知识图谱技术进展及展望*

覃　晓[1],廖兆琪[1],施　宇[1],元昌安[2**]
(1.南宁师范大学,广西南宁　530299,2.广西科学院,广西南宁　530007)

摘要:随着大数据的发展,知识图谱(Knowledge Graph)关键技术及其应用成为人工智能最热门的研究领域之一。本文从知识图谱的定义、架构以及常见的知识库出发,对知识图谱构建的知识表达和知识自动获取技术进行总结和回顾,讨论其研究要点和发展趋势,介绍知识图谱技术常见的应用场景,并结合本团队的研究对知识图谱的发展趋势进行展望。

关键词:自然语言处理　知识图谱　知识表达　知识抽取　模型改进

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1002-7378(2020)03-0242-10
DOI:10.13657/j.cnki.gxkxyxb.20201027.009

0　引言

知识图谱(Knowledge Graph)的研究最早可追溯到1977年,在第五届国际人工智能会议上,美国计算机科学家 Feigenbaum B. A. 首次提出知识工程(Knowledge Engineering)的概念。知识工程即针对用户提出的问题用知识库中已有的知识来求解的系统,其中最经典的是专家系统。

2012年5月17日,谷歌(Google)发布知识图谱项目,并宣布以此为基础构建下一代智能化搜索引擎[1]。该项目通过对客观真实世界中各种实体及其关系的描绘,形成一张巨大的语义结构网络图,使各种庞杂无关的知识联系起来,从而达到便捷地获取知识的目的。

如果按使用范围划分,知识图谱可划分为领域知识图谱和通用知识图谱。对于特定领域知识,通过知识库的理论进行组织和管理较为有效。根据全球机构库统计网站开放获取知识库名录的数据,截至2014年4月,大约有2 616个知识库已在该网站注册,其中包含机构知识库2 212个,占总数的84.56%。在国内,始建于2007年的中国科学院知识库为全民提供大量的知识学习资源。另外,许多高校也开始构造或已经构造自己的知识库系统。

通用知识图谱指的是由世界知识构成的语义网络。从2006年开始,随着大规模百科资源的出现以及知识提取方法的进步,知识工程取得重大进展。典型的例子是谷歌收购 Freebase 后在2012年推出的知识图谱。最具代表性的大规模网络知识获取工作包括 DBpedia、Freebase、KnowItAll、WikiTaxonomy 和 YAGO,以及 BabelNet、ConceptNet、DeepDive、

*国家自然基金项目(61962006)和广西创新驱动重大项目(AA18118047)资助。

【作者简介】
覃　晓(1973—),女,副教授,主要从事人工智能、图像处理研究。
【**通信作者】
元昌安(1964—),男,博士,教授,主要从事人工智能与数据挖掘研究,E-mail:68852917@qq.com。

【引用本文】
覃晓,廖兆琪,施宇,等.知识图谱技术进展及展望[J].广西科学院学报,2020,36(3):242-251.
QIN X,LIAO Z Q,SHI Y,et al. Progress and Prospect of Knowledge Graph Technology [J]. Journal of Guangxi Academy of Sciences,2020,36(3):242-251.

NELL、Probase、Wikidata、XLORE、Zhishi.me、CNDBpedia等。这些知识图谱遵循RDF数据模型,包含数以千万级或者亿级规模的实体,并且这些实体被组织到各种客观世界的概念中。

1 知识图谱研究基础

1.1 知识图谱定义

知识图谱是将大量收集的数据整理成机器能处理的知识库,并实现可视化的展示。知识图谱本质上是一种大规模的语义网络,其主要目的是对真实世界里实体或概念之间的关联关系进行描述。

三元组是知识图谱的一种基本表示方式,即$G=(E,R,S)$,其中$E=\{e_1,e_2,\cdots,e_{|E|}\}$是知识库中的实体集,共包含$|E|$种实体;$R=\{r_1,r_2,\cdots,r_{|E|}\}$是知识库中的关系集合,共包含$|R|$种关系;$S\subseteq E\times R\times E$代表知识库中的三元组集合。三元组的主要结构是实体-关系-实体,以及各种概念、属性和属性值等,其中实体是其最基本的元素。概念主要指集合、类别、对象类型等;属性主要指对象可能具有的属性、特征、特性等;属性值主要指对象指定属性的值。实体可以通过特有的标签来表示,关系则用来联系两个实体[2]。

1.2 知识图谱架构

知识图谱的体系架构是指构造该图谱模型的结构,如图1所示。其中虚线框内的部分为知识图谱的模块构造过程。

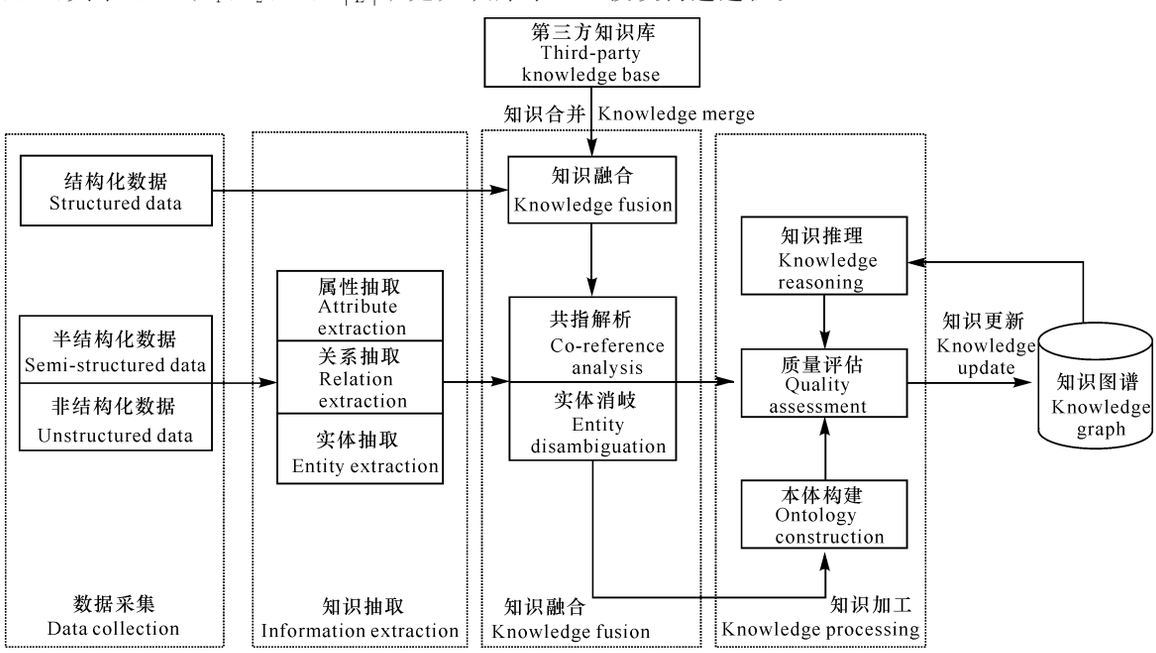

图1 知识图谱的体系架构

Fig.1 Architecture of the knowledge graph

知识图谱主要有自顶向下与自底向上两种构造方式。自顶向下指的是先定义所需要的模式,再将各种实体知识加入知识库中。自底向上指的是先从各种数据中抽取实体,再筛选出置信度较高的实体去构造顶层的模式[3]。

知识图谱的体系架构展现了构造知识图谱的几个关键步骤,包括数据采集、知识抽取、知识融合、知识加工、知识更新等过程,其中,从数据采集到知识抽取还需要恰当的知识表达技术。本文着重就知识表达和知识抽取两个关键技术进行阐述。

1.3 知识图谱研究常见的知识库

为了高效存储与利用结构化知识,人们结合专家手工标注与计算机自动标注等方式,面向开放领域和垂直领域构建了各种大规模知识图谱。如来自罗马萨皮恩萨大学的Roberto Navigli是BabelNet的创始人[4],BabelNet目前是最大的高质量多语言百科全书计算机辞典,一个覆盖广泛的大型多语言语义网络。BabelNet网络能够自动将最大的多语Web百科全书——维基百科,链接到最常用的英语计算词典WordNet。除此之外,机器翻译也能够让所有语种的词汇信息资源等更丰富,已有的BabelNet(v3.7)已覆盖271种语言,包括全部的欧洲语言、大多数亚洲语言及拉丁语。在新的标准数据集和现有的标准数据集上进行实验,结果也证明这个资源具有很高的品质和很广的覆盖范围。

来自 Max-Planck 信息学研究所的 Hoffart 等[5]提出的 YAGO2,是 YAGO 知识库的一个拓展;实体、事实和事件在 YAGO2 知识库里都被按照时间和空间的顺序进行排序;YAGO2 涵盖 980 万个实体的 4.47 亿个事实,这些事实数据都在 GeoNames、维基百科以及 WordNet 上自动构建形成,经过专家的评估确认,其中有 95%的事实是正确的。

此外,还有 WikiData[6]、Freebase[7]、DBpedia[8]、WordNet[9]等经典知识库。以 WikiData 为例,目前其已经包含 5 700 多万个实体。与此同时,国内外各大互联网公司也均有各自的知识图谱产品,如谷歌知识图谱、百度知心、同方、搜狗知立方和微软(Microsoft) Bing Satori 等。

2 知识图谱关键技术研究现状

2.1 知识表示技术

知识表示是知识图谱研究首先需要讨论的技术。鄢珞青[10]对知识表达方面的相关知识做了细致的研究,提出知识点的概念并讨论了各种知识表达的类型等。王知津等[11]在对知识组织各个方面进行分析后提出多维性原则、科学性原则等十大原则。王军等[12]着重对互联网环境下知识的组织结构进行系统化讨论,针对网络知识组织系统的各种应用层面进行细致的介绍。知识表达组织需要根据整个知识库系统的需求及其框架来确定。当今,比较常用的知识表达框架主要基于面向对象,将知识分解为实体与实体间的关系。

近年来,知识表示学习由于深度学习的发展也获得了相应的成果,并逐渐成为前沿研究的热点。知识表示学习主要是对知识库中的实体以及它们之间的关系进行学习,将其中的语义知识信息向量化,从而在低维空间中实现高效计算实体和关系的语义联系,不但有效解决数据稀疏的问题,而且使知识获取、融合和推理的效果更为有效。国外关于知识库的研究更侧重实践方面,并且主要针对网络知识组织系统进行相关的研发工作,例如对在线图书馆的研究等[13]。

2.1.1 知识表示学习经典模型

(1)神经张量模型

神经张量模型[14]的基本思想:在不同维度下,将实体联系起来,表示实体间复杂的语义联系。模型为知识库中的每个三元组 (h,r,t) 定义了以下形式的评价函数:

$$f_r(h,t)=\mu_r^T g(l_h M_r l_t+M_{r,1}l_h+M_{r,2}l_t+b_r),$$

式中,$\mu_r^T\in R^k$ 为关系 r 的向量化表示;$g()$ 为 $tanh$ 函数;$M_r\in R^{d\times d\times k}$ 是一个三阶张量;$M_{r,1}$、$M_{r,2}\in R^{d\times k}$ 是通过关系 r 定义的两个投影矩阵。

神经张量模型在构造实体的向量表示时,是将该实体中的所有单词的向量取平均值,这样一方面可以反复使用单词向量构造实体,另一方面将有利于增强低维向量的稠密程度以及实体与关系的语义计算。

(2)矩阵分解模型

通过矩阵分解的方式可得到低维的向量表示,因此相关模型被开发出来,其中的典型代表是 RESACL 模型[15]。

在 RESCAL 模型中,知识库中的三元组 (h,r,t) 集合被表示为一个三阶张量,如果该三元组存在,张量中对应位置的元素被置为 1,否则置为 0。通过张量分解算法,可将张量中每个三元组 (h,r,t) 对应的张量值 X_{hrt} 分解为双线性模型中的知识表示形式 $l_h^T M_r l_t$,并使 $|X_{hrt}-l_h^T M_r l_t|_{L_2}$ 尽量小。

(3)翻译模型

受平移不变现象的启发,有研究团队提出 TransE 模型[16],即将知识库中实体之间的关系看成是从实体间的某种平移,并用向量表示。关系 l_r 可以看作是从头实体向量 l_h 到尾实体向量 l_t 的翻译。对于知识库中的每个三元组 (h,r,t),TransE 都希望满足以下关系:$l_h+l_r\approx l_t$,其损失函数为

$$f_r(h,r,t)=|l_h+l_r-l_t|_{L_1/L_2},$$

即向量 l_h+l_r 与 L_t 的 L_1 或 L_2 距离。该模型的参数较少,计算的复杂度显著降低,并且该模型具有较好的性能与扩展性。

2.1.2 知识表示学习改进模型

尽管知识表示学习经典模型具有很好的效率和结果,并被广泛应用于知识表示学习任务中,但经典模型仍存在难以表达复杂关系、未充分利用多步关系路径信息的不足。有研究人员尝试将复杂关系、多步路径关系信息进行嵌入表达,如 Tang 等[17]针对知识表示学习中的复杂关系建模进行研究,提出一种基于距离的链接预测知识图嵌入方法。这个方法先是使用正交关系变换把 RotatE 拓展到高维空间上,然后把图结构的信息集成到距离评分函数中,用于训练和推理过程中度量三元组的相似性。Nguyen 等[18]提出基于关系记忆网络的 Embedding 模型,这个模型充分利用三元组之间潜在的依赖关系,其中包含多头注意力机制编码,并且在三元组分类中验证了模型的效果比当前最新的模型好。Zhang 等[19]提出一种名

为 CrossE 的新型知识图谱嵌入,该模型可以正确地模拟交叉交互。它不仅能像大多数已有方法一样为每个实体和关系学习生成一个通用嵌入,还为这两者之间生成多个三重特定的嵌入,即交互嵌入。通过对典型链接预测任务上的嵌入评估,发现 CrossE 可以在很复杂的数据集上得到良好效果。同时从新的角度评估嵌入,然后对头尾实体之间的可靠闭合路径给出解释,完成三元组的预测。Lin 等[20]提出将关系路径信息嵌入知识表示学习模型 PTransE。PTransE 提供一种知识图谱的新型表示方法,通过编码关系路径将实体和关系嵌入一个低维空间之中,与传统方法相比,PTransE 在知识图谱补全和文本关系抽取任务上取得了显著的改进效果。

复杂关系知识表示模型 TransR 和关系路径知识表示 PTransE 是关系表示学习的成功改进模型。

(1)复杂关系知识表示模型 TransR

经典知识表示模型的三元组(h,r,t)中,关系 r 仅代表一种语义。但实际情况下,同一个实体在不同的关系场景下具有的语义是有区别的。比如"绣球"在民族文化语义中是广西壮族手工艺品,是壮家人的定情物和吉祥物,但在植物科目中"绣球"却是蔷薇目虎耳草科植物;"韦启初"是环江韦氏仿古壮族铜鼓铸造厂厂长,又是广西壮族自治区级非物质文化遗产"壮族铜鼓铸造技艺"代表性传承人。为扩展经典知识表示模型复杂关系的表达能力,Lin 等[21]提出 TransR 模型,该模型基于复杂关系知识表示建模,为每一种关系 r 定义单独的语义空间,并使用不同的映射矩阵 Mr 将经典模型中的实体映射到关系空间中(图 2)。

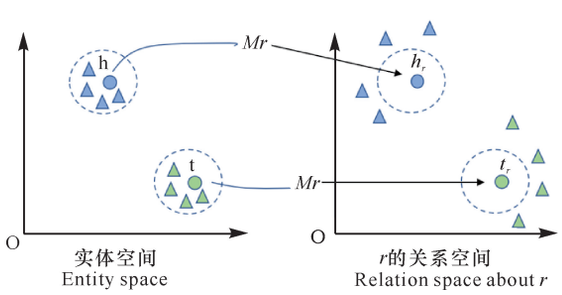

图 2 实体空间到关系空间映射

Fig. 2 Mapping from entity space to relation space

以翻译模型 TransE 为基础,考虑复杂关系的知识表示模型,在关系 r 所在的空间中,h_r 和 t_r 满足的损失函数与 TransE 相同。即

$$f_r(h,r,t)=|l_{hr}+l_r-l_{tr}|_{L_1/L_2}。$$

(2)关系路径知识表示模型 PTransE

知识图谱嵌入的主要目的是学习实体和关系的分布式表示形式,而交叉交互有助于再预测新的三元组时选择相关信息。经典知识表示模型往往只考虑实体之间的直接关系,但现实世界中,知识图谱中的实体具有多步路径关系现象,且多步路径关系包含大量的语义信息。 例如:关系路径 $h \xrightarrow{(\text{文化传承人})} e_1 \xrightarrow{(\text{文化所属民族})} e_2 \xrightarrow{(\text{民族所属地})} e_3 \xrightarrow{(\text{城市归属})} t$ 隐含实体 h 和 t 的居住地关系,即$(h,$居住地$,t)$。可见,将关系路径的特征表达嵌入知识表示中,是知识表示的一个重要研究工作。

PTransE 为关系三元组定义的损失函数考虑实体间多步关系路径信息:

$$f(h,r,t)=E(h,r,t)+E(h,P,t),$$

其中,$E(h,r,t)$代表实体 h 和 t 之间直接关系的相关性,而 $E(h,P,t)$则刻画多步路径所蕴含的关系信息。PTransE 模型将 P 看作是多条关系路径 p 的嵌入表示,每一条关系路径 p 看作多步关系信息得到的实体间关系的近似。于是 $E(h,P,t)$就可以定义为

$$E(h,P,t)=\sum\nolimits_{p\in P(h,t)} R(p\mid h,t)E(h,p,t),$$

其中,$R(p|h,t)$表示路径 p 上信息量,而 $E(h,p,t)$为 p 与 r 的接近程度,$E(h,p,t)=\|p-r\|_{L_1/L_2}$。

2.2 知识自动获取技术

近年来,尽管很多大型知识图谱,如 Freebase、DBpedia、YAGO 等在问答系统、文本检索等领域取得显著效果,但是这些大型知识图谱所涵盖的知识,与现实世界无穷无尽的知识比较起来,是不够完善的。因此,知识自动获取成为丰富知识图谱知识和提高知识获取效率的重要课题。在知识的自动获取技术上,关系抽取是其核心。关系抽取的目标是解决实体间语义链接的问题,最初的关系抽取是通过人为构造规则的方法,随后,实体间的关系模型逐渐替代人工预定义的语法与规则。文献[22]提出面向开放域的信息抽取框架(Open Information Extraction,OIE)。但 OIE 方法在对实体的隐含关系抽取方面性能低下,因此部分学者提出基于马尔可夫逻辑网(Markov Logic Network,MLN)以及基于本体推理的深层隐含关系抽取方法[23]。

2.2.1 传统的关系抽取模型

(1)开放式实体关系抽取

开放式实体关系抽取可分为二元开放式关系抽取和 n 元开放式关系抽取。在二元开放式关系抽取中,早期的研究有 KnowItAll[24] 与 TextRunner[25] 系统,但是表现一般。Suchanek 等[26] 提出一种基于 Wikipedia 的 OIE 方法——WOE,经自监督学习得到提取器,准确率较 TextRunner 有显著提高。

(2)基于联合推理的实体关系抽取

联合推理的实体关系抽取中的典型方法是马尔可夫逻辑网 MLN[27],其最核心的思想即将马尔可夫网络与逻辑相结合,同时也是在 OIE 中融入推理的一种重要实体关系抽取模型。基于该模型,Liu 等[28] 提出一种无监督学习模型 StatSnowball,不同于传统的 OIE,该方法可自动产生或选择样例生成提取器。在 StatSnowball 的基础上,杨博等[22] 和 Liu 等[28] 提出一种实体识别与关系抽取相结合的模型 EntSum,该模型主要由扩展的 CRF 命名实体识别模块与基于 StatSnowball 的关系抽取模块构成,在保证准确率的同时也提高了召回率。

(3)有监督的实体关系抽取

传统的有监督的实体关系抽取模型主要基于统计方法,包括特征工程方法[29-31]、核函数方法[32-34]、图模型方法[35-37]等。有监督的实体关系抽取模型虽然取得有目共睹的效果,但是这些方法大多依赖于大量的标注数据,而取得大规模标注数据需要极高代价的人力和物力。为取得大规模的标注数据用于关系抽取模型的训练,Mintz 等[38] 提出远程监督模型,用于自动标注训练数据。远程监督模型基于一个强假设条件来标注数据,即假设在一个小型的知识图谱中,两个实体之间存在某种关系 R,那么远程监督模型认为,现实世界中只要这两个实体同时出现在一个句子中,则两个实体间的关系就一定是 R。

远程监督的强假设条件不可避免地造成数据的错误标注问题。为了解决远程监督数据集的噪声问题,Bunescu 等[39] 将弱监督学习与多实例学习相结合,并将其扩展到关系抽取上。Riedel 等[40] 将远程监督的关系抽取问题形式化为多实例单标签问题。但这些方法还是基于传统的自然语言处理工具生成特征,其效果仍然受到特征提取错误的附加影响。

2.2.2 关系抽取学习模型

伴随着深度学习的快速发展,基于深度学习的关系抽取模型得到广泛关注和研究。有监督的关系抽取深度学习模型的研究,主要受计算机视觉任务中各种卷积神经网络的启发,诸多基于变种卷积神经网络的关系抽取模型相继被提出[41-43],研究人员同时也关注了深度学习应用与消除远程监督模型噪声数据的研究。

(1)句子级别的关系抽取深度学习模型

深度学习中注意力(Attention)机制可以使神经网络具备专注于其输入(或特征)子集的能力:选择特定的输入。Lin 等[44] 将注意力机制应用于远程监督等关系抽取任务中,提出一种基于句子级别选择性注意力机制等神经网络模型(ATT)。ATT 为每个句子计算注意力得分,并以该得分衡量句子在表达实体间关于关系的信息量。假设 $S(h,t)$ 为包含 n 个句子的集合,其中每一个句子中都含有实体对 (h,t),即 $S_{(h,t)}=\{s^1_{(h,t)},\cdots,s^n_{(h,t)}\}$。为便于计算句子 $s^i_{(h,t)}$ 中有关实体对 (h,t) 之间关系的信息量,ATT 将句子集合 $S(h,t)$ 表达为句子向量 $s^i_{(h,t)}$ 的加权平均,即

$$S_{(h,t)}=\sum_i \alpha_i s^i_{(h,t)},$$

其中的 α_i 在 ATT 模型中由选择性注意力机制定义。

ATT 模型通过从远程监督的噪声数据识别有效实例,减轻远程监督中错误标注带来的影响。但这种仅对每一类关系使用单独模型来处理噪声数据的方法,忽略了实体关系间丰富的关联信息,而这些关联信息对关系抽取具有重要意义。

Yang 等[45] 在 ATT 模型基础上引进关系的层次信息,提出层次注意力模型的关系自动抽取模型(HATT)。与 ATT 模型相比较,HATT 模型引入关系内在的层次结构,并规定底层关系具有特定的关系特征(如宁明花山景点),而高层关系则为泛化概念,较为笼统和普遍(如地域)。HATT 模型在关系层次上逐层计算包含同样实体对的句子权重,因而在不同层次的关系上具有不同粒度的信息选择与噪声处理能力。与传统的去噪模型相比,注意力机制通过学习句子关系信息量的权重,能够动态降低噪声句子的影响,有效提升关系抽取的性能。而层次注意力机制能够更好地利用关系间丰富的联系,进一步提升关系抽取模型的整体效果。

(2)多语言关系抽取

在互联网时代,承载信息的自由文本资源来源丰富,实体间的关系不仅存在于一种语言文本中,而且常常是多语言的。如同一个景区景点的介绍,常常存在多国语言版本,因此,不同语言文本之间,实体关系

具有潜在的互补性和一致性。Lin等[46]基于实体关系通常在各种语言中存在不同的表达模式这个事实,基于当前存在的单语言关系抽取方法,提出一个基于多语言交叉Attention机制实体关系抽取方法(MNRE),即针对不同语言中实体间关系的不同表达模式,设计相应的关系权重计算方法,可以充分利用不同语言中的关系模式,从而增强关系模式的学习。Wang等[47]针对MNRE模型不能够很好地捕捉不同语言间关系模式的一致性和多样性的问题,将对抗网络引入多语言关系抽取模型学习中,提出基于对抗训练的多语言神经关系抽取模型(AMNRE)。该模型将不同语言文本映射到相应的特有语言空间进行语言特性的提取,并采用对抗机制以保证能够有效抽取出语言一致性特征,从而解决关系模式一致性和多样性的学习问题。

3 知识图谱的应用及发展趋势

3.1 知识图谱的应用

3.1.1 基于知识图谱的对话系统

对话系统,传统上分为目标导向Agent和闲聊Agent两种。所谓目标导向Agent,即帮助用户去完成某项任务,例如帮忙预定餐桌或安排代驾等。闲聊Agent即智能对话,具有互动性、娱乐性和话题性。

近年来涌现出太多关于深度神经网络构建端到端(不需要特定通道)对话系统的工作。然而,现在越来越明显的趋势是,无论在目标导向Agent还是闲聊Agent中都需要拥有一些知识,前者需要领域知识,后者需要常识知识。知识图谱将提高Agent对话的可解释性。在实际应用中,一个任务型对话系统一般会涉及多个领域的知识,分别对应不同领域的知识库。这些知识库往往有着不同的来源。这些不同的知识源往往由不同的技术人员进行维护,且具有异构的分布和属性。这会导致知识库很难甚至无法直接应用于任务型对话系统中。所以,需要借助于知识融合模型,将这些异源的知识库融合为一个知识库,然后再将融合后的知识库应用于任务型对话系统中[48]。

3.1.2 知识图谱情报案例分析

漆桂林实验团队的前沿研究现状是对知识图谱在情报案例中的分析[49]。该团队为推动知识图谱发展,强调中文开放知识图谱联盟OpenKG发展的必要性。该联盟旨在推动中文知识图谱的开放与互联,推动知识图谱技术在中国的普及与应用,为中国人工智能的发展以及创新创业做出贡献。

该实验团队举例的情报案例分析包括股票投研情报分析、公安情报分析、反欺诈情报分析。对于股票投研情报分析,主要是从各种股票相关的半结构化及非结构化数据中批量自动抽取股票相关人员的信息,构建公司知识图谱,为投资研究人员做更深层次的分析与决策提供可视化的分析依据。对于公安情报分析,主要是构建融合企业与个人信息的资金关系知识图谱,通过分析资金流向,为公安人员判断是否为非法集资提供分析依据。对于反欺诈情报分析,主要是通过融合来自不同数据源的信息构成知识图谱,同时引入领域专家建立业务专家规则,利用构建的知识图谱分析识别可能潜在的诈骗风险[49]。

3.1.3 基于知识图谱的产品案例

除上述将知识图谱技术应用于辅助特定业务分析之外,国内有关知识图谱技术应用的成熟智能产品也在市场中不断涌现。其中科大讯飞、云知声等企业的基于知识图谱的智慧产品在市场中表现尤为活跃。

科大讯飞基于学生学情、学科教学内容等数据,构建教育领域知识图谱,借助教育知识图谱,帮助老师预设教学重点,打造课前、课中、课后以生为本的教学闭环场景,构建实时线上互动的智慧课堂,显著提升教学效率,实现精准教学。同时,知识点图谱与自适应推荐引擎可为学生构建线上线下可视化的自主学习场景,支持学生按图索骥式学习,从而实现因材施教,提升学习效率,达到自主学习的目的。而通过构建基于司法案件宗卷数据的司法领域知识图谱,科大讯飞实现了智慧司法的产品研发和实际应用。在公安、检察机关、法院以及政法业务等领域,提供多种一体化智慧建设方案,实现案件宗卷语义理解,规范司法管理流程,打通公检法司数据流程,服务各级机关执法办案,确保办案证据标准符合法定定案标准。

与科大讯飞相比,云知声的AI能力始于智能语音处理技术,在知识感知、表达、理解、分析和决策等认知技术广泛部署,并朝着多模态人工智能系统方向发展。同样是智能教育方案,云知声专注于利用自然语言理解技术,构建自然语言语义及语音的关系知识库,并联合应用语音评测技术、云计算技术等,为用户提供智能化的语言学习产品后台服务。云知声在智能家居、智慧医疗等行业,将领域知识与语音识别、语音交互技术有效结合起来,有效解决现实领域内智能服务产品的技术实现,真正带给用户良好流畅的交流体验和卓越的应用价值。如云知声提供的智慧医疗

方案,能够实现智能语音交互的知识问答和病例查询,从而进行健康风险预测等患者病例分析,能够从真正意义上实现病例的精准录入。

3.2 知识图谱的发展趋势

知识图谱在未来的智能机器中将发挥大脑的作用,对自然语言处理、信息检索以及人工智能的发展将产生深远的影响。知识图谱关键技术及应用研究将会在很长一段时间成为大数据、人工智能的热门研究方向。未来的知识图谱关键技术及应用仍需针对以下 3 个方面展开深入研究。

第一,高质量知识的获取。如何在互联网大数据以及其他纷繁浩瀚的数据来源里面获取高质量的知识,是构建知识图谱的难题之一。目前在抽取知识的准确率、有效性和效率等方面都不尽如人意,影响知识图谱系统构建的有效性。在旅游文化知识图谱中,文化知识的来源主要为百科科普型网站、旅游网站以及相关的书籍等,各方面来源的知识汇聚到一起使得知识量非常庞大,出现冗余或者错误的知识比较普遍,因此如何构建知识图谱的本体成为难题,这时需要旅游相关的专家来进行实体的定义。如何定义实体并建立实体之间的关系,以及用什么方法把实体在知识中抽取出来,这些都是建立知识图谱非常关键的过程。同时为了保证知识的高质量,在实体抽取的过程中需要大量的人力资源进行校对修改。因此,如何有效地获取高质量的知识,应作为知识图谱的重要研究主题。

第二,知识的融合。从不同来源获取的知识可能存在大量的噪声或者冗余,不同语种中对同类型知识也可能存在不同的描述方式,使用什么方法把这些知识有效地融合到一起,以建立更大规模的知识图谱,是完成大数据智能的必经之路。在旅游文化知识图谱系统中,某个实体或者概念在知识库中可能存在不同的描述信息,在现实中也存在相同事物有多种不同叫法的情况。为了确保知识图谱系统的质量,以便于给用户提供准确的信息,需要把多个来源同一个实体或者概念的不同描述信息进行融合映射,这需要解决实体命名模糊、数据格式不一致等问题。同时,因为知识量过于庞大,如何准确有效地把知识进行融合映射也是其中的难点之一。

第三,民族文化知识图谱构建及应用。知识图谱在智能产品中的应用案例分析,揭示了知识图谱是从大数据到人工智能实现的技术桥梁这一事实。知识图谱的构建为领域数据分析提供具有可解释性的推理过程,因而基于知识图谱的解决方案更符合人类认知的规律。目前已有基于教育、司法、医疗、交通等诸多领域知识图谱的成功应用案例公开报道,但是关于民族文化知识图谱的研究还非常匮乏。民族文化知识是世界知识的一个子集,构建民族文化知识图谱对于丰富世界知识图谱具有重要意义。

民族文化知识图谱的构建能够应用于数字文化旅游中的特色文化推荐、同源文化演变分析,以及文化跨媒体数据有效管理和检索等场景,具有重要的应用价值。民族文化知识图谱的构建工作同样在于知识库中实体类型、属性、实体关系类型和属性的定义,以及海量知识数据的标注。民族文化知识本身的多样性和丰富性,使得文化知识的实体类型、关系类型及其属性的定义存在较大难度,需要通过阅读大量文献,并与领域专家共同探讨进行约定。同时,承载文化知识的媒体数据繁多,如何有效降低数据标注的人工成本,研究文化知识数据自动标注方法,是当前知识图谱领域的一个研究热点。

笔者所在研究团队针对上述民族文化知识图谱构建中的两个主要工作——民族文化知识实体、实体关系及其属性的定义和文化知识数据标注展开深入研究。目前对旅游行业的景观文化、民族服饰等不同类型的文化实体、实体关系进行定义,构建关于广西旅游景区文化知识、壮族服饰和瑶族服饰等几类小型知识图谱,并基于广西民族文化旅游知识图谱,设计相应的问答系统。后续将针对民族文化知识数据深入研究其知识表达和知识自动获取关键技术,提高民族文化知识图谱构建效率,拓展民族文化知识图谱的应用场景,使其在地方经济建设,特别是旅游产业建设中发挥重要作用。

参考文献

[1] Wikipedia. Knowledge graph [EB/OL]. (2016-05-09). https://en.wikipedia.org/wiki/Knowledge_Graph.

[2] 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.

[3] 刘峤,李杨,杨段宏,等. 知识图谱构建技术综述[J]. 计算机研究与发展,2016,53(3):582-600.

[4] NAVIGLI R,PONZETTO S P. BabelNet:The automatic construction,evaluation and application of a wide-coverage multilingual semantic network [J]. Artificial Intelligence,2012,193(1):217-250.

[5] HOFFART J,SUCHANEK F M,BERBERICH K,et al. YAGO2:A spatially and temporally enhanced knowl-

edge base from Wikipedia [J]. Artificial Intelligence, 2013,194:28-61.

[6] VRANDECIC D,KROTZSCH M. WikiData:A free collaborative knowledgebase [J]. Communications of the ACM,2014,57(10):78-85.

[7] BOLLACKER K D,EVANS C,PARITOSH P,et al. Freebase:A collaboratively created graph database for structuring human knowledge [C]. Proceedings of the ACM SIGMOD International Conference on Management of Data. Vancouver BC,Canada,2008:1247-1250.

[8] AUER S,BIZER C,KOBILAROV G,et al. DBpedia:A nucleus for a web of open data [C]. 6th International Semantic Web Conference,2nd Asian Semantic Web Conference,ISWC 2007+ASWC 2007. Busan,Korea,2007:722-735.

[9] MILLER G A. WordNet: A lexical database for English [J]. Communications of the ACM,1995,38(11):39-41.

[10] 鄢珞青.知识库的知识表达方式探讨[J].情报杂志,2003(4):63-64.

[11] 王知津,王璇,马婧.论知识组织的十大原则[J].国家图书馆学刊,2012,21(4):3-11.

[12] 王军,张丽.网络知识组织系统的研究现状和发展趋势[J].中国图书馆学报,2008,34(173):65-69.

[13] HODGE G. Next generation knowledge organization systems: Integration challenges and strategies [C]. ACM/IEEE Joint Conference on Digital Libraries. New York:ACM,2005.

[14] SOCHER R,CHEN D,MANNING C D,et al. Reasoning with neural tensor networks for knowledge base completion [C]. NIPS'13:Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1,2013:926-934.

[15] NICKEL M,TRESP V,KRIEGEL H. A three-way model for collective learning on multi-relational data [C]. Proceedings of the 28th International Conference on Machine Learning,ICML 2011. Washington,2011:809-816.

[16] BORDES A,USUNIER N,GARCIA-DURAN A,et al. Translating embeddings for modeling multi-relational data [C]. NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2,2013:2787-2795.

[17] TANG Y,HUANG J,WANG G,et al. Orthogonal relation transforms with graph context modeling for knowledge graph embedding [C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,2020:2713-2722. DOI:10.18653/v1/2020.acl-main.241.

[18] NGUYEN D Q,NGUYEN T D,PHUNG D. A relational memory-based embedding model for triple classification and search personalization [C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 3429-3435. DOI: 10.18653/v1/2020.acl-main.313.

[19] ZHANG W,PAUDEL B,ZHANG W,et al. Interaction embeddings for prediction and explanation in knowledge graphs [C]. WSDM'19: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining,2019:96-104.

[20] LIN Y K,LIU Z Y,SUN M S,et al. Modeling relation paths for representation learning of knowledge bases [C]. Proceedings of NLP, 2015: 705-714. DOI: 10.18653/v1/D15-1082.

[21] LIN Y K,LIU Z Y,SUN M S,et al. Learning entity and relation embeddings for knowledge graph completion [C]. AAAI'15: Proceedings of the Twenty-ninth AAAI Conference on Artificial Intelligence,2015:2181-2187.

[22] 杨博,蔡东风,杨华.开放式信息抽取研究进展[J].中文信息学报,2014,28(4):1-11,36.

[23] ETZIONI O,CAFARELLA M,DOWNEY D,et al. Unsupervised named-entity extraction from the web: An experimental study [J]. Artificial Intelligence, 2005,165(1):91-134.

[24] WU F,WELD D S. Open information extraction using Wikipedia [C]. ACL'10: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Sweden:ACL,2010:118-127.

[25] DOMINGOS P,LOWD D. Markov logic: An interface layer for artificial intelligence [M]. San Rafael CA: Morgan & Claypool,2009.

[26] SUCHANEK F M,KASNECI G,WEIKUM G. YAGO: A large ontology from wikipedia and wordnet [J]. Journal of Web Semantics,2008,6(3):203-217.

[27] ZHU J,NIE Z Q,LIU X J,et al. StatSnowball:A statistical approach to extracting entity relationships [C]. WWW'09:Proceedings of the 18th International Conference on World Wide Web. Switzerland,2009:101-110.

[28] LIU X J,YU N H. People summarization by combining named entity recognition and relation extraction [J]. Journal of Convergence Information Technology, 2010,5(10):233-241.

[29] ZHOU G D,SU J,ZHANG J,et al. Exploring various

knowledge in relation extraction [C]. ACL'05:Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics,2005:427-434.

[30] JIANG J,ZHAI C X. A systematic exploration of the feature space for relation extraction [C]. Human language technologies 2007:The conference of the North American Chapter of the Association for Computational Linguistics. New York,2007:113-120.

[31] KAMBHATLA N. Combininglexical,syntactic,and semantic features with maximum entropy models for extracting relations [C]. ACLdemo'04:Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions,2004:22.

[32] BUNESCU R C,MOONEY R J. A shortest path dependency kernel for relation extraction [C]. HLT'05:Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing,2005:724-731.

[33] MOONEY R J,BUNESCU R C. Subsequence kernels for relation extraction [C]. NIPS'05:Proceedings of the 18th International Conference on Neural Information Processing Systems,2005:171-178.

[34] ZHAO S B,GRISHMAN R. Extracting relations with integrated information using kernel methods [C]. ACL'05:Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics,2005:419-426.

[35] ROTH D,YIH W T. Probabilistic reasoning for entity & relation recognition [C]. COLING 2002:The 19th International Conference on Computational Linguistics,2002:1-7.

[36] SARAWAGI S,COHEN W W. Semi-Markov conditional random fields for information extraction [C]. NIPS'04:Proceedings of the 17th International Conference on Neural Information Processing Systems,2005:1185-1192.

[37] YU X F,LAM W. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach [C]. COLING'10:Proceedings of the 23rd International Conference on Computational Linguistics: Posters,2010:1399-1407.

[38] MINTZ M,BILLS S,SNOW R,et al. Distant supervision for relation extraction without labeled data [C]. ACL'09:Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP:Volume 2,2009:1003-1011.

[39] BUNESCU R,MOONEY R. Learning to extract relations from the web using minimal supervision [C]. ACL 2007,Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,2007:576-583.

[40] RIEDEL S,YAO L M,MCCALLUM A. Modeling relations and their mentions without labeled text [C]. ECML PKDD'10: Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases:Part Ⅲ,2010:148-163.

[41] HUANG Y Y,WANG W Y. Deep residual learning for weakly-supervised relation extraction [C]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,2017:1803-1807.

[42] LIU C Y,SUN W B,CHAO W H,et al. Convolution neural netword for relation extraction [C]. International Conference on Advanced Data Mining and Applications,2013:231-242.

[43] SANTOS C N D,XIANG B,ZHOU B W. Classifying relations by ranking with convolutional neural networks [C]. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,2015:626-634.

[44] LIN Y K,SHEN S Q,LIU Z Y,et al. Neural relation extraction with selective attention over instances [C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,2016:2124-2133.

[45] YANG Z C,YANG D,DYER C,et al. Hierarchical attention networks for document classification [C]. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics,2016. DOI:10.18653/v1/N16-1174.

[46] LIN Y K,LIU Z Y,SUN M S. Neural relation extraction with multi-lingual attention [C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,2017:34-43.

[47] WANG X Z,HAN X,LIN Y K,et al. Adversarial multi-lingual neural relation extraction [C]. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe,New Mexico,USA:Association for Computational Linguistics,2018:1156-1166.

[48] 郝增光. 基于张量分解的知识图谱融合研究及其在对话中的应用[D]. 青岛:山东大学,2020.

[49] 漆桂林,高桓,吴天星. 知识图谱研究进展[J]. 情报工程,2017,3(1):4-25.

Progress and Prospect of Knowledge Graph Technology

QIN Xiao[1],LIAO Zhaoqi[1],SHI Yu[1],YUAN Changʹan[2]
(1. Nanning Normal University,Nanning,Guangxi,530299,China;2. Guangxi Academy of Sciences,Nanning,Guangxi,530007,China)

Abstract:With the development of big data, the key technology and application research of knowledge graph have become one of the most popular research fields of artificial intelligence. Based on the definition,architecture and common knowledge base of knowledge graph, this article summarizes and reviews the knowledge representation and automatic knowledge acquisition technology constructed by knowledge graph,and discusses its research points and development trends. Then the common application scenarios of knowledge graph technology are introduced. Finally, combined with the research of the team, we discuss the trends of knowledge graph.

Key words: natural language processing,knowledge graph,knowledge representation,knowledge extraction,model improvement

责任编辑:陆 雁