事件信息,使用从整体结构中学到的全局特征来提升局部信息的预测能力。Chen 等人^[61]提出了一个具有门控多级注意力机制的分层和偏置标记网络

框架,该框架解决了仅利用词或者句子信息,忽略了篇章信息的问题。

对信息抽取中的实体抽取,关系抽取和事件抽取的不断研究,部分学者开始进行多任务联合学习的研究,多任务联合学习解决了各任务独立学习时忽略了依存关系问题。Lee 等人^[62]提出了一个新颖的共指解析系统,它可以跨文档的联合实体和事件,使用迭代方式构建实体和事件提及的集群,用线性回归来建模集群合并操作。Barhom 等人^[63]受文献^[62]的启示,提出了一种跨文档共指解析的神经架构模型 ECB+用来联合建模实体和事件共指。它的结果优于文献^[62]提出的模型。Xi 等人^[64]提出了一种 BERD 模型,该模型通过结合实体上下文的参数角色来预测生成参数角色,从而提升隐式参数分布模式中更准确的事件。Han 等人^[65]提出了一种 Neural SSVM 模型,该模型通过将事件和关系共享上下文嵌入来使事件的表示得到改进。Han 等人^[66]进一步提出了一个以概率领域知识构建分布约束来增强深层神经网络框架。Tang 等人^[67]提出了一种用于事件关系抽取的多层知识投影网络 (MKPNet),可以有效地利用多层话语知识进行事件关系的抽取。

事件抽取一般从属于实体、关系才有明显的意义,所以目前一般采用联合学习的方式结合实体、关系抽取所获得的信息来进一步指导事件的抽取。

4 信息抽取的研究趋势

NER、RE 与 EE 是知识图谱信息抽取的 3 个子任务。图 2 与图 3 是我们针对 2015–2020 年 NLP 领域的两个顶级会议 ACL 和 EMNLP 上的信息抽取各子任务的论文数量的统计。从图中可以看出信息抽取 3 个子任务的研究热度逐年上升。

我们将知识图谱中信息抽取的主要技术整理成表 2–表 4。如表 2 所示,实体抽取开始于基于规则和字典的方法。随后采用基于监督学习的方法进行研究,取得了大量研究成果。但是该方法需要对语料进行大量的标注,研究的方法主要围绕如何降低人工标注的数量获得准确的抽取。鉴于深度学习能够很好地发现隐藏特征,可以降低特征的人工抽取,所以目前大量的实体抽取的研究围绕深度学习展开。它在当前的研究热点在于如何引入语言学的成就在词汇、句法、语义特征等方面来寻找合适的神经网络结构来提升实体抽取的能力。

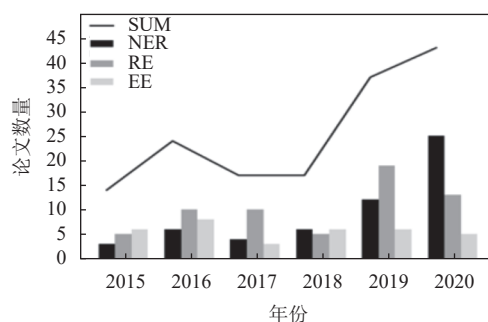


图 2 ACL 会议中信息抽取子任务的论文数量

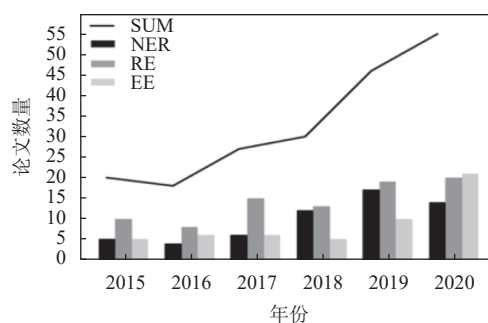


图 3 EMNLP 会议中信息抽取子任务的论文数量

表 2 实体抽取研究发展趋势

实体抽取方法	策略
早期方法	基于规则和字典的方法
基于机器学习	HMM, ME, CRF
基于深度学习	BiLSTM、CRF、CNNs结合, BERT、CRF结合, ERNIE, MT-DNN, XLNet, RoBERTa, SpanBERT, ALBERT

表 3 关系抽取研究发展趋势

关系抽取方法	策略
基于监督学习	核函数方法, 特征向量方法
基于深度学习	CNNs, RNNs, BiLSTM, GCN, 参数共享, 标注策略
基于远程监督	注意力机制, 强化学习
基于开放领域	OIE系统, WOE系统, PATTY系统, OLLIE系统

表 4 事件抽取研究发展趋势

事件抽取方法	策略
基于模式匹配	ExDisco, GenPAM,
基于机器学习	ME
基于深度学习	DMCNN, JRNN
联合学习方法	共指解析系统, ECB+系统, SSVM系统

如表 3 所示,当前关系抽取的方法主要有 4 种。研究的趋势主要是采用各种技术来降低人工提取关系特征。在这里主要分 3 种:一是利用深度学习方法具有的学习隐知识的能力,从词汇、句法结构、语句块以及引入图像处理方面的知识来改进深度学习中的神经网络。

络的结构;二是引入外部知识库中与待抽取关系中重叠知识来降低复杂度,主要采用增强学习来处理噪声与语义漂移;第三就是采用监督学习以及启发式规则等方式对开放领域中新信息的汇入造成的新关系引入、旧关系的偏移进行研究。

事件抽取主要方法如表4所示,主要分为基于模式匹配、机器学习、深度学习与联合学习的方法。研究者仍然是围绕如何降低人工标注的工作量展开。同时,为了利用实体抽取、关系抽取中获得的知识,现在将事件抽取与上述两个任务整合一起进行联合抽取已逐渐成为研究热点。

5 结束语

知识图谱构建过程中信息抽取是必不可少的环节。本文详细介绍了近年来信息抽取中实体抽取、关系抽取和事件抽取的技术进展,梳理了它们的发展趋势。在应对减少人工干预信息抽取的3大挑战中,目前的研究主要集中在针对领域语料采用深度学习进行。它表现在利用句法结构、注意力机制等语言学知识和图像处理知识来寻找合适的神经网络结构以改进深度学习。而在融合已有知识图谱中知识以及在开放领域中减少人工工作方面,目前的研究成果较少。因此,在知识图谱的信息抽取研究中继续进行深度学习的研究是一个重要方向。而引入机器学习中的降噪技术结合信息抽取的特点做已有相似实体、关系与事件的融合是一个可行的有前景的方向。另一个非常有前景的方向就是对开发领域中已有标注的语料结合新信息、新语料利用半监督学习的成果进行信息抽取的研究。希望能有更多的学者就这两个方向展开研究取得成果。

参考文献

- 1 Singhal A. Official Google blog: Introducing the knowledge Graph: Things, not strings. Official Google Blog. <https://www.blog.google/products/search/introducing-knowledge-graph-things-not/>. (2012-05-16).
- 2 Ji SX, Pan SR, Cambria E, *et al.* A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 2021: 1–21. [doi: 10.1109/TNNLS.2021.3070843]
- 3 黄恒琪, 于娟, 廖晓, 等. 知识图谱研究综述. *计算机系统应用*, 2019, 28(6): 1–12. [doi: 10.15888/j.cnki.csa.006915]
- 4 Wu XD, Chen HH, Wu GQ, *et al.* Knowledge engineering with big data. *IEEE Intelligent Systems*, 2015, 30(5): 46–55. [doi: 10.1109/MIS.2015.56]
- 5 刘焯宸, 李华昱. 领域知识图谱研究综述. *计算机系统应用*, 2020, 29(6): 1–12. [doi: 10.15888/j.cnki.csa.007431]
- 6 郭喜跃, 何婷婷. 信息抽取研究综述. *计算机科学*, 2015, 42(2): 14–17, 38. [doi: 10.11896/j.issn.1002-137X.2015.02.003]
- 7 Wu XD, Zhu XQ, Wu GQ, *et al.* Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(1): 97–107.
- 8 Mu XF, Wang W, Xu AP. Incorporating token-level dictionary feature into neural model for named entity recognition. *Neurocomputing*, 2020, 375: 43–50. [doi: 10.1016/j.neucom.2019.09.005]
- 9 Peng ML, Xing XY, Zhang Q, *et al.* Distantly supervised named entity recognition using positive-unlabeled learning. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, 2019. 2409–2419.
- 10 Azalia FY, Bijaksana MA, Huda AF. Name indexing in Indonesian translation of hadith using named entity recognition with naïve Bayes classifier. *Procedia Computer Science*, 2019, 157: 142–149. [doi: 10.1016/j.procs.2019.08.151]
- 11 Ghiasvand O, Kate RJ. Learning for clinical named entity recognition without manual annotations. *Informatics in Medicine Unlocked*, 2018, 13: 122–127.
- 12 Sintayehu H, Lehal GS. Named entity recognition: A semi-supervised learning approach. *International Journal of Information Technology*, 2021, 13(4): 1659–1665. [doi: 10.1007/s41870-020-00470-4]
- 13 Hao ZF, Lv D, Li ZJ, *et al.* Semi-supervised disentangled framework for transferable named entity recognition. *Neural Networks*, 2021, 135: 127–138. [doi: 10.1016/j.neunet.2020.11.017]
- 14 Li J, Sun AX, Han JL, *et al.* A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 2022, 34(1): 50–70.
- 15 Hammerton J. Named entity recognition with long short-term memory. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL (Volume 4)*. Stroudsburg: Association for Computational Linguistics, 2003. 172–175.
- 16 Lample G, Ballesteros M, Subramanian S, *et al.* Neural architectures for named entity recognition. *Proceedings of the*

- 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016. 260–270.
- 17 Ma XZ, Hovy EH. End-to-end sequence labeling via bi-directional LSTM-CNN-CRF. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL Press, 2016. 1064–1074.
- 18 Luo L, Yang ZH, Yang P, *et al.* An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, 2018, 34(8): 1381–1388. [doi: [10.1093/bioinformatics/btx761](https://doi.org/10.1093/bioinformatics/btx761)]
- 19 Devlin J, Chang MW, Lee K, *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186.
- 20 Souza F, Nogueira R, Lotufo R. Portuguese named entity recognition using BERT-CRF. arXiv: 1909.10649, 2019.
- 21 谢腾, 杨俊安, 刘辉. 基于 BERT-BiLSTM-CRF 模型的中文实体识别. *计算机系统应用*, 2020, 29(7): 48–55. [doi: [10.15888/j.cnki.csa.007525](https://doi.org/10.15888/j.cnki.csa.007525)]
- 22 Sun YS, Wang SH, Li YK, *et al.* ERNIE: Enhanced representation through knowledge integration. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. ACL, 2019. 1441–1451.
- 23 Liu XD, He PC, Chen WZ, *et al.* Multi-task deep neural networks for natural language understanding. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 4487–4496.
- 24 Yang ZL, Dai ZH, Yang YM, *et al.* XLNet: Generalized autoregressive pretraining for language Understanding. arXiv: 1906.08237, 2019.
- 25 Liu YH, Ott M, Goyal N, *et al.* RoBERTa: A robustly optimized BERT pretraining approach. arXiv: 1907.11692, 2019.
- 26 Joshi M, Chen DQ, Liu YH, *et al.* SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 2020, 8: 64–77. [doi: [10.1162/tacl_a_00300](https://doi.org/10.1162/tacl_a_00300)]
- 27 Lan ZZ, Chen MD, Goodman S, *et al.* ALBERT: A lite BERT for self-supervised learning of language representations. arXiv: 1909.11942, 2019.
- 28 Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
- 29 Feng XC, Feng XC, Qin B, *et al.* Improving low resource named entity recognition using cross-lingual knowledge transfer. Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm: IJCAI, 2018. 4071–4077.
- 30 Zhou JT, Zhang H, Jin D, *et al.* Dual adversarial neural transfer for low-resource named entity recognition. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 3461–3471.
- 31 李冬梅, 张扬, 李东远, 等. 实体关系抽取方法研究综述. *计算机研究与发展*, 2020, 57(7): 1424–1448. [doi: [10.7544/issn1000-1239.2020.20190358](https://doi.org/10.7544/issn1000-1239.2020.20190358)]
- 32 Wang DS, Tiwari P, Garg S, *et al.* Structural block driven enhanced convolutional neural representation for relation extraction. *Applied Soft Computing*, 2020, 86: 105913. [doi: [10.1016/j.asoc.2019.105913](https://doi.org/10.1016/j.asoc.2019.105913)]
- 33 Lin YK, Shen SQ, Liu ZY, *et al.* Neural relation extraction with selective attention over instances. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. ACL, 2016. 2124–2133.
- 34 Lin C, Miller T, Dligach D, *et al.* Self-training improves recurrent neural networks performance for temporal relation extraction. Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis. Brussels: Association for Computational Linguistics, 2018. 165–176.
- 35 Xiao MG, Liu C. Semantic relation classification via hierarchical recurrent neural network with attention. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka: The COLING 2016 Organizing Committee, 2016. 1254–1263.
- 36 Xu Y, Mou LL, Li G, *et al.* Classifying relations via long short term memory networks along shortest dependency paths. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015. 1785–1794.
- 37 Schlichtkrull M, Kipf TN, Bloem P, *et al.* Modeling relational data with graph convolutional networks. Proceedings of the 15th European Semantic Web Conference. Heraklion: Springer, 2018. 593–607.

- 38 Zhang YH, Qi P, Manning CD. Graph convolution over pruned dependency trees improves relation extraction. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 2205–2215.
- 39 Zhu H, Lin YK, Liu ZY, *et al.* Graph neural networks with generated parameters for relation extraction. Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1331–1339.
- 40 Song LF, Zhang Y, Wang ZG, *et al.* N-ary relation extraction using graph-state LSTM. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 2226–2235.
- 41 Guo ZJ, Zhang Y, Lu W. Attention guided graph convolutional networks for relation extraction. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 241–251.
- 42 Zheng SC, Hao YX, Lu DY, *et al.* Joint entity and relation extraction based on a hybrid neural network. Neurocomputing, 2017, 257: 59–66. [doi: [10.1016/j.neucom.2016.12.075](https://doi.org/10.1016/j.neucom.2016.12.075)]
- 43 Miwa M, Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin: Association for Computational Linguistics, 2016. 1105–1116.
- 44 Zheng SC, Wang F, Bao HY, *et al.* Joint extraction of entities and relations based on a novel tagging scheme. Proceedings of the 55th Annual Meeting of the Association for Computational linguistics (ACL). Vancouver: Association for Computational Linguistics, 2017. 1227–1236.
- 45 Bekoulis G, Deleu J, Demeester T, *et al.* Joint entity recognition and relation extraction as a multi-head selection problem. Expert Systems with Applications, 2018, 114: 34–45. [doi: [10.1016/j.eswa.2018.07.032](https://doi.org/10.1016/j.eswa.2018.07.032)]
- 46 Bekoulis G, Deleu J, Demeester T, *et al.* Adversarial training for multi-context joint entity and relation extraction. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 2830–2836.
- 47 Nayak T, Ng HT. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8528–8535. [doi: [10.1609/aaai.v34i05.6374](https://doi.org/10.1609/aaai.v34i05.6374)]
- 48 Li XY, Yin F, Sun ZJ, *et al.* Entity-relation extraction as multi-turn question answering. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1340–1350.
- 49 Wei ZP, Su JL, Wang Y, *et al.* A novel cascade binary tagging framework for relational triple extraction. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 1476–1488.
- 50 Sun CZ, Gong YY, Wu YB, *et al.* Joint type inference on entities and relations via graph convolutional networks. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1361–1370.
- 51 Fu TJ, Li PH, Ma WY. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1409–1418.
- 52 Ji GL, Liu K, He SZ, *et al.* Distant supervision for relation extraction with sentence-level attention and entity descriptions. Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017. 3060–3066.
- 53 Feng J, Huang ML, Zhao L, *et al.* Reinforcement learning for relation classification from noisy data. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 5779–5786.
- 54 鄂海红, 张文静, 肖思琪, 等. 深度学习实体关系抽取研究综述. 软件学报, 2019, 30(6): 1793–1818. [doi: [10.13328/j.cnki.jos.005817](https://doi.org/10.13328/j.cnki.jos.005817)]
- 55 Wu F, Weld DS, *et al.* Open information extraction using Wikipedia. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala: Association for Computational Linguistics, 2010. 118–127.
- 56 Nakashole N, Weikum G, Suchanek F, *et al.* PATTY: A taxonomy of relational patterns with semantic types. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island: Association for Computational Linguistics, 2012. 1135–1145.
- 57 Mausam, Schmitz M, Bart R, *et al.* Open language learning for information extraction. Proceedings of the 2012 Joint

- Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island: Association for Computational Linguistics, 2012. 523–534.
- 58 Akbik A, Löser A. KrakeN: N-ary facts in open information extraction. Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX). Montréal: Association for Computational Linguistics, 2012. 52–56.
- 59 Chen YB, Xu LH, Liu K, *et al.* Event extraction via dynamic multi-pooling convolutional neural networks. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing: Association for Computational Linguistics, 2015. 167–176.
- 60 Nguyen TH, Cho K, Grishman R. Joint event extraction via recurrent neural networks. NAACL HLT 2016, the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016. 300–309.
- 61 Chen YB, Yang H, Liu K, *et al.* Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 1267–1276.
- 62 Lee H, Recasens M, Chang A, *et al.* Joint entity and event coreference resolution across documents. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island: Association for Computational Linguistics, 2012. 489–500.
- 63 Barhom S, Shwartz V, Eirew A, *et al.* Revisiting joint modeling of cross-document entity and event coreference Resolution. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 4179–4189.
- 64 Xi XY, Wei Y, Zhang SK, *et al.* Capturing event argument interaction via a bi-directional entity-level recurrent decoder. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021. 210–219.
- 65 Han RJ, Ning Q, Peng NY. Joint event and temporal relation extraction with shared representations and structured prediction. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 434–444.
- 66 Han RJ, Zhou YC, Peng NY. Domain knowledge empowered structured neural net for end-to-end event temporal relation extraction. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 5717–5729.
- 67 Tang JL, Lin HY, Liao M, *et al.* From discourse to narrative: Knowledge projection for event relation extraction. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021. 732–742.

(校对责编: 牛欣悦)