

知识图谱构建技术综述

张吉祥, 张祥森, 武长旭, 赵增顺

(山东科技大学 电子信息工程学院, 山东 青岛 266590)

摘要: 知识图谱在医疗、金融、农业等领域得到快速发展与广泛应用,其可以高效整合海量数据的有效信息,为实现语义智能化搜索以及知识互联打下基础。随着深度学习的发展,传统基于规则和模板的知识图谱构建技术已经逐渐被深度学习所替代。梳理知识抽取、知识融合、知识推理3类知识图谱构建技术的发展历程,重点分析基于卷积神经网络、循环神经网络等深度学习的知识图谱构建方法,并归纳现有方法的优劣性与发展思路。此外,深度学习虽然在自然语言处理、计算机视觉等领域取得了较大成果,但自身存在依赖大规模样本、缺乏推理性与可解释性等缺陷,限制了其进一步发展。为此,对知识图谱应用于深度学习以改善深度学习自身缺陷的相关方法进行整理,分析深度学习的可解释性、指导性以及因果推理性,归纳知识图谱的优势以及发展的必要性。在此基础上,对知识图谱构建技术以及知识图谱应用于深度学习所面临的困难和挑战进行梳理和分析,并对该领域的发展前景加以展望。

关键词: 知识图谱;信息抽取;语义网;深度学习;自然语言处理

开放科学(资源服务)标志码(OSID):



中文引用格式:张吉祥,张祥森,武长旭,等.知识图谱构建技术综述[J].计算机工程,2022,48(3):23-37.

英文引用格式:ZHANG J X, ZHANG X S, WU C X, et al. Survey of knowledge graph construction techniques [J]. Computer Engineering, 2022, 48(3): 23-37.

Survey of Knowledge Graph Construction Techniques

ZHANG Jixiang, ZHANG Xiangsen, WU Changxu, ZHAO Zengshun

(College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)

[Abstract] Knowledge graph has been rapidly developed and widely used in the medical, financial, agricultural, and other fields. It can efficiently integrate the effective information of massive data and lay the foundation for semantic intelligent search and knowledge interconnection. With the development of deep learning, the traditional knowledge graph construction technology based on rules and templates has been gradually replaced by deep learning. This paper studies the development process of three types of knowledge graph construction technologies: knowledge extraction, knowledge fusion, and knowledge reasoning; focuses on knowledge graph construction methods based on deep learning such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN); and summarizes the advantages and disadvantages of existing methods and development ideas. In addition, although deep learning has made great achievements in Natural Language Processing (NLP), computer vision, and other fields, its own defects such as reliance on large-scale samples, lack of reasoning, and interpretability limit its further development. Therefore, this paper sorts out the relevant methods for applying knowledge graph to deep learning to address the defects of the latter; analyzes the interpretability, guidance, and causal reasoning of deep learning; and summarizes the advantages of knowledge graph and the necessity of development. On this basis, this paper studies and analyzes the construction technology of knowledge graph and the difficulties and challenges faced by the application of knowledge graph in deep learning and looks forward to the development prospect of this field.

[Key words] knowledge graph; information extraction; semantic Web; deep learning; Natural Language Processing (NLP)

DOI: 10.19678/j.issn.1000-3428.0061803

0 概述

随着互联网的不断发展,人类从简单的具备内

容获取与查询功能的 Web 1.0,步入到可参与互联网并进行内容制造的 Web 2.0,以及以知识互联为支柱的 Web 3.0^[1],万物互联的时代使人们的生活更便

基金项目: 中国博士后科学基金特别项目(2015T80717);山东省自然科学基金(ZR2020MF086)。

作者简介: 张吉祥(1997—),男,硕士研究生,主研方向为知识图谱、自然语言处理;张祥森、武长旭,硕士研究生;赵增顺(通信作者),副教授、博士。

收稿日期: 2021-05-31 **修回日期:** 2021-08-10 **E-mail:** zhaozengshun@163.com

利,信息获取更快速。但是,由于互联网的内容多源、数据多样,大量的信息不能得到有效地利用,知识互联面临着极大挑战。知识组织的原则中表明知识的充分性、有序性和标准化原则^[2],这就需要人们以一种新的视角去整合互联网的异源和异构知识信息,从而适应用户的认知需求,而知识图谱的诞生为实现语义智能化检索以及知识互联打下了坚实的基础。

互联网信息量的快速增长给深度学习带来了巨大的数据资源,通过对大规模数据的标注和使用,深度学习在自然语言处理(Natural Language Processing, NLP)、计算机视觉(Computer Vision, CV)等领域取得了较好的成绩。但是,深度学习自身存在着很大的局限性,数据红利的消耗殆尽也限制了其进一步发展,具体体现在3个方面:深度学习的效果在很大程度上依赖大规模的样本,缺乏先验知识,导致某些结果可能背离人类知识或专家知识;深度学习本质上是一种映射,是输入和输出之间的特征关系,不具备因果推理性;深度学习缺乏可解释性,只是一种端到端模型,包含了众多的神经元和参数,人们无法清楚地解释每一个参数的意义,这也是深度学习最大的缺陷之一。基于以上原因,人们开始尝试将知识图谱与深度学习相结合,旨在打破人工智能发展中所出现的瓶颈。

徐增林等^[3]对知识图谱的定义、当前的大规模知识图谱、知识图谱构建技术以及知识图谱典型应用进行了分析与讨论,将知识图谱的构建分为知识抽取、知识表示、知识融合、知识推理4个方面。李涓子等^[4]总结归纳知识表示及构建技术。文献[3-4]涉及深度学习技术的部分内容,均从深度学习相关技术应用于知识图谱这一角度出发,而本文从深度学习用于构建知识图谱、知识图谱用于深度学习推理、知识图谱指导深度学习、知识图谱提高深度学习可解释性等多个角度,对最近几年知识图谱与深度学习相结合的最新研究进展进行整理分析。除此之外,还有一些综述是针对知识图谱构建的子任务,如命名实体识别(Named Entity Recognition, NER)^[5]、关系抽取(Relation Extraction, RE)^[6-8]等。其中:文献[5]对2014年—2019年的命名实体识别、命名实体消歧和命名实体链接技术进展进行了详细的分析,以文本预处理、命名实体识别、命名实体消歧、命名实体链接为主要脉络进行综述;文献[6]主要针对关系抽取中的远程监督方法进行归纳,该方法适用于结构化或半结构化的数据;文献[7-8]针对深度学习用于关系抽取进行分析综述,文献[7]分析了卷积神经网络在关系抽取中的应用,文献[8]分析了卷积神经网络、循环神经网络以及混合网络的

应用。

本文阐述知识图谱构建技术的发展历程,对相关模型进行讨论,归纳知识抽取、知识融合、知识推理等相关研究成果,分析知识图谱用于深度学习时的可解释性、指导性以及因果推理性,在此基础上,对知识图谱未来的发展方向加以展望。

1 知识图谱的定义与架构

1.1 知识图谱的定义

知识图谱的概念由 Google 于 2012 年提出,用于完善搜索引擎,是一种典型的多边关系图,由节点(实体)和边(实体之间的关系)组成。知识图谱本质上是一种语义网络,用于揭示万物之间的关系。如图 1 所示,知识图谱旨在从多种类型的复杂数据中抽取概念、实体和关系,是事物关系的可计算模型。按照知识的覆盖范围和领域的不同,知识图谱整体可以划分为通用性知识图谱和领域性知识图谱。随着科技的不断发展,知识图谱在 NLP 领域应用广泛,如语义搜索^[9]、智能问答^[10]、辅助决策^[11]等方面,其已经成为人工智能发展的重要动力。

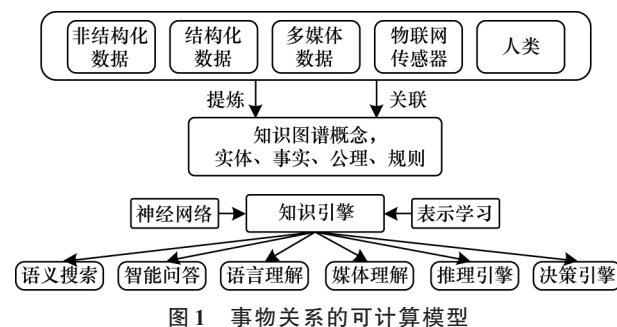


图1 事物关系的可计算模型

Fig.1 Computable model of relationship between things

知识图谱的一种通用表示形式是三元组形式,即 $G=(Entity_{head}, Relation, Entity_{tail})$, $Entity_{head}$ 为三元组 G 中的头实体, $Entity_{tail}$ 为尾实体, $Relation$ 为 2 个实体之间的关系,其中, $Entity=[Entity_1, Entity_2, \dots, Entity_n]$ 表示实体的集合,其包含了 n 种实体的概念, $Relation=[Relation_1, Relation_2, \dots, Relation_n]$ 表述实体之间的关系集合,其包含了 n 种不同的关系。

1.2 知识图谱的体系架构

知识图谱的体系架构分为 3 个部分,如图 2 所示:第一部分是源数据的获取,即在各个类型的数据中获取有用的资源信息;第二部分是知识融合,用于关联多数数据源的知识,扩大知识范围;第三部分是知识计算与知识应用,知识计算是知识图谱能力输出的主要方式^[12],而知识应用是将知识图谱与特定领域或业务相结合,从而提高业务效率。由于构建知识图谱的技术和深度学习紧密相关,因此本文重点分析知识图谱构建技术。

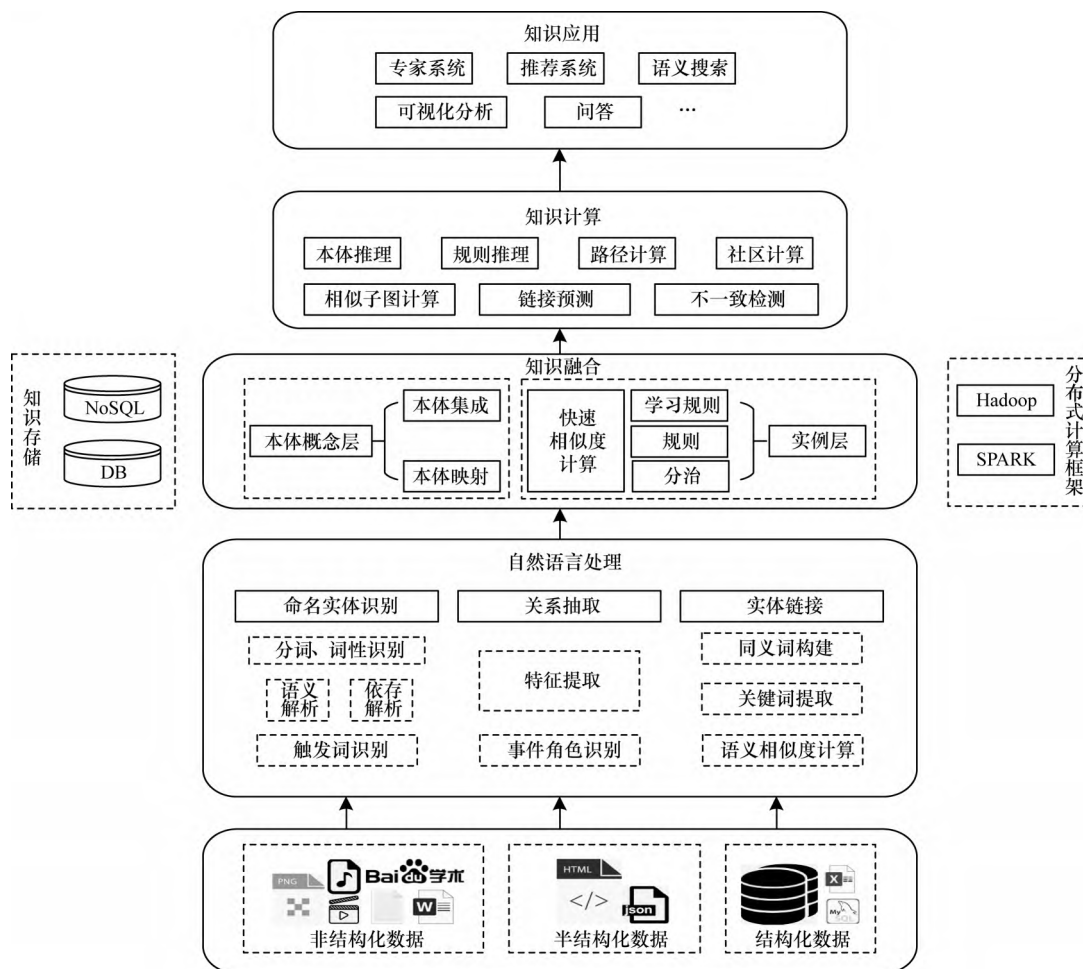


图2 知识图谱的体系架构

Fig.2 Architecture of knowledge graph

2 知识图谱构建技术

知识图谱的构建需要应用到多方面信息处理技术。知识抽取从多种数据源中提取知识并存入知识图谱,是构建大规模知识图谱的基础。知识融合可以解决不同知识图谱的异构问题,通过知识融合,能够使得不同数据源的异构知识图谱相互联通、相互操作,从而提高知识图谱的质量。知识计算是知识图谱的主要输出能力,其中,知识推理是最重要的能力之一,

其为知识精细化工作以及辅助决策的实现方式。本节根据知识图谱的体系架构详细介绍知识抽取、知识融合、知识推理的相关研究。

2.1 知识抽取

知识抽取主要分为命名实体识别和关系抽取2个方面。按照其发展历程,主要可分为3类方法,分别是基于传统规则和模板、基于统计机器学习、基于深度学习的知识抽取,本文主要对第3类方法展开分析。实体关系抽取的发展历程如图3所示。

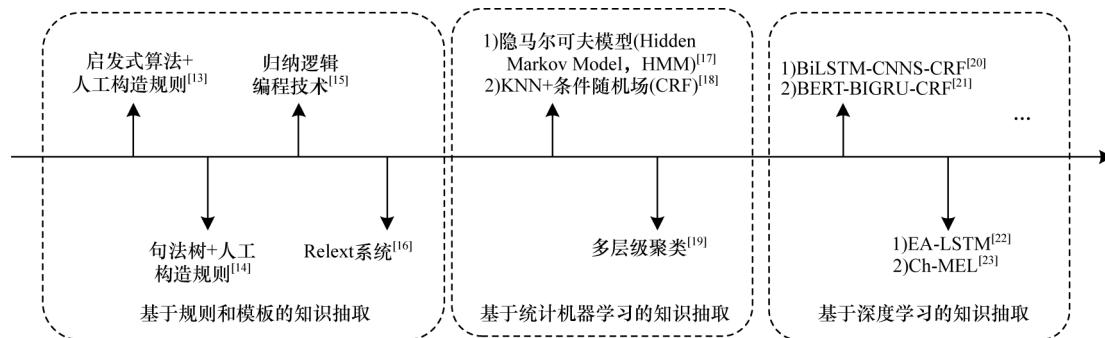


图3 实体关系抽取的发展历程

Fig.3 The development of entity relationship extraction

2.1.1 命名实体识别

在命名实体识别方面,3类知识抽取方法具体如下:

1) 基于传统规则和模板的方法

在早期技术发展不成熟时,命名实体识别和关系抽取主要采用人工编写规则和模板的方法来实现^[13-14]。对于命名实体识别任务,首先由特定领域的专家构建大量的实体识别规则,如“人名:(姓氏+名字),地址名:(省+市+县+...)”,然后将规则和文本字符相匹配从而抽取实体。1991年,RAU^[13]首次利用启发式算法和人工构造规则,从财经新闻中自动提取公司名称,准确率超过了95%,远高于人工抽取的准确率。但是,通过构造规则的方法会耗费大量的人力物力,并且一套规则只能适用于一种领域,其迁移性和泛化性很低。

2) 基于传统机器学习的方法

基于机器学习的方法主要利用标注的数据进行模型训练,采用的模型有最大熵马尔科夫模型(Maximum Entropy Markov Model, MEMM)、条件随机场(Conditional Random Field, CRF)等。2004年,ZHOU等^[17]通过HMM和基于HMM的实体识别器,对词的形态模式、词性等进行集成,通过KNN算法解决了数据稀疏问题,在GENIA V3.0中总体F1值为66.6%。除此之外,2011年,LIU等^[18]在半监督学习框架下结合KNN分类器和CRF模型进行实体识别,该方法缓解了训练数据匮乏的问题,并且将KNN与半监督学习策略相结合,提升了模型的效果。基于机器学习的方法需要构造特征,其误差传播也是一个问题,因此,研究人员开始将命名实体识别技术转向深度学习领域。

3) 基于深度学习的方法

基于深度学习的方法对人工构造特征的依赖性大幅降低,解决了特征提取误差传播的问题,对于命名实体识别而言,主要的方法有卷积神经网络(Convolutional Neural Network, CNN)和循环神经网络(Recurrent Neural Network, RNN)两大类。2011年,COLLOBERT等^[24]将单层的CNN模型用于命名实体识别,但是其丢失了长距离单词的有效信息。针对传统CNN模型学习速率低的问题,2019年,QIU等^[25]提出结合CRF的残差膨胀卷积神经网络(RD-CNN-CRF),其利用RD-CNN捕获上下文特征,最后通过CRF捕获相邻标签的相关性,在CCKS-2017

数据集中,RD-CNN-CRF的性能高于BI-LSTM-CRF,并且学习速率更快。随着注意力机制的兴起和广泛运用,2021年,JUN等^[26]针对传统CNN无法捕捉句子中长期信息的问题,将注意力机制和CNN相融合,提出新的卷积神经网络模型ALL CNN(ACNN),其利用融合不同卷积核以及残差结构的CNN来捕捉不同尺度的上下文信息,引入注意力机制增强模型捕捉上下文信息的能力。

除CNN模型外,RNN模型及其变体同样在NER任务中被广泛应用。HUANG等^[27]提出了LSTM、BI-LSTM、BI-LSTM-CRF等模型,BI-LSTM可以综合考虑过去和未来的特征,因此,在CoNLL-2000和CoNLL-2003数据集中,BI-LSTM-CRF相对其他模型准确率更高。受到上述研究的启发,文献[20]提出端到端的命名实体识别模型,其通过CNN网络将单词级和字符级的嵌入向量整合在一起并同时送入BI-LSTM-CRF模型中,取得了较好的效果。2018年,GREGORIC等^[28]采用多个独立的BI-LSTM单元,通过模型间的正则化提高了各个LSTM单元之间的多样性,大幅减少了参数量,在CoNLL-2003数据集中比文献[20]方法的F1值高0.27%。2020年,RONRAN等^[29]改进HUANG等所提的模型,研究词嵌入、字符特征和词特征对实体识别的有效性,并利用CNN-BI-LSTM-CRF模型进行实体识别,其准确率得到有效提升。

除CNN与RNN之外,近年来,将Transformer系列模型应用于命名实体识别也成为研究重点。Transformer完全依赖于注意力机制,准确率更高并且训练时间更少,典型代表有文献[30]提出的针对命名实体识别的Transformer(Transformer Encoder for NER, TENER)、Google^[31]提出的BERT(Bidirectional Encoder Representations from Transformers)以及BERT-BIGRU-CRF^[21]等衍生模型。2021年,曾青霞等^[32]提出一种结合自注意力机制的BI-LSTM-CRF(SelfAtt-BI-LSTM-CRF),其在BI-LSTM-CRF的基础上引入自注意力机制,可以获取句子的全局依赖性并捕捉其结构特征。罗熹等^[33]同样将自注意力机制与BI-LSTM-CRF相结合,使用一种新的融合领域字典的字符表示方法,有效提升了模型的实体识别能力。注意力机制的引入为命名实体识别等NLP技术注入了新的活力,拓展出了新的研究方向。近年来所出现的命名实体识别方法对比如表1所示。

表 1 命名实体识别方法对比
Table 1 Comparison of named entity recognition methods

方法类别	年份	方法	数据集	模型	方法特点	适用场景	评测指标	评测值/%
基于传统规则的方法	1991 ^[13]	启发式方法	财经新闻	启发式算法+规则	较准确地自动提取实体,但构造规则的方法会耗费大量的人力,可移植性很差	财经新闻	Acc	97.50
基于传统机器学习的方法	2004 ^[17]	HMM	GENIA V3.0	HMM+实体识别器	集成了构词模式、形态模式、词性、中心名词、特殊动词、别称这 6 个特征,特征丰富,但需要人工构造特征	生物医学领域	F1	66.60
基于传统机器学习的方法	2011 ^[18]	CRF	Tweets	KNN 分类器+CRF	在半监督框架下进行实体识别,有效缓解了训练数据匮乏的问题,但需要人工构造特征	推特博文等通用领域	F1	80.20
基于深度学习的方法	2011 ^[24]	CNN	CoNLL-2003	Conv-CRF	首次引入 CNN 进行实体识别,但丢失了长距离单词的有效信息	通用领域	F1	88.67
基于深度学习的方法	2019 ^[25]	CNN	CCKS-2017	RD-CNN-CRF	将实体识别视为序列标注任务,利用残差膨胀卷积捕获上下文,有效提高了训练效率	医学临床领域	F1	88.51
基于深度学习的方法	2021 ^[26]	CNN	CCKS-2017	ALL CNN	构建多级 CNN+注意力机制捕获不同尺度的上下文信息,提高了模型效率	医学临床领域	F1	90.49
基于深度学习的方法	2015 ^[27]	RNN	CoNLL-2003	BI-LSTM-CRF	首次应用 BI-LSTM 捕获过去和未来的特征,但需要大量的特征工程	通用领域	F1	88.83
基于深度学习的方法	2016 ^[20]	RNN	CoNLL-2003	LSTM-CNNs-CRF	不需要人工构造特征,将 BI-LSTM 与 CNN 相结合,是完全端到端的模型	通用领域	F1	91.21
基于深度学习的方法	2018 ^[28]	RNN	CoNLL-2003	并行 RNN 模型	采用多个独立的 BI-LSTM,大幅减少了参数量,提高了训练效率	通用领域	F1	91.48
基于深度学习的方法	2020 ^[29]	RNN	CoNLL-2003	CNN-BI-LSTM-CRF	研究单词和字符特征对实体识别的有效性,采用两层 BI-LSTM 减少输入序列,以克服长输入序列难以预测的问题	通用领域	F1	91.10
基于深度学习的方法	2019 ^[30]	Trans-former	CoNLL-2003	TENER	引入相对位置编码,可以分别在词级与字符级表示	通用领域	F1	91.52
基于深度学习的方法	2019 ^[31]	Trans-former	CoNLL-2003	BERT	采用 Transformer-Encoder 结构,可以深度挖掘上下文相关信息,但模型参数量大,训练速率较慢	通用领域	F1	92.80
基于深度学习的方法	2021 ^[32]	Trans-former	CoNLL-2003	SelfAtt-BI-LSTM-CRF	引入自注意力机制,更好地处理实体之间的长距离依赖关系	通用领域	Acc	90.47
基于深度学习的方法	2021 ^[33]	Trans-former	CCKS-2017	MHA-BiLSTM-CRF	将中文字符特征与临床知识特征相结合,对医学临床文本更具针对性	医学临床领域	F1	91.97

2. 1. 2 关系抽取

在关系抽取方面,3 类知识抽取方法具体如下:
1) 基于传统规则和模板的方法
传统的关系抽取主要采用人工编写规则和模板的方法来实现^[24-25],一般是由特定领域的专家手动

编写模板来匹配关系,这种基于规则和模板的方法会耗费领域专家大量的时间和精力,且可移植性较差,无法适应数据的变化。
2) 基于传统机器学习的方法
对于关系抽取而言,传统的机器学习方法可以

分为有监督、半监督和无监督3类。有监督的关系抽取算法受到标注数据集的制约,其准确率受到标注数据质量和数量的影响,并且不能拓展新关系^[34]。因此,学术界开始转向研究半监督和无监督的学习方法,这2种方法对标注数据的依赖性较弱,适合缺少语料数据的关系抽取任务。无监督学习具有领域无关性,非常适合大规模开放领域的关系抽取,如文献[19]通过多层级聚类方法抽取关系,在基础模板上不断衍射生成新的子类模板,在美国的12家报纸文章中进行关系抽取,获得了较好的效果。基于传统机器学习的方法缺点同样明显,该类方法存在特征提取误差传播问题,因此,研究人员开始将深度学习和实体关系抽取相结合。

3) 基于深度学习的方法

基于深度学习的关系抽取方法改善了特征提取误差传播的问题,是近些年的研究热点。本文将基于深度学习的关系抽取方法分为流水线方法和实体关系联合抽取方法两类。

(1) 流水线方法

流水线方法分2步抽取信息,先抽取实体再抽取关系,最后整合三元组输出。流水线方法一般采用CNN、RNN及其改进模型进行关系抽取,由于它们不需要手动构造特征并且预测精度很高,因此得到广泛关注和应用。

文献[35]首次利用CNN提取词汇特征和句子特征并进行关系抽取,大幅提升了模型的分类准确率,但其提出的模型在长句序列中会受到不相关子序列的影响和干扰。2015年,XU等^[36]利用CNN从最短依赖路径中学习关系表示,并在模型中加入了负采样策略,其比文献[35]模型的F1值提高了1.3%,大幅提升了关系抽取性能。CNN模型存在2个问题:一方面,CNN模型卷积核的大小固定不变,因此,其对句子语义的提取不充分,为解决该问题,GUO等^[37]提出一种新的Att-RCNN模型以提取文本特征与关系分类,利用注意力机制来强化关键词与特征,提高了分类性能,2020年,闫雄等^[38]将CNN与自注意力机制相结合进行关系抽取,并取得了较好的效果;另一方面,单一CNN对非连续性词的特征抽取效果较差,因此,龚乐君等^[39]将CNN与BI-GRU相结合,改善了单一模型特征提取不足的问题,文献[40]将BI-LSTM与CNN相结合,并引入注意力机制来捕捉句子的内部语义信息,在SemEval-2010数据集中,该融合模型的F1值为79.38%,优于单一模型。

除了采用CNN进行关系抽取外,RNN同样被用于关系抽取。2012年,SOCHER等^[41]利用RNN获取句子的向量表示,提高了关系抽取的性能。由于RNN存在梯度消失和梯度爆炸的问题,因此研究人员开始将LSTM等RNN网络的变体引入关系抽取任务。ZHANG等^[42]采用Bi-LSTM进行关系分类,

Bi-LSTM可以获取当前单词之前和之后的信息,在关系抽取任务中具有有效性。2019年,GAN等^[22]提出子序列级实体注意力LSTM网络(EA-LSTM),其可以集中地关注2个实体间的重要信息,在特征融合阶段引入注意力机制来提高模型的上下文信息处理能力。文献[43]对句子本身进行句法分析,构建双向树结构的BI-LSTM分析句法,提高了关系抽取性能。针对关系抽取中过度依赖句子树本身信息的问题,文献[44]将注意力机制与句子树结构相结合,生成权重矩阵,构造注意力图长短时记忆神经网络(Attention Graph Long Short Term Memory Neural Network,AGLSTM),同时利用句子的时序结构和自身结构进行关系抽取,在SemEval-2010数据集中,其F1值为85.3%。

(2) 实体关系联合抽取方法

传统的流水线方法先抽取实体再抽取关系,其存在错误传播的问题,实体识别模块的效果直接影响关系抽取模块的抽取效果。另外,实体识别模块中抽取的实体对不一定完全存在关系,没有关系的实体对会带来冗余信息,影响分类效果。联合抽取方法将实体识别和关系抽取模型相融合,直接在文本中抽取实体关系三元组。联合抽取方法可分为2种,分别是基于参数共享和基于序列标注的联合抽取方法。

① 基于参数共享的实体关系联合抽取

2016年,MIWA等^[45]首次提取单词序列和依赖树结构上实体之间的关系以用于模型训练,在模型结构中,实体识别和关系抽取共享BI-LSTM编码层信息,但是,这种方法只是简单地共享底层参数和编码表示,2个任务之间没有关联,并非真正意义上的联合抽取。2017年,KATYAR等^[46]首次将注意力机制和BI-LSTM用于实体关系联合抽取,实体识别任务采用BI-LSTM来完成,关系分类任务利用实体识别任务输出的实体序列和共享的编码表示作为输入,引入了注意力模型,有效改善了文献[45]方法依赖于词性标签等复杂特征的缺点。

② 基于序列标注的实体关系联合抽取

基于参数共享的实体关系联合抽取任务弥补了传统流水线模型中的错误传播、信息丢失等缺点,但是,基于参数共享的模型本质上还是先抽取实体再抽取关系,会产生不存在关系的冗余实体信息。为了弥补这一缺陷,ZHANG等^[47]提出了基于序列标注的实体关系联合抽取方法,其可以将实体关系联合抽取任务转换为序列标注任务,如图4所示,序列标注输出包含3种信息:实体词的位置信息,采用{B I E S O}进行标注;关系类型信息,如{CP: Country-President, CF: Company-Founder};实体对角色信息{1, 2},分别表示{实体1, 实体2}。通过这种新的标注策略可以同时识别标注出实体和关系,避免构建复杂特征,有效减少了基于参数共享方法中的实体识别冗余信息。

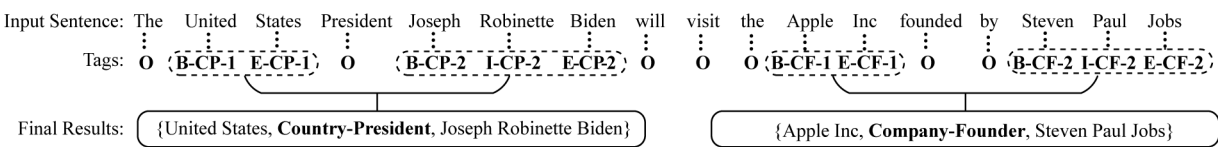


图 4 实体关系联合抽取标注策略

Fig.4 Annotation strategy of entity relationship joint extraction

但是,文献[47]方法同样存在缺陷,即无法解决实体关系重叠问题(EntityPairOverlap, SingleEntity Overlap, EPO SEO)。针对关系重叠问题,文献[48]提出一种基于多标签分类的联合分类模型,其编码器由BI-LSTM+DCNN组成,解码器由实体预测、关系预测以及实体关系联合预测组成,将关系抽取任务转换为多标签分类任务,有效解决了实体关系重叠问题。2019年,WEI等^[49]提出了级联二进制标记框架(CasRel),其将关系看作主语映射到宾语的函数 $f(\cdot)$,通过 Subject Tagger 与 Relation-Specific Object Taggers 获取实体关系三元

组,有效解决了关系重叠问题。文献[50]同样将三元组抽取分为头实体抽取和尾实体抽取2个部分,解决了SEO问题,在进行尾实体抽取时,采用3种策略(EPO_None/Two/Three_Tagging)对尾实体进行标记,在NYT数据集中取得了较好的效果。2020年,XIAO等^[23]设计一种章节级联合学习的模型Ch-MEL,该模型利用BI-LSTM与自注意力机制增强章节学习表示能力,将参数共享与联合解码相结合,在SemEval-2010数据集中,其F1值比文献[45]模型高0.4%。近年来所出现的关系抽取方法对比如表2所示。

表 2 关系抽取方法对比

Table 2 Comparison of relationship extraction methods

方法类别	方法细分	年份	方法	数据集	模型	方法特点	适用场景	评测指标	评测值/%
基于传统规则和模板的方法	—	2005 ^[16]	本体扩展	football domain	RelExt	利用本体层次结构与概念之间的关系进行关系抽取,但可移植性较差	足球领域	Precision	20.20
基于传统机器学习的方法	—	2006 ^[19]	多层次聚类	news	Unrestricted Relation Discovery	无监督关系抽取,在基础模板上不断衍生成新的子类模板	新闻报纸领域	Acc	75.00
基于深度学习的方法	流水线方法	2014 ^[35]	CNN	SemEval-2010	CDNN	采用CDNN自动提取词汇与句子级别的特征,但容易受到不相关子序列的干扰	通用领域	F1	82.70
基于深度学习的方法	流水线方法	2015 ^[36]	CNN	SemEval-2010	depLCNN+NS	引入最短依赖路径与负采样策略来避免无关序列的干扰,并解决了关系的方向性问题	通用领域	F1	84.00
基于深度学习的方法	流水线方法	2019 ^[37]	CNN	SemEval-2010	Att-RCNN	结合CNN与BI-GRU提取更丰富的上下文表示,利用词级注意力与句子级注意力强化特征	通用领域	F1	86.60
基于深度学习的方法	流水线方法	2020 ^[38]	CNN	SemEval-2010	Self-Att-CNN	通过引入自注意力机制得到词序列之间的关系,使得词向量信息更加丰富	通用领域	F1	84.46
基于深度学习的方法	流水线方法	2020 ^[39]	CNN	DDIExtraction 2013	BI-GRU+CNN	BI-GRU与CNN相对于单一模型可以更充分地提取句子蕴含的语义信息与特征	生物医学领域	F1	75.00
基于深度学习的方法	流水线方法	2012 ^[41]	RNN	SemEval-2010	MV-RNN	利用RNN获取句子的向量表示,但无法处理长期依赖问题	通用领域	F1	82.40
基于深度学习的方法	流水线方法	2015 ^[42]	RNN	SemEval-2010	BI-LSTM	采用BI-LSTM获取单词的双向信息	通用领域	F1	84.30
基于深度学习的方法	流水线方法	2019 ^[22]	RNN	SemEval-2010	EA-LSTM	重点关注实体之间的文本,在特征融合阶段引入注意力机制获取最终的特征表示	通用领域	F1	84.70

续表

方法类别	方法细分	年份	方法	数据集	模型	方法特点	适用场景	评测指标	评测值/%
基于深度学习的方法	流水线方法	2021 ^[44]	RNN	SemEval-2010	AGLSTM	通过软修剪策略与注意力机制学习句子的结构信息	通用领域	F1	85.30
基于深度学习的方法	联合抽取方法	2016 ^[45]	参数共享	SemEval-2010	SPTree	实体识别与关系抽取共享BI-LSTM编码层,但2个任务之间不存在关联	通用领域	F1	85.50
基于深度学习的方法	联合抽取方法	2017 ^[47]	序列标注	NYT	LSTM-LSTM-Bias	提出一种新的标注策略同时抽取实体与关系,但无法解决实体关系重叠问题	通用领域	F1	49.50
基于深度学习的方法	联合抽取方法	2019 ^[48]	序列标注	NYT	BI-LSTM+DCNN	将关系抽取任务转化为多标签分类任务,有效解决了实体关系重叠问题	通用领域	F1	53.10
基于深度学习的方法	联合抽取方法	2020 ^[49]	序列标注	NYT	CasRel	将关系看作头实体到尾实体的映射,有效解决了实体关系重叠问题	通用领域	F1	89.60
基于深度学习的方法	联合抽取方法	2020 ^[50]	序列标注	NYT	Bert+BI-LSTM	在进行尾实体抽取时采用3种策略,解决了关系重叠问题	通用领域	F1	86.00
基于深度学习的方法	联合抽取方法	2020 ^[23]	序列标注	SemEval-2010	Ch-MEL	从章节级粗粒度角度进行跨句或跨段的实体关系抽取	通用领域	F1	85.90

2.2 知识融合

知识图谱的构建数据来源十分广泛,不同数据源之间的知识缺乏深入的关联,知识重复问题很严重^[51]。知识融合将来自不同数据源的异构化、多样化的知识在同一个框架下进行消歧、加工、整合等,达到数据、信息等多个角度的融合。知识融合的核心在于映射的生成^[12],目前,知识融合技术可以分为本体融合和数据融合2个方面。

2.2.1 本体融合

在知识融合技术中,本体层占据着重要部分。到目前为止,研究人员已经提出了多种解决本体异构的方法,主要分为本体集成和本体映射^[52]两大类,如图5所示。本体集成是将多个不同数据源的异构本体集成为一个统一的本体,本体映射则是在多个本体之间建立映射规则,使信息在不同本体之间进行传递。

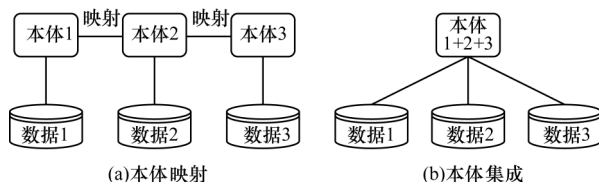


图5 本体映射和本体集成

Fig.5 Ontology mapping and ontology integration

对于本体集成,文献^[53]将本体集成划分为一系列的子活动,包括决定本体集成的方式、确定识别模块、确定每个模块中需要表示的知识、识别和获取候选本体、候选本体研究和分析、选择源本体,最后执行集成过程。由于本体集成步骤比较复杂,耗费大量的人力

物力,缺乏自动方法的支持,因此本体映射方法应运而生,研究人员从不同的角度采用不同的方法建立本体之间的映射,主要分为以下4种映射方法:

1) 基于NLP的方法,其比较映射对象的相似度,如PORTER^[54]提出的Stemming算法,用于寻找词形的变化。

2) 基于结构的方法,如Stanford大学开发的AnchorPROMPT本体工具集^[55]。

3) 基于实例的方法,大多采用机器学习方法,让算法学习到正确的映射。该类方法可以较好地解决异构本体之间的映射问题,但不适用于本体关系间的映射。

4) 综合方法,即综合使用多种映射方法,如QOM,其通过合理利用各种分配算法,可以提高映射效率,有效处理大规模本体映射问题。

2.2.2 数据融合

数据方面的知识融合包括实体合并、实体对齐、实体属性融合等方面。其中,实体对齐是多源知识融合的重要部分,用于消除实体指向不一致性与冲突问题。知识图谱的对齐算法可分为3类,分别是成对实体对齐、局部实体对齐和全局实体对齐。

1) 成对实体对齐

成对实体对齐方法包括基于传统概率模型和基于机器学习的实体对齐方法。文献^[56]利用属性相似度将实体匹配问题转换为分类问题,为后来的研究奠定了基础。2017年,张伟莉等^[57]将实体对齐建模为带约束的二分类问题,将特征空间划分为2个视图,引入半监督协同算法分别在2个视图上训练分类器,提升了实体对齐的效果。

2) 局部实体对齐

局部实体对齐方法引入实体属性并为其分配不同的权重,再进行加权求和计算实体的相似度。文献[58]提出一种大规模实体匹配方法,其主要采用TF-IDF为向量中的每个分量分配权重并建立索引,实体相似度通过余弦相似性进行判别。这种算法具有较快的计算速度和较高的召回率,但准确性较低,因为其只是将实体关系看作实体的一类属性,并未实现真正的对齐。

3) 全局实体对齐

全局实体对齐综合考虑多种匹配策略来判别实体相似度,包括基于相似性传播和概率模型的实体对齐方法。2007年,BHATTACHARYA等^[59]提出一种关系聚类算法用于实体对齐,LACOSTE-JOURNAL等^[60]基于该算法进一步提出SiGMA算法,将实体对齐问题转换为全局匹配评分目标优化问题,引入实体关系与属性并不断迭代发现所有的实体匹配对。基于概率模型的实体对齐方法通常利用统计关系进行计算,如CRF、马尔科夫逻辑网(Markov Logic Networks, MLNs)模型等。2012年,WICK等^[61]将CRF引入判别式层次模型中,其中将实体作为树结构、观察值作为叶子节点,以进行推理,其适用于大规模的知识图谱实体对齐。针对大规模数据集对齐效率问题,文献[62]基于MLNs提出一种通用的实体匹配算法,其在实体的有限子集运行匹配器,在实例之间传递所构建的消息集以控制匹配器的交互,从而显著提高实体匹配的召回率。

2.3 知识推理

知识推理根据已有的实体关系信息来推断新的事实结论,从而进一步丰富知识图谱,满足上游任务的需求。本文将知识推理方法主要分为3种类型,分别为基于逻辑规则的推理、基于分布式特征表示的推理和基于深度学习的推理。

2.3.1 基于逻辑规则的推理

基于逻辑规则的推理包括谓词逻辑推理、本体推理和随机推理。2010年,SCHOENMACKERS等^[63]提出的一阶归纳学习就是一种典型的谓词逻辑推理,其可以自动提取高质量的事实并去除噪声,但是,谓词逻辑推理效率很低。2016年,CHEN等^[64]提出一种本体寻路(Ontological Pathfinding, OP)算法,该算法通过一系列的并行优化技术实现大规模的知识图谱,在Freebase数据库中挖掘了第一个规则集,取得了较好的效果。文献[65]为解决实体关系推理中准确率和召回率低的问题,提出双层随机游走(Two-tier Random Walk, TRWA)算法,TRWA利用无向图来表述知识图谱,除此之外,TRWA在全局模式和局部模式2种角度下对路径特征进行评估,提高了算法的准确率和召回率。

2.3.2 基于分布式特征表示的推理

基于分布式特征表示的推理包括基于翻译模型的知识推理、基于张量分解的知识推理以及基于语

义匹配模型的知识推理。

1) 基于翻译模型的知识推理

在基于翻译模型的知识推理中,经典模型有TransE、TransR等,但是,TransE、TransR等模型认为三元组之间是独立的。2018年,WANG等^[66]提出TransN模型以整合三元组周围的邻域信息,其采用对象嵌入和上下文嵌入表示实体与关系,提升了知识推理的性能。由于TransE、TransR等模型都无法完全满足所有关系的建模,因此SUN等^[67]提出Rotate模型,如图6所示,其将关系看作头实体向量向尾实体向量的旋转角度,可以建模和推断各种关系的模式,并提出一种较为新颖的自对抗负采样技术,使得模型效果大幅提升。

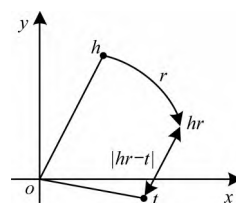


图6 Rotate模型中的关系建模

Fig.6 Relationship modeling in Rotate model

文献[68]在TransE的基础上引入三元组邻域信息,提出聚合邻域信息的表示模型TransE-NA,为实体选择最相关的属性作为邻域信息,有效缓解了数据稀疏问题。2021年,宋浩楠等^[69]针对知识推理可解释性差的问题,将知识表示与强化学习相结合,提出RLPTransE模型,其将知识推理问题转化为马尔科夫序列决策问题,增强了知识推理的可解释性。

2) 基于张量分解的知识推理

在基于张量分解的知识推理中,一般将知识图谱中的实体关系三元组通过张量分解方法进行表示学习,将分解得到的向量重构为张量,元素值高于一定阈值的作为候选推理结果。2011年,NICKEL等^[70]提出RESCAL,如图7所示,其将高维关系数据分解为三阶张量,在降低数据维度的同时又保持了数据原有特征,在知识推理中取得了较好的效果。

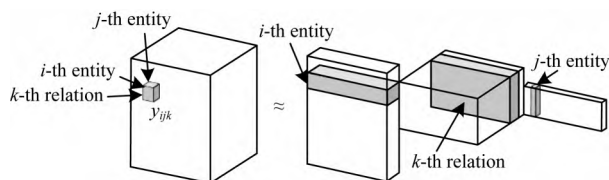


图7 高维关系数据的分解

Fig.7 Decomposition of high dimensional relational data

由于RESCAL计算复杂,所占内存很大,因此文献[71]提出新的知识推理模型TRESICAL,其引入实体类型信息来约束并排除不满足关系的三元组,显著降低了训练时间并提高了预测性能。2017年,吴运兵等^[72]通过对知识图谱进行路径张量分解,旨在解决知识图谱推理中忽略实体间关系路径的问

题,该方法可以高效地挖掘实体间的关系与新事实,有助于完善知识图谱。

3) 基于语义匹配模型的知识推理

文献[73]提出 DistMult 模型,其将 RESCAL 中的每一个关系向量转换为对角矩阵,从而对 RESCAL 进行简化,减少参数,在对现有知识库的推理中表现出了较好的效果。但是,无论是 RESCAL 还是 DistMult,都忽略了实体和关系的语义多样性。2017年,刘峤等^[74]认为每个关系都反映了相应实体的某些语义关系,可以通过选择性的加权来对这些关系进行表示和区分,因此,其提出统一加权模型(Unified Weighted Model, UWM)和独立加权模型(Independent Weighted Model, IWM)关系推理算法,计算效率较高。2019年,ZHANG 等^[75]引入超复数来建模实体和关系,同样将关系看作超平面内头实体到尾实体的旋转,相对于 Rotate 只有一个旋转平面,QuatE 有 2 个旋转平面,其自由度及泛化能力更好。

2.3.3 基于深度学习的推理

目前,神经网络已被广泛应用于 NLP 领域,并取得了显著的成效。神经网络可以自动捕捉特征,通过非线性变换将输入数据从原始空间映射到另一个特征空间并自动学习特征表示,适用于知识推理这种抽象任务。

2013年,SOCHER 等^[76]提出一种新的神经张量网络模型(Neural Tensor Network, NTN)以建模关系信息,如图 8 所示,该模型采用双线性张量层直接将 2 个实体向量跨多个维度联系起来,刻画实体之间复杂的语义关系,大幅提高了推理性能。

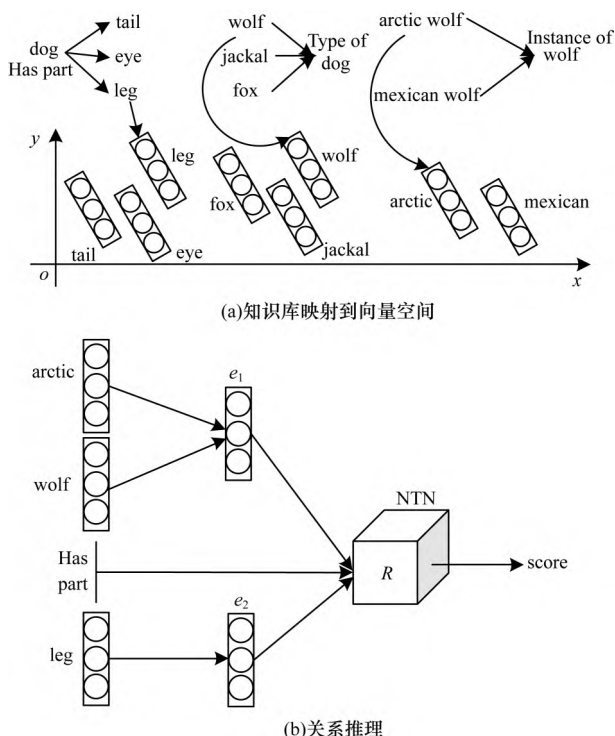


图 8 NTN 网络中的关系建模

Fig.8 Relationship modeling in NTN network

DAS 等^[77]提出一种具有单一性和高容量性的 RNN 模型,该模型的所有目标关系共享 RNN 的关系类型表示和组合矩阵,减小了训练参数量,在大规模知识图谱推理中具有更高的准确性和实用性。2018年, GUO 等^[78]设计了知识图谱的深度序列模型(Deep Sequential model for KG, DSKG),分别用独立的 RNN 单元处理实体层和关系层,取得了较好的效果。2020年,CHEN 等^[79]针对实体句中的词序特征,提出基于 LSTM 的知识图谱嵌入方法(Learning Knowledge Graph Embedding with Entity Descriptions based on LSTM Networks, KGDL),其采用 LSTM 实体描述的句子进行编码,然后联合 TransE 与 LSTM 模型将实体描述的句子嵌入与三元组编码为实体描述,从而实现知识推理。除 RNN 之外, CNN 也被引入知识推理任务。2018年, DETTMERS 等^[80]提出一种用于知识推理的卷积神经网络模型 ConvE,该模型采用二维卷积的嵌入来对知识图谱中的新链接进行推理,在现有的知识图谱数据集中获得了较好的结果。由于 ConvE 的交互数量有限,因此 VASHISHTH 等^[81]提出 InteractE, InteractE 通过特征置换、特征重塑以及圆形卷积来捕捉额外的异构特征,增加实体关系间的交互次数。李少杰等^[82]认为 ConvE 丢失了三元组的整体结构信息,因此,提出基于 CNN 的知识表示模型(Convolutional Knowledge Embeddings, ConvKE),该模型将三元组的各个元素整合到一起提取整体的结构信息,通过维度变换策略增加卷积滑动窗口步数,增强了知识之间的信息交互。

3 知识图谱在深度学习中的应用

3.1 深度学习的可解释性

深度学习在 CV、NLP 等领域获得了巨大的成功,但是可解释性一直是深度学习的一个弱点,深度学习复杂性高,参数多,人们一般无法解释这种“端到端”模型做出决策的原理,也无法判断决策的可靠性。大部分的深度学习模型都采用数据驱动的方法,很少关注知识驱动,因此,将知识图谱与深度学习相结合构建具有可解释性的深度学习模型是一个新的研究课题。

2016年, CATHERINE 等^[83]将深度学习和知识图谱相结合,构建了一个具有可解释性的通用概率逻辑系统(Programming with Personalized PageRank, ProPPR),以完成推荐任务,其提出 EntitySim,由于 EntitySim 仅使用知识图谱链接来了解用户的喜好,并没有利用实体的类型等额外信息,因此文中作者将其扩展为 TypeSim,并开发了基于图的潜在因子模型 GraphLF,其可以展现对象的隐藏维度以提高推荐性能。2019年, CAO 等^[84]考虑到知识图谱总是不完整的,因此,他们在将知识图谱引入推荐系统时还加入了知识图谱补全技术,利用知识图谱的先验

知识来增强用户-系统的交互,通过用户-系统的交互来补全知识图谱,实现了相互增强的效果,提高了推荐的可解释性。文献[85]提出一种图模型,用于揭示预训练模型CNN中的隐藏层知识层次,其通过自动学习上万个节点的解释图来揭示CNN的层次和知识结构,解释图中的节点代表CNN中卷积层的部分模式,利用知识图谱来解释决策。

3.2 深度学习的指导

知识图谱可以作为约束来优化目标,将知识图谱中的知识表达作为优化目标的后验正则项,从而提升深度学习模型的性能。其中,比较典型的是文献[86]将深度神经网络和结构化逻辑规则相结合,如图9所示,将一阶逻辑规则用于增强各种类型的神经网络,该文中所提框架 teacher-student network 可以使神经网络在标记的数据和逻辑规则中同时进行学习,通过一个迭代的规则知识提取过程,将逻辑规则中编码的结构化信息转换为网络参数。其中,利用 teacher network 建模逻辑规则,student network 通过反向传播算法与 teacher network 的约束来学习规则,该框架应用于CNN与RNN,在命名实体识别和情感分析中获得了较高的性能。

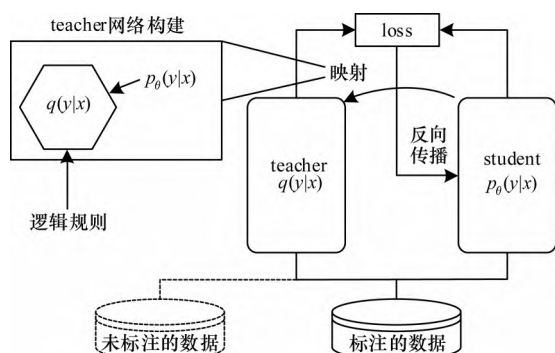


图9 逻辑规则与神经网络结合框架

Fig.9 Combining framework of logic rules and neural networks

2018年,WANG等^[87]提出深度概率逻辑系统(Deep Probabilistic Logic,DPL),DPL将概率逻辑和深度学习相结合,使用概率逻辑表示间接监督的监督模块,使用深度神经网络作为最终的预测模块,DPL模型将决策标记为潜在变量,先验知识采用加权一阶逻辑公式表示,最后采用变分期望最大化交替学习深度神经网络和改善间接监督中不确定权重,在生物医学机器阅读实验中表现出了优良的性能。针对神经网络需要大量数据集进行训练的问题,LI等^[88]利用逻辑规则将先验知识引入神经网络中用于指导训练和预测。不同于文献[86]使用结构化规则的teacher network,该文直接将结构化的知识整合到神经网络结构中,通过将规则转换为可微分的计算图,使用其节点约束来系统地扩充现有的网络结构。该文提出的框架在机器理解、自然语言推理、文字分块等任务中都可以改进现有神经网络的

性能,尤其适用于数据集有限的情况。2019年,WANG等^[89]在信息抽取任务中,考虑到实体之间的关系与实体类型紧密相关,这种依赖关系可以看作约束并表示为逻辑规则,同样采用一阶逻辑的形式将逻辑知识集成到深度学习网络中。该文提出的模型由深度学习模块和逻辑模块组成,深度学习模块用于学习每个单词的特征表示,逻辑模块是通过一阶逻辑(First-Order Logic,FOL)表示的一组逻辑规则,其指明了输出标签的空间复杂关系。文中通过映射函数将神经网络的输出映射到逻辑系统并最终产生逻辑输出,在意见目标抽取任务和关系抽取任务中效果较好。

3.3 深度学习推理

深度学习本质上是一种映射,并不具有推理性,缺乏“归纳推广”的能力。贝叶斯网络之父JUDEA PEARL在文献[90]中也明确提出,当前机器学习的突破口在于“因果革命”。基于此,在因果推理领域,最典型的方法是BATTAGLIA等^[91]面向关系推理而提出的图网络,该网络将图与深度学习相结合,图网络具有较好的归纳推广和推理能力,同时也具备深度学习“端到端”的灵活性,其主要的计算单元为图网络模块(graph-to-graph),输入为图的格式,对结构进行计算并返回图作为输出,图网络相对于普通的神经网络具有归纳推广和组合概括的能力,使得深度学习也能进行因果推理。图神经网络除了具有较好的归纳推广和因果推理能力之外,其最大的优点在于可以处理非欧几里德结构数据。图神经网络可以分为不同的类别,如图卷积网络(Graph Convolution Networks,GCN)、图注意力网络(Graph Attention Network,GAT)等。

在推荐模型方面,文献[92]提出一种混合推荐模型,其包含2个模型,分别是由知识图谱和深度学习相结合的RCKD模型和知识图谱与协同过滤相结合的RCKC模型。其中:RCKD模型首先利用已知的三元组 (h,r,t) 构成图,再搜索知识图谱中的推理路径,利用TransE模型将推理路径嵌入成路径向量,最后利用深度学习获取路径向量中的语义预测评分;RCKC模型则利用TransE模型获取知识图谱中的实体向量,再利用协同过滤思想对用户进行物品推荐,其推荐准确率相对较高。

4 存在的困难与挑战

知识图谱的概念于2012年由Google提出,其发展还不算成熟,属于建立在多个技术领域上的一门实用技术。本文对知识图谱的构建以及知识图谱在深度学习中的应用进行分析,总结出该领域目前面临的困难和挑战具体如下:

1) 知识图谱构建技术方面

(1)对于知识抽取技术,现有的实体关系联合抽取技术并未解决关系重叠问题,算法准确率和召回

率都较低。除此之外,基于开放域的关系抽取还处于初级阶段,多语种、大范围的多元实体关系抽取是当前面临的主要困难之一,无监督关系抽取具有较高的可移植性和泛化性,为开放域关系抽取任务提供了新的思路。

(2)对于知识融合技术,实体对齐是最主要的手段,如何实现高质量的实体对齐、共指消解是一项巨大的挑战,开放领域下的实体对齐、共指消解以及多源数据库融合是当前主要的研究重点,如何在短文本文情况下准确地将实体链接到知识库中亟需解决。小样本学习是最具潜力的研究方法之一,在计算机视觉中取得了较好的效果,使得模型具有高效的学习能力,将小样本学习应用于实体对齐具有重要的研究价值。

(3)对于知识推理技术,现有的知识推理技术大多只关注静态数据,忽略了时间信息,知识图谱的信息应该随着时间的推移而变化,因此,动态知识图谱推理还需要进行研究探索,如 EvolveGCN^[93],其使用 RNN 来演化 GCN 的参数,捕捉图序列的动态信息。除此之外,如何利用文本、音频、图像等多源信息来进行知识推理以及跨语言的知识推理,也是一个新的研究方向。

2)知识图谱应用于深度学习方面

目前,将知识图谱应用于深度学习还处于起步阶段,存在一些问题亟需解决,例如,知识图谱由点和边组成,符号化的知识在向量化后不可避免地会导致语义缺失问题,因此,如何对知识图谱进行高质量的表示仍是一个挑战。

5 结束语

作为知识工程的一个重要分支,知识图谱已经成为人工智能发展的核心动力和重要领域。本文对比分析知识图谱构建技术,讨论知识图谱与深度学习相结合的方法,并给出研究难点与潜在解决思路。从当前研究进展可以看出,将深度学习用于知识图谱构建已成为主流方法并取得了较好的效果,但无论是知识图谱构建还是深度学习,目前都存在一些缺陷和挑战。在未来,基于深度学习的开放领域知识抽取、多源知识融合、动态知识推理,以及利用知识图谱弥补深度学习的某些缺陷,将是知识图谱领域的重要研究方向。

参考文献

- [1] MIN C, MAO S, LIU Y. Big data: a survey[J]. *Mobile Networks & Applications*, 2014, 19(2): 171-209.
- [2] 付小红. 论知识组织的原则[J]. *情报资料工作*, 2001(5): 11-15.
FU X H. Principles on knowledge organization [J]. *Information and Documentation Services*, 2001(5): 11-15. (in Chinese)
- [3] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述[J]. *电*

子科技大学学报, 2016, 45(4): 589-606.

- XU Z L, SHENG Y P, HE L R, et al. Review on knowledge graph techniques[J]. *Journal of University of Electronic Science and Technology of China*, 2016, 45(4): 589-606. (in Chinese)
- [4] 李渭子, 侯磊. 知识图谱研究综述[J]. *山西大学学报(自然科学版)*, 2017, 40(3): 454-459.
LI J Z, HOU L. Reviews on knowledge graph research[J]. *Journal of Shanxi University (Natural Science Edition)*, 2017, 40(3): 454-459. (in Chinese)
- [5] AI-MOSLMI T, OCAA M G, OPDAHL A L, et al. Named entity extraction for knowledge graphs: a literature overview[J]. *IEEE Access*, 2020, 8: 32862-32881.
- [6] SMIRNOVA A, CUDRE-MAUROUX P. Relation extraction using distant supervision: a survey[J]. *ACM Computing Surveys*, 2019, 51(5): 1-35.
- [7] KUMAR S. A survey of deep learning methods for relation extraction[EB/OL]. [2021-04-05]. <https://arxiv.org/pdf/1705.03645.pdf>.
- [8] 庄传志, 靳小龙, 朱伟建, 等. 基于深度学习的关系抽取研究综述[J]. *中文信息学报*, 2019, 33(12): 1-18.
ZHUANG C Z, JIN X L, ZHU W J, et al. Deep learning based relation extraction: a survey[J]. *Journal of Chinese Information Processing*, 2019, 33(12): 1-18. (in Chinese)
- [9] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a Web-scale approach to probabilistic knowledge fusion [C]//*Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA: ACM Press, 2014: 601-610.
- [10] HAO Y, ZHANG Y, LIU K, et al. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge[C]//*Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. [S. l.]: ACL, 2017: 221-231.
- [11] GONG F, WANG M, WANG H, et al. SMR: medical knowledge graph embedding for safe medicine recommendation[J]. *Big Data Research*, 2021, 23: 100174.
- [12] 王昊奋, 漆桂林, 陈华钧. 知识图谱: 方法, 实践与应用[M]. 北京: 电子工业出版社, 2019.
WANG H F, QI G L, CHEN H J. Knowledge graph: method, practice and application[M]. Beijing: Publishing House of Electronics Industry, 2019. (in Chinese)
- [13] RAU L F. Extracting company names from text [C]//*Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications*. Washington D. C., USA: IEEE Press, 1991: 12-23.
- [14] HUMPHREYS K, GAIZAUSKAS R, AZZAM S, et al. University of sheffield: description of the LaSIE-II system as used for MUC-7 [M]. [S. l.]: Association for Computational Linguistics, 1998.
- [15] AITKEN J S. Learning information extraction rules: an inductive logic programming approach[C]//*Proceedings of the 15th European Conference on Artificial Intelligence*. Berlin, Germany: Springer, 2002: 102-112.
- [16] SCHUTZ A, BUITELAAR P. RelExt: a tool for relation extraction from text in ontology extension[C]//*Proceedings of the 4th International Semantic Web Conference*. Berlin, Germany: Springer, 2005: 593-606.
- [17] ZHOU G D, ZHANG J, SU J, et al. Recognizing names in biomedical texts: a machine learning approach[EB/OL].

- [2021-04-05]. <https://academic.oup.com/bioinformatics/article/20/7/1178/245780>.
- [18] LIU X, ZHANG S, WEI F, et al. Recognizing named entities in tweets [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. [S. l.]: ACL, 2011; 14-18.
- [19] SHINYAMA Y, SEKINE S. Preemptive information extraction using unrestricted relation discovery [C]//Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. [S. l.]: ACL, 2006; 304-311.
- [20] MA X, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [EB/OL]. [2021-04-05]. https://www.aclweb.org/old_anthology/P/P16/P16-1101.pdf.
- [21] 杨飘,董文永. 基于BERT嵌入的中文命名实体识别方法[J]. 计算机工程, 2020, 46(4): 40-45, 52.
YANG P, DONG W Y. Chinese named entity recognition method based on BERT embedding [J]. Computer Engineering, 2020, 46(4): 40-45, 52. (in Chinese)
- [22] GAN T, GAN Y, YANMIN H E. Subsequence-level entity attention LSTM for relation extraction [C]//Proceedings of 2019 International Computer Conference on Wavelet Active Media Technology and Information Processing. Washington D. C., USA: IEEE Press, 2019; 102-113.
- [23] XIAO J, ZHOU Z. Chapter-level entity relationship extraction method based on joint learning [C]//Proceedings of 2020 International Conference on Intelligent Human-Machine Systems and Cybernetics. Washington D. C., USA: IEEE Press, 2020; 145-156.
- [24] RONAN C, JASON W, LÉON B, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12: 2493-2537.
- [25] QIU J, ZHOU Y, WANG Q, et al. Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field [J]. IEEE Transactions on NanoBioscience, 2019, 12: 306-315.
- [26] KONG J, ZHANG L, JIANG M, et al. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition [J]. Journal of Biomedical Informatics, 2021, 116: 103737.
- [27] HUANG Z, WEI X, KAI Y. Bidirectional LSTM-CRF models for sequence tagging [EB/OL]. [2021-04-05]. <https://arxiv.org/pdf/1508.01991.pdf>.
- [28] UKOV-GREGORI A, BACHRACH Y, COOPE S. Named entity recognition with parallel recurrent neural networks [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. [S. l.]: ACL, 2018; 69-74.
- [29] RONRAN C, LEE S. Effect of character and word features in bidirectional LSTM-CRF for NER [C]//Proceedings of 2020 IEEE International Conference on Big Data and Smart Computing. Washington D. C., USA: IEEE Press, 2020; 14-25.
- [30] YAN H, DENG B, LI X, et al. TENER: adapting transformer encoder for name entity recognition [EB/OL]. [2021-04-05]. <https://arxiv.org/pdf/1911.04474.pdf>.
- [31] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2021-04-05]. <https://aclanthology.org/N19-1423.pdf>.
- [32] 曾青霞,熊旺平,杜建强,等. 结合自注意力的BiLSTM-CRF的电子病历命名实体识别[J]. 计算机应用与软件, 2021, 38(3): 159-162, 242.
ZENG Q X, XIONG W P, DU J Q, et al. Electronic medical record named entity recognition combined with self-attention BiLSTM-CRF [J]. Computer Applications and Software, 2021, 38(3): 159-162, 242. (in Chinese)
- [33] 罗熹,夏先运,安莹,等. 结合多头自注意力机制与BiLSTM-CRF的中文临床实体识别[J]. 湖南大学学报(自然科学版), 2021, 48(4): 45-55.
LUO X, XIA X Y, AN Y, et al. Chinese CNER combined with multi-head self-attention and BiLSTM-CRF [J]. Journal of Human University (Natural Sciences), 2021, 48(4): 45-55. (in Chinese)
- [34] 于浏洋,郭志刚,陈刚,等. 面向知识图谱构建的知识抽取技术综述[J]. 信息工程大学学报, 2020, 21(2): 227-235.
YU L Y, GUO Z G, CHEN G, et al. Summary of knowledge graph construction oriented knowledge extraction technology [J]. Journal of Information Engineering University, 2020, 21(2): 227-235. (in Chinese)
- [35] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network [EB/OL]. [2021-04-05]. http://www.nlpr.labs.gov.cn/cip/liukang.files/camera_coling2014_final.pdf.
- [36] XU K, FENG Y, HUANG S, et al. Semantic relation classification via convolutional neural networks with simple negative sampling [J]. Computer Science, 2015, 71: 941-949.
- [37] GUO X Y, ZHANG H Y, XU H J, et al. A single attention-based combination of CNN and RNN for relation classification [J]. IEEE Access, 2019, 7: 12467-12475.
- [38] 闫雄,段跃兴,张泽华. 采用自注意力机制和CNN融合的实体关系抽取[J]. 计算机工程与科学, 2020, 42(11): 2059-2066.
YAN X, DUAN Y X, ZHANG Z H. Entity relationship extraction fusing self-attention mechanism and CNN [J]. Computer Engineering & Science, 2020, 42(11): 2059-2066. (in Chinese)
- [39] 龚乐君,刘晓林,高志宏,等. 基于双向GRU和CNN的药物相互作用关系抽取[J]. 陕西师范大学学报(自然科学版), 2020, 48(6): 108-113.
GONG L J, LIU X L, GAO Z H, et al. Extraction of drug-drug interaction based on bidirectional GRU and CNN [J]. Journal of Shaanxi Normal University (Natural Science Edition), 2020, 48(6): 108-113. (in Chinese)
- [40] YIN B, SUN Y, WANG Y. Entity relation extraction method based on fusion of multiple information and attention mechanism [C]//Proceedings of 2020 IEEE International Conference on Computer and Communications. Washington D. C., USA: IEEE Press, 2020; 145-160.
- [41] SOCHER R, HUVAL B, MANNING C D, et al. Semantic compositionality through recursive matrix-vector spaces [C]//Proceedings of Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) & Computational Natural Language Learning (CoNLL). Washington D. C., USA: IEEE Press, 2012; 1201-1211.
- [42] SHU Z, ZHENG D, HU X, et al. Bidirectional long short-term memory networks for relation classification [C]//Proceedings of the 29th Pacific Asia Conference on

- Language, Information and Computation. Washington D. C., USA: IEEE Press, 2015: 73-78.
- [43] GENG Z, CHEN G, HAN Y, et al. Semantic relation extraction using sequential and tree-structured LSTM with attention[J]. Information Sciences, 2020, 509: 183-192.
- [44] 张勇, 高大林, 巩敦卫, 等. 用于关系抽取的注意力图长短时记忆神经网络[J]. 智能系统学报, 2021, 16(3): 518-527.
- ZHANG Y, GAO D L, GONG D W, et al. Attention graph long short term memory neural network for relation extraction[J]. CAAI Transactions on Intelligent Systems, 2021, 16(3): 518-527. (in Chinese)
- [45] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. [S. l.]: ACL, 2016: 1105-1116.
- [46] KATYAR A, CARDIE C. Going out on a limb: joint extraction of entity mentions and relations without dependency trees [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. [S. l.]: ACL, 2017: 917-928.
- [47] ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme[EB/OL]. [2021-04-05]. <https://arxiv.org/pdf/1706.05075.pdf>.
- [48] LIU Y, LI A, HUANG J, et al. Joint extraction of entities and relations based on multi-label classification [C]// Proceedings of 2019 IEEE International Conference on Data Science in Cyberspace. Washington D. C., USA: IEEE Press, 2019: 102-113.
- [49] WEI Z, SU J, WANG Y, et al. A novel hierarchical binary tagging framework for relational triple extraction[EB/OL]. [2021-04-05]. <https://arxiv.org/pdf/1909.03227v2.pdf>.
- [50] WANG C, LI A, TU H, et al. An advanced BERT-based decomposition method for joint extraction of entities and relations [C]// Proceedings of 2020 IEEE International Conference on Data Science in Cyberspace. Washington D. C., USA: IEEE Press, 2020: 145-163.
- [51] 林海伦, 王元卓, 贾岩涛, 等. 面向网络大数据的知识融合方法综述[J]. 计算机学报, 2017, 40(1): 1-27.
- LIN H L, WANG Y Z, JIA Y T, et al. Network big data oriented knowledge fusion methods: a survey[J]. Chinese Journal of Computers, 2017, 40(1): 1-27. (in Chinese)
- [52] KALFOGLOU Y, SCHORLEMMER M. Ontology mapping: the state of the art[J]. The Knowledge Engineering Review, 2003, 18(1): 1-31.
- [53] PINTO H S, MARTINS J P. A methodology for ontology integration [C]// Proceedings of the 1st International Conference on Knowledge Capture. Washington D. C., USA: IEEE Press, 2001: 123-146.
- [54] PORTER M F. An algorithm for suffix stripping [J]. Program Electronic Library and Information Systems, 1980, 14(3): 130-137.
- [55] NOY N F, MUSEN M A. The PROMPT suite: interactive tools for ontology merging and mapping[J]. International Journal of Human-Computer Studies, 2003, 59(6): 983-1024.
- [56] NEWCOMBE H B, KENNEDY J M, AXFORD S J, et al. Automatic linkage of vital records[J]. Science, 1959, 130(3381): 954-959.
- [57] 张伟莉, 黄廷磊, 梁霄. 基于半监督协同训练的百科知识库实体对齐[J]. 计算机与现代化, 2017(12): 88-93.
- ZHANG W L, HUANG T L, LIANG X. Instance alignment algorithm between encyclopedia based on semi-supervised co-training[J]. Computer and Modernization, 2017(12): 88-93. (in Chinese)
- [58] LI J, WANG Z, XIAO Z, et al. Large scale instance matching via multiple indexes and candidate selection[J]. Knowledge-Based Systems, 2013, 50: 112-120.
- [59] BHATTACHARYA I, GETOOR L. Collective entity resolution in relational data[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 5-10.
- [60] LACOSTE-JULIEN S, PALLA K, DAVIES A, et al. SiGMA: simple greedy matching for aligning large knowledge bases [C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2013: 10-13.
- [61] WICK M, SINGH S, MCCALLUM A. A discriminative hierarchical model for fast coreference at large scale [C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. [S. l.]: ACL, 2012: 379-388.
- [62] RASTOGI V, DALVI N, GAROFALAKIS M. Large-scale collective entity matching[J]. Proceedings of the VLDB Endowment, 2011, 4(4): 208-218.
- [63] SCHOENMACKERS S, DAVIS J, ETZIONI O, et al. Learning first-order horn clauses from Web text [C]// Proceedings of Conference on Empirical Methods in Natural Language Processing. New York, USA: ACM Press, 2010: 1088-1098.
- [64] CHEN Y, GOLDBERG S, WANG D Z, et al. Ontological pathfinding: mining first-order knowledge from large knowledge bases [C]// Proceedings of 2016 ACM SIGMOD Conference on Management of Data. New York, USA: ACM Press, 2016: 835-846.
- [65] 刘峻, 韩明皓, 江浏祎, 等. 基于双层随机游走的关系推理算法[J]. 计算机学报, 2017, 40(6): 1275-1290.
- LIU Q, HAN M H, JIANG L Y, et al. Two-tier random walk based relational inference algorithm[J]. Chinese Journal of Computers, 2017, 40(6): 1275-1290. (in Chinese)
- [66] WANG C, CHENG P. Translating representations of knowledge graphs with neighbors [C]// Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York, USA: ACM Press, 2018: 917-920.
- [67] SUN Z Q, DENG Z H, NIE J Y, et al. Rotate: knowledge graph embedding by relational rotation in complex space [EB/OL]. [2021-04-05]. <http://arxiv.org/pdf/1902.10197.pdf>.
- [68] 彭敏, 黄婷, 田纲, 等. 聚合邻域信息的联合知识表示模型[J]. 中文信息学报, 2021, 35(5): 46-54.
- PENG M, HUANG T, TIAN G, et al. Neighborhood aggregation for knowledge graph representation[J]. Journal of Chinese Information Processing, 2021, 35(5): 46-54. (in Chinese)
- [69] 宋浩楠, 赵刚, 王兴芬. 融合知识表示和深度强化学习的知识推理方法[J]. 计算机工程与应用, 2021, 57(19): 189-197.

- SONG H N, ZHAO G, WANG X F, et al. Knowledge reasoning method combining knowledge representation with deep reinforcement learning [J]. Computer Engineering and Applications, 2021, 57(19): 189-197. (in Chinese)
- [70] NICKEL M, TRESP V, KRIEGER H P. A three-way model for collective learning on multi-relational data [C]// Proceedings of International Conference on Machine Learning. Washington D. C., USA: IEEE Press, 2011: 809-816.
- [71] CHANG K W, YIH W T, YANG B, et al. Typed tensor decomposition of knowledge bases for relation extraction [EB/OL]. [2021-04-05]. <https://mirror.aclweb.org/emnlp2014/papers/pdf/EMNLP2014165.pdf>.
- [72] 吴运兵, 朱丹红, 廖祥文, 等. 路径张量分解的知识图谱推理算法[J]. 模式识别与人工智能, 2017, 30(5): 473-480.
- WU Y B, ZHU D H, LIAO X W, et al. Knowledge graph reasoning based on paths of tensor factorization[J]. Pattern Recognition and Artificial Intelligence, 2017, 30(5): 473-480. (in Chinese)
- [73] YANG B, YIH W T, HE X, et al. Embedding entities and relations for learning and inference in knowledge bases[EB/OL]. [2021-04-05]. http://scotttyih.org/files/ICLR2015_updated.pdf.
- [74] 刘峤, 韩明皓, 杨晓慧, 等. 基于表示学习和语义要素感知的关系推理算法[J]. 计算机研究与发展, 2017, 54(8): 1682-1692.
- LIU Q, HAN M H, YANG X H, et al. Representation learning based relational inference algorithm with semantical aspect awareness [J]. Journal of Computer Research and Development, 2017, 54(8): 1682-1692. (in Chinese)
- [75] ZHANG S, TAY Y, YAO L N, et al. Quaternion knowledge graph embeddings [C]// Proceedings of the 33rd Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2019: 2731-2741.
- [76] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion [C]// Proceedings of the 26th International Conference on Neural Information Processing. New York, USA: ACM Press, 2013: 926-930.
- [77] DAS R, NEELAKANTAN A, BELANGER D, et al. Chains of reasoning over entities, relations, and text using recurrent neural networks [C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. [S. l.]: ACL, 2017: 105-110.
- [78] GUO L B, ZHANG Q H, GE W Y, et al. DSKG: a deep sequential model for knowledge graph completion [C]// Proceedings of 2018 China Conference on Knowledge Graph and Semantic Computing. Berlin, Germany: Springer, 2018: 65-77.
- [79] CHEN W R, HONG D P, ZHENG C. Learning knowledge graph embedding with entity descriptions based on LSTM networks [C]// Proceedings of 2020 IEEE International Symposium on Product Compliance Engineering-Asia. Washington D. C., USA: IEEE Press, 2020: 1-7.
- [80] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2D knowledge graph embeddings [C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2017: 1811-1818.
- [81] VASHISHTH S, SANYAL S, NITIN V, et al. InteractE: improving convolution-based knowledge graph embeddings by increasing feature interactions [C]// Proceedings of AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2020: 3009-3016.
- [82] 李少杰, 陈曙东, 郝悦星, 等. 基于卷积神经网络的高效知识表示模型[J]. 高技术通讯, 2020, 30(9): 901-907.
- LI S J, CHEN S D, HAO Y X, et al. A novel knowledge representation model based on convolutional neural network [J]. Chinese High Technology Letters, 2020, 30(9): 901-907. (in Chinese)
- [83] CATHERINE R, COHEN W. Personalized recommendations using knowledge graphs: a probabilistic logic programming approach [C]// Proceedings of ACM Conference on Recommender Systems. New York, USA: ACM Press, 2016: 11-22.
- [84] CAO Y, XIANG W, HE X, et al. Unifying knowledge graph learning and recommendation: towards a better understanding of user preferences [EB/OL]. [2021-04-05]. <https://arxiv.org/pdf/1902.06236.pdf>.
- [85] ZHANG Q, CAO R, SHI F, et al. Interpreting CNN knowledge via an explanatory graph [EB/OL]. [2021-04-05]. <https://arxiv.org/pdf/1708.01785.pdf>.
- [86] HU Z, MA X, LIU Z, et al. Harnessing deep neural networks with logic rules [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. [S. l.]: ACL, 2016: 2410-2420.
- [87] WANG H, POON H. Deep probabilistic logic: a unifying framework for indirect supervision [EB/OL]. [2021-04-05]. <https://arxiv.org/pdf/1808.08485.pdf>.
- [88] LI T, SRIKUMAR V. Augmenting neural networks with first-order logic [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. [S. l.]: ACL, 2019: 102-113.
- [89] WANG W, PAN S J. Integrating deep learning with logic fusion for information extraction [EB/OL]. [2021-04-05]. <https://arxiv.org/pdf/1912.03041v1.pdf>.
- [90] PEARL J. Theoretical impediments to machine learning with seven sparks from the causal revolution [C]// Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York, USA: ACM Press, 2018: 3-9.
- [91] BATTAGLIA P W, HAMRICK J B, BAPST V, et al. Relational inductive biases, deep learning, and graph networks [EB/OL]. [2021-04-05]. <https://arxiv.org/pdf/1806.01261.pdf>.
- [92] 康雁, 李涛, 李浩, 等. 融合知识图谱与协同过滤的推荐模型[J]. 计算机工程, 2020, 46(12): 73-79, 87.
- KANG Y, LI T, LI H, et al. Recommendation model fusing with knowledge graph and collaborative filtering [J]. Computer Engineering, 2020, 46(12): 73-79, 87. (in Chinese)
- [93] PAREJA A, DOMENICONI G, CHEN J, et al. EvolveGCN: evolving graph convolutional networks for dynamic graphs [EB/OL]. [2021-04-05]. <https://arxiv.org/pdf/1902.10191.pdf>.

编辑 吴云芳