

面向知识库问答的问句语义解析研究综述

仇韞琦^{1,2}, 王元卓^{1,3}, 白 龙^{2,4}, 尹芷仪¹, 沈华伟^{1,2}, 白 硕¹

(1. 中国科学院计算技术研究所数据智能系统研究中心, 北京 100190; 2. 中国科学院大学计算机科学与技术学院, 北京 101408;
3. 中科大数据研究院, 河南郑州 450046; 4. 中国科学院计算技术研究所网络数据科学与技术重点实验室, 北京 100190)

摘要: 知识库问答(Knowledge Base Question Answering, KBQA)借助知识库中精度高、关联性强的结构化知识,为给定的复杂事实型问句提供准确、简短的答案。语义解析是知识库问答的主流方法之一,该类方法在给定的问句语义表征形式下,将非结构化的问句映射为结构化的语义表征,再将其改写为知识库查询获取答案。目前,面向知识库问答的语义解析方法主要面临三个挑战:首先是如何选择合适的语义表征形式以表达问句的语义,然后是如何解析问句的复杂语义并输出相应的语义表征,最后是如何应对特定领域中数据标注成本高昂、高质量数据匮乏的问题。本文从上述挑战出发,分析了知识库问答中常用的语义表征的特点与不足,然后梳理现有方法并总结分析其如何应对问句的复杂语义,接着介绍了当前方法在标注数据匮乏的低资源场景下的尝试,最后展望并讨论了面向知识库问答的语义解析的未来发展方向。

关键词: 知识库; 问答; 语义表征; 语义解析; 低资源

中图分类号: TP391; TP182

文献标识码: A

文章编号: 0372-2112(2022)09-2242-23

电子学报 URL: <http://www.ejournal.org.cn>

DOI: 10.12263/DZXB.20220212

A Survey of Question Semantic Parsing for Knowledge Base Question Answering

QIU Yun-qi^{1,2}, WANG Yuan-zhuo^{1,3}, BAI Long^{2,4}, YIN Zhi-yi¹, SHEN Hua-wei^{1,2}, BAI Shuo¹

(1. *Research Center for Data Intelligence Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;*
2. *School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China;*
3. *Big Data Academy, Zhongke, Zhengzhou, Henan 450046, China;*
4. *CAS Key Laboratory of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China*)

Abstract: Knowledge base question answering(KBQA) provides accurate and short answers to complex factoid questions with the help of high-precision and highly relevant structured knowledge in the knowledge base(KB). Semantic parsing has become one of the mainstream methods of KBQA. Under the given form of question meaning representation, this kind of method maps unstructured questions into structured meaning representations, and then rewrites them as KB queries to obtain answers. At present, semantic parsing for KBQA mainly faces three challenges: first how to choose a suitable meaning representation form to express the semantics of questions, then how to parse the complex semantics of questions and output the corresponding meaning representations, and finally how to deal with the high cost of labeling datasets and the lack of annotated data in specific domains. Starting from the above challenges, this paper first analyzed the characteristics and shortcomings of meaning representations commonly used in KBQA and then combed out how existing methods deal with the complex semantics of questions. After that, this paper introduced the current attempts in low-resource scenarios and finally discussed the future directions of semantic parsing for KBQA.

Key words: knowledge base; question answering; meaning representation; semantic parsing; low resource

1 引言

随着人工智能技术的快速发展,人机交互的现实应用日益广泛.知识库问答(Knowledge Base Question Answering,KBQA)作为人机交互的重要支撑,成为工业界与学术界共同关注的研究热点之一.相较于基于非结构化文本数据的问答(如文档检索、阅读理解等),知识库问答能借助知识库中精度高、关联性强的结构化知识,为给定的复杂事实型问句提供精确的语义理解或解析,并在知识库中查询推理来得到准确、简短的答案.如问句“法国在第一次世界大战时的总统是谁?”中需要对法国历任总统的就任与卸任时间分别同第一次世界大战的开始与结束时间进行数值比较,从而得到答案“雷蒙·普恩加莱”.

目前主要有两类知识库问答方法:一类基于语义解析(semantic parsing)^[1-3],将自然语言问句按照特定文法(grammar)或者处理步骤映射为结构化的语义表征(meaning representation),再将其改写为知识库查询从而获得最终答案;另一类基于信息检索(information retrieval)^[4-6],首先通过实体链接,识别出问句中关心的实体作为主题实体(topic entity),并获取以主题实体为中心的知识库子图,将子图中的每个节点都视作候选答案,再学习问句和候选答案的向量表示,最后通过排序模型或分类模型对候选答案进行筛选,从而获得最终答案.基于语义解析的方法以符号化表征表示问句的语义,基于信息检索的方法以稠密向量表示问句的语义.相比基于信息检索的方法,语义解析的方法能够应对更多类型的问句,例如含有实体约束、类别约束、数值比较、数值排序等的问句,从而有更好的应用前景.

早期的知识库问答系统以基于符号和规则的专家系统为主,具有代表性的如BASEBALL^[7],将问句解析为规格列表(specification list),从而回答美国棒球比赛相关的问句.但是专家系统的核心是强依赖领域专家的规则,这导致其应用领域狭窄,可回答问句类型受限等问题.

近年来,随着自然语言处理和机器学习技术的进步,这些问题逐步得到缓解,面向知识库问答的语义解析技术也涌现出了大量高水平研究论文,在诸多评测数据集上显示出良好的性能.尽管如此,现有方法依旧面临了如下挑战.

(1) 如何选择合适的语义表征形式以表达问句的语义.语义解析的第一步是为问句选择具体的语义表征形式,现有方法采用了多种不同形式的语义表征.有些语义表征的表达能力强,能覆盖多数类型的自然语言问句,但结构复杂,构造困难;有些语义表征的结构紧凑,构造相对容易,但表达能力有所欠缺.因此,采用

合适的语义表征才能平衡好表达完备性与构造便捷性,在覆盖更多类型的问句的同时,降低模型构造该语义表征的难度.

(2) 如何解析问句的复杂语义并输出相应的语义表征.复杂问句的语义丰富,包含了知识库中的多个关系,实体或类别约束,比较或排序等函数算子等.同时,相同的语义成分按照不同的顺序或结构组合,表达的语义则大不相同.因此,准确地识别出问句包含的语义成分,并将不同的语义成分按照特定顺序和结构组合,才能准确地反映问句的真实语义.

(3) 如何应对特定领域中数据标注成本高昂、高质量数据匮乏的问题.现有方法主要是数据驱动,但是给问句标注相应的语义表征或最终答案需要标注人员熟悉知识库的本体(ontology)和实例(instance),高昂的标注成本导致特定领域仅有少量标注数据,甚至没有标注数据,从而让问答模型训练困难.因此,在标注样本匮乏的情况下实现模型的有效训练与泛化,是知识库问答落地的必要条件.

现有的部分知识库问答综述^[8-11]对语义表征的分析较少,对语义解析方法的介绍较笼统,未能体现不同解析方式的特点与适用场景.为了更好地帮助领域内研究者,本文通过梳理面向知识库问答的语义解析研究,总结和分析当前方法应对上述挑战的途径和发展趋势,为实现高质量的知识库问答系统提供可借鉴的思路.

2 背景知识

2.1 知识库问答

知识库是用于存储结构化事实信息的计算机系统.早期的知识库以关系型数据库(Relational Database,RDB)为主,将现实世界的实体、实体本身的属性、实体之间的联系以表格形式的关系模型来表示.近年来,以语义网络(semantic web)和知识图谱(knowledge graph)为根基的图数据库(Graph DataBase,GDB)成为知识库的研究主流,其中,实体、类别、属性值作为节点,分别以特定的关系或属性边相连.图1给出的图数据库实例展示了中国部分省份的相关信息,如三元组“(Jiangsu, capital, Nanjing)”表明江苏的省会是南京.典型的知识库包括 WordNet^[12], DBpedia^[13], Freebase^[14], Wikidata^[15]等.

基于知识库中的结构化信息,知识库问答系统能针对用户给定的自然语言问句,给出简洁、准确的回答.目前主要有基于语义解析和基于信息检索这两类知识库问答方法.以问句“Which provinces border Jiangsu?”为例,基于语义解析的方法将该问句映射为结构化的语义表征,如 λ 演算(lambda calculus)表达

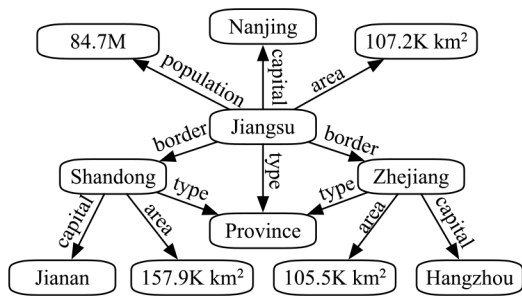


图1 图数据库示意图

式^[16,17] $\lambda x.province(x) \wedge border(x, Jiangsu)$, 再将其改写为知识库查询得到答案, 如 SPARQL 查询^[18]`SELECT ?x WHERE {Jiangsu border ?x . ?x type province . }`; 基于信息检索的方法将图1所示的以主题实体(即 Jiangsu)为中心的知识库子图中的每个节点都视作候选答案, 学习问句和候选答案的向量表示, 再通过排序模型计算二者之间的相似度, 或者通过分类模型对候选答案进行分类, 从而得到答案。

早期的基于信息检索的方法^[4]以词袋(bag-of-words)模型表示问句, 以候选答案及其关联子图内的实体和关系作为答案特征, 并分别通过两个线性映射得到问句和候选答案的向量表示, 再以向量点积作为相似度得分。随着深度学习的发展, 基于信息检索的方法广泛使用神经网络模型学习问句和候选答案的向量表示^[5,19-23]。例如, Dong 等人^[5]提出多列卷积神经网络, 从主题实体到候选答案的路径、候选答案的邻居子图、候选答案的类别这3个方面考虑候选答案的特征, 并为问句同样学习了3个向量与之对应。Hao 等人^[20]通过长短时记忆网络^[24](Long Short-Term Memory, LSTM)获得问句的表示, 并利用交叉注意力(cross-attention)机制建模候选答案与问句之间的关联性。Sun 等人^[22,23]结合外部文本(如维基百科)和知识库子图, 构建问句相关的子图, 并利用图神经网络在该子图上更新候选答案的向量表示。部分工作^[6,25-28]引入多跳推理(multi-hop reasoning)机制, 迭代更新问句或候选答案的向量表示, 从而更好地应对多关系问句。例如, Miller 等人^[6]提出键值记忆网络(key-value memory network), 将问句关联的三元组存入键值记忆网络中, 迭代计算存储的三元组与问句的关联性, 并利用三元组信息更新问句的向量表示。

由此可见, 基于信息检索的方法无需研究者了解问句语义表征和知识库查询语言的细节, 较为便捷, 但是可回答问句的类型受限, 难以回答带有实体约束或函数算子的问句。如问句“Which is the largest province that borders Jiangsu?”需要对江苏省相邻省份的面积进行排序操作。而基于语义解析的方法生成的语义表征可以包含问句中的约束和算子, 并利用知识库本身的查询引擎获得准确答案。本文主要介绍面向知识库问

答的语义解析方法, 对信息检索方法不做进一步展开。

2.2 语义解析

语义解析是将自然语言句子映射为结构化语义表征, 解析得到的语义表征通常可以在某个环境中被执行并获得相应的结果, 如知识库问答中查询知识库获得问句的答案。除知识库问答外, 语义解析技术在其他多种人机交互场景中也得到广泛应用, 如机器人引导(robot instruction)、任务导向型对话(task-oriented dialogue)等。相较于这些场景, 面向知识库问答的语义解析要求输出的语义表征中所有语义成分与知识库中的信息对应。图2给出了面向知识库问答的语义解析的一般流程: 首先为问句选择具体的语义表征形式; 然后根据该语义表征形式设计相应的语义解析方法, 从而完成自然语言问句到其对应的语义表征的映射; 最后将生成的问句语义表征改写为知识库查询, 从知识库中获取问句的最终答案。

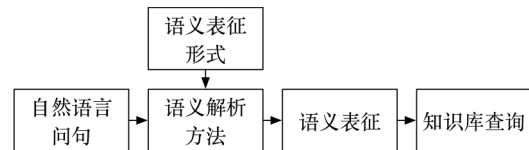
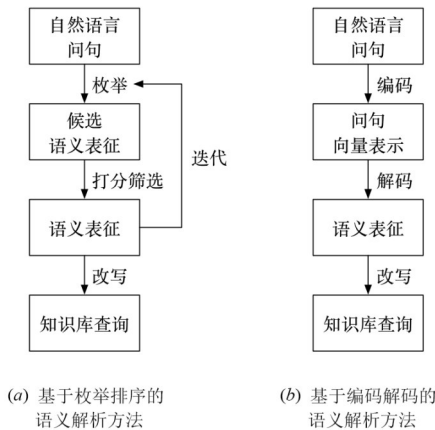


图2 面向知识库问答的语义解析的一般流程

面向知识库问答的语义解析方法按照解析方式可以分为两类: 枚举排序和编码解码。如图3所示, 基于枚举排序的方法是通过特定文法或者人工设计的处理步骤枚举所有候选语义表征, 再通过一个打分函数对枚举的所有候选排序, 筛选高分候选输出。考虑到候选的搜索空间巨大, 一部分工作采用迭代式枚举排序, 每轮迭代基于上一轮筛选得到的语义表征子式进行扩展, 再进入下一轮迭代。最终输出的语义表征被改写为相应的知识库查询。相比枚举排序这类方法依赖特定文法、人工规则或者模板, 编码解码的方法采用基于编码器-解码器(encoder-decoder)架构的模型直接学习自然语言到语义表征的映射, 首先利用编码器得到问句的向量表示, 再利用解码器根据问句向量输出符号化的语义表征。

在语义表征到知识库查询的改写阶段, 语义表征中的成分需要与知识库对应, 即语义表征中的实体需要被链指到知识库中的对应实体, 语义表征中实体间的关系需要被映射到知识库中对应实体间的关系。一部分工作在解析得到语义表征时已经完成与特定知识库的对应, 因此可直接将其改写为知识库查询, 这类语义表征被称为有根基的语义表征(grounded meaning representation)。相比直接生成有根基的语义表征, 一部分工作认为先将问句解析为知识库无关的通用表征能够实现解析模型在不同知识库上的迁移与复用, 因为



(a) 基于枚举排序的语义解析方法 (b) 基于编码解码的语义解析方法
图3 面向知识库问答的语义解析方法分类

通用表征的实体和关系没有与特定知识库对应, 所以称之为无根基的语义表征 (ungrounded meaning representation). 无根基的语义表征在改写为知识库查询时, 往往通过结构修改、实体链指及关系映射等方式先获

得有根基的语义表征, 再将其改写为知识库查询. 不同解析方式的实现方法和优缺点将在第 4 节和第 5 节进一步展开介绍.

2.3 常用数据集及评价指标

在数十年的发展中, 知识库问答领域涌现了大量的数据集. 早期以关系型数据库为基准知识库的阶段, 被广泛采用的问答数据集包括面向美国地理领域的 GEO880 (<https://www.cs.utexas.edu/users/ml/nldata/geoquery.html>)、面向工作领域的 JOBS640 (<https://www.cs.utexas.edu/users/ml/nldata/jobquery.html>)、面向旅行计划领域的 ATIS^[29] 等. 近年来, 以图数据库为基准知识库的问答数据集成为知识库问答领域的主流. 表 1 列举了近年来被广泛采用的以图数据库为基准的英文知识库问答数据集^[2, 25, 30-44], 分别介绍了各数据集名称、采用的基准知识库、数据集的样例规模、是否含有需要多个三元组查询或函数约束的复杂问句、每个问句的具体标注形式、样例中问句的来源.

表 1 以图数据库为基准知识库的部分英文知识库问答数据集

数据集名称	基准知识库	样例规模	含有复杂问句	问句标注形式	问句来源
Free917 ^[30]	Freebase	917	有	λ 演算	人工撰写
WebQuestions ^[2]	Freebase	5 810	有	最终答案	Google Suggest API
SimpleQuestions ^[31]	Freebase	108 442	无	三元组	人工撰写
WebQuestionsSP ^[32]	Freebase	4 737	有	SPARQL	其他数据集
ComplexQuestions ^[33]	Freebase	2 100	有	最终答案	其他数据集、人工撰写
GraphQuestions ^[36]	Freebase	5 166	有	SPARQL	模板填充+人工复述
ComplexQuestions ^[34]	Freebase	150	有	最终答案	WikiAnswers
30M Factoid Questions ^[37]	Freebase	30 000 000	无	三元组	神经模型生成
PathQuestion ^[25]	Freebase	9 731	有	三元组	模板填充
CFQ ^[38]	Freebase	239 357	有	SPARQL	规则构造
GrailQA ^[39]	Freebase	64 331	有	SPARQL	模板填充+人工复述
ComplexWebQuestions ^[40]	Freebase	34 689	有	SPARQL	模板填充+人工复述
QALD series(http://qald.aksw.org/)	DBpedia	50 ~ 558	有	SPARQL	人工撰写
LC-QuAD ^[41]	DBpedia	5 000	有	SPARQL	模板填充+人工复述
LC-QuAD 2.0 ^[42]	DBpedia & Wikidata	30 000	有	SPARQL	模板填充+人工复述
CSQA ^[35]	Wikidata	1 600 000	有	最终答案	人工撰写
KQA Pro ^[43]	Wikidata	117 970	有	SPARQL	模板填充
MetaQA ^[44]	WikiMovies	407 513	有	最终答案	模板填充

表 1 列举了现有知识库问答数据集主要采用的基准知识库, 分别为 Freebase, DBpedia, Wikidata 和 Wikimovies^[6], 均为图数据库. 从构建方式上, Freebase, DBpedia 和 Wikidata 均为众包模式构建, 即允许用户创建、修改、查询, 同时也采用信息抽取模型进行自动扩充 (Freebase 已于 2015 年停止更新, 原数据迁移至 Wikidata); 而 Wikimovies 来源于已有电影数据库 Open Movie DataBase (OMDB). 从面向领域上, Freebase, DB-

pedia 和 Wikidata 为通用领域, 即包含多个特定领域的事实信息; 而 Wikimovies 仅面向电影领域. 从数据规模上, Freebase 包含超过 4 600 万个实体和 5 323 种关系, 共计 31 亿个三元组; DBpedia 的最新发布版本 (2016-10 版) 包含 600 万个实体和 9 亿个三元组; 当前 Wikidata 包含超过 9 000 万个实体和超过 9 000 种关系, 共计超过 100 亿个三元组; Wikimovies 为小规模知识库, 包含 43 234 个实体和 9 种关系, 共计 13 万个三元组.

从表1可以看出,知识库问答数据集的构造方法主要分两种:(1)先有问句,再有答案,即通过人工撰写或者现有网站获取真实的自然语言问句,再通过人工标注获得问句的最终答案或问句相应的知识库查询^[2,30-35]; (2)先有答案,再有问句,即先从知识库中构造出语义表征,再通过预定义模板或规则将其转变为模板问句,考虑到这些模板问句表述单一且可能存在语法错误,部分工作再通过众包,将模板问句复述为具有相同含义且表述更自然的问句^[25,36-44]. 这两种构造方法中,前者需要标注人员熟悉基准知识库的本体概念与具体实例,在以大规模知识库为基准时,因实体和关系数量庞大,标注难度较高,且收集来的部分问句可能无法用知识库中的事实信息来回答;后者需要构造合适的语义表征,从而才能生成有实际意义的问句,同时需要标注人员提供尽可能多的表述方式,防止单一问句使得解析模型过拟合且缺乏泛化性.

除了构造方式,不同数据集的区别主要在样例规模和是否含有复杂问句上. 遵循现有工作^[9-11]的定义,知识库问答中的复杂问句指问句对应的知识库查询中含有多个知识库关系、实体或类别约束、比较或排序等函数算子,例如问句“When’s the last time the Mets won the World Series?”需要对纽约大都市棒球队夺冠的日期进行排序操作. 与之相对的,简单问句仅依赖单个三元组查询,例如问句“What city was Alex Golfis born in?”只需要(Alex Golfis, place_of_birth, ?a)组成的知识库查询即可回答. 为了更全面地检验知识库问答模型的能力,新近构造的数据集往往呈现出规模越来越大、复杂问句的种类越来越多、难度越来越高的趋势. 例如, Berant等人^[2]指出Free917人工撰写问句并标注的方式难以进一步扩展数据集规模,因此利用Google suggest API搜集事实性问句,并提交给众包平台直接标注问句的最终答案,从而得到数据集WebQuestions,数据规模超过Free917的5倍;Yih等人^[32]删去WebQuestions中无法利用Freebase回答的问句,并为每个问句提供了SPARQL标注,形成数据集WebQuestionsSP; Bao等人^[33]和Talmor等人^[40]指出WebQuestionsSP中复杂问句偏少,因此按照不同约束类别添加了新的复杂问句,其中Bao等人^[33]将复杂问句中的约束条件细化分类,并人工撰写一部分问句作为补充,形成数据集ComplexQuestions,而Talmor等人^[40]则是在WebQuestionsSP标注的SPARQL中加入额外约束条件,再通过模板填充与人工复述得到更加复杂的问句,形成数据集ComplexWebQuestions.

表1展示了知识库问答数据集为每个问句提供的具体标注形式. 一部分数据集仅标注了问句的最终答案,例如WebQuestions;一部分数据集采用知识库查询

语言SPARQL作为标注形式,并且标注的SPARQL可在对应的基准知识库中查询得到问句的最终答案,例如数据集WebQuestionsSP中,问句“When’s the last time the Mets won the World Series?”对应的SPARQL查询标注是SELECT ?x WHERE {New_York_Mets championships ?x . ?x end_date ?sk0 . } ORDER BY DESC(date-time(?sk0)) LIMIT 1. 少量数据集采用问句涉及的具体三元组作为标注形式,问句的最终答案为标注的三元组或三元组序列的尾实体,例如数据集SimpleQuestions中,问句“What city was Alex Golfis born in?”对应的三元组为(Alex Golfis, place_of_birth, Athens)(数据集的实际标注中,实体及关系均以其在知识库中的ID形式呈现,本文将ID替换为其具体名称,以便理解). 基于上述数据集衡量一个知识库问答模型的效果时,通常以问句的最终答案作为评价标准,并且要将模型生成的问句语义表征改写为基准知识库可读的查询,获取模型预测的答案. 而数据集Free917以 λ 演算作为问句的标注形式,并且以问答模型解析得到的语义表征与标注的 λ 演算相比较,两者完全对应上才视为解析正确,从而避免伪正例现象(即错误的语义表征得到正确的查询结果)导致评价失准.

以问句的最终答案作为评价标准衡量一个知识库问答模型的效果时,通常可以考虑以下指标:精确率(precision)、召回率(recall)、 F_1 值、准确率(accuracy)等. 对于一个给定问句 q ,其标准答案集合记为 A_q ,模型预测的答案集合记为 P_q ,则模型在给定问句上的精确率为模型预测答案集合中正确元素的占比,即

$$\text{precision} = \frac{|A_q \cap P_q|}{|P_q|}$$

而模型在给定问句上的召回率为问句标准答案集合中被模型预测出来的元素的占比,即

$$\text{recall} = \frac{|A_q \cap P_q|}{|A_q|}$$

同时考虑到精确率与召回率,模型在给定问句上的 F_1 值的计算方式为

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

对模型在数据集中所有问句上的 F_1 值求均值,可得到模型在当前数据集上的平均 F_1 值(average F_1),又称为宏 F_1 值(macro F_1). 而准确率是指模型在数据集中所有问句中完全回答正确的问句的占比,只有当 $A_q = P_q$ 时,问句 q 才被视为完全回答正确. 相较于平均 F_1 值,准确率这一指标更为严格,对模型要求更高.

3 语义表征

在设计具体的语义解析模型前,研究者需要先确

定采用何种语义表征作为模型的输出以表示问句的语义. 现有方法采用了多种不同形式的语义表征. 有些语义表征的表达能力强,能覆盖多数类型的自然语言问句,但结构复杂,构造困难;有些语义表征的结构紧凑,构造相对容易,但表达能力有所欠缺. 选取合适的语义表征对提升解析模型的效果起着至关重要的作用,因此需要平衡好语义表征的表达完备性与构造便

捷性,在覆盖更多类型的问句的同时,降低模型构造该语义表征的难度. 如表 2 所示,目前知识库问答中常用的语义表征可分为 3 类,分别是 λ 演算及其衍生^[16,17,45]、图状逻辑表达式^[3,46-53]、程序语言^[18,43,54-57]. 表 2 总结了这些常用语义表征的特点和相应的不足之处. 下文分别介绍这三类语义表征,并对其进行了小结与讨论.

表 2 知识库问答中常用的语义表征

语义表征分类	语义表征名称	主要特点
λ 演算及其衍生	λ 演算 ^[16,17]	图灵完备,具有很强的表达完备性,但表达较为繁琐,可读性差,构造较难
	λ -DCS ^[45]	在 λ 演算基础上隐去变量,表达更紧凑,构造难度比 λ 演算小
图状逻辑表达式	语义图 ^[46]	基于问句句法结构,构造过程中依赖特定文法或现有句法解析工具得到无根基的语义表征;再通过自定义
	语义查询图 ^[47]	操作、实体关系映射等映射到知识库子图,得到有根基的语义表征,构造较为便捷,可读性强,但图结构限制了其表达完备性
	抽象语义表征 ^[48]	
程序语言	查询图 ^[3]	直接从知识库子图中构造,不依赖特定文法或现有解析工具,构造更便捷,可读性强,但图结构限制了其表达完备性
	SPARQL ^[18]	具有较强的表达完备性,一般通过识别问句中的实体和关系完成 SPARQL 模板填充
	Prolog ^[54]	基于一阶谓词逻辑,具有较强的语义完备性,结构相对紧凑,但无法像图状逻辑表达式一样直接利用知识库的拓扑结构信息,在生成嵌套结构时易出错
	FunQL ^[56]	
	Lisp ^[55]	以前缀表达式列表为表达形式,可通过更新自定义函数来增强其表达能力,具有很强的表达完备性,但无法像图状逻辑表达式一样利用知识库的拓扑结构信息,在生成嵌套结构时易出错
	KoPL ^[43]	

3.1 λ 演算及其衍生

λ 演算^[16]源自数理逻辑,是一套以变量绑定和替换的规则来研究函数如何抽象化定义、函数如何被施用以及递归的形式系统. 在此基础上,有类型的 λ 演算^[17] (typed lambda calculus) 在语义解析工作中得到广泛使用,它主要包含 3 个基本类型,分别是实体、真值、实数,并在基本类型之上构建函数类型,如将实体映射为真值的函数. 本文后续部分均以 λ 演算代指有类型的 λ 演算. 以问句“Which is the largest province that borders Jiangsu?”为例,其 λ 演算表达式 $\arg \max (\lambda x . \text{province}(x) \wedge \text{border}(x, \text{Jiangsu}), \lambda x . \text{size}(x))$ 中, Jiangsu 就是实体; province 和 size 均是函数,前者将实体映射到真值上,表明这个实体的类别是否是省份,后者将实体映射到实数上,返回这个省份的面积. 总的来说,在知识库问答中常用的有类型的 λ 演算的组成元素包括:(1)常量,包括实体、实数、函数;(2)逻辑联结词,即合取(\wedge)、析取(\vee)、否定(\neg)、蕴含(\rightarrow);(3)量词,包括全称量词(\forall)、存在量词(\exists),如 $\exists x . \text{province}(x) \wedge \text{border}(x, \text{Jiangsu})$ 为真,当且仅当至少存在一个省份与江苏省相邻;(4) λ 表述 (lambda expression),接收参数并返回结果值,如 $\lambda x . \text{border}(x, \text{Jiangsu})$ 将实体映射为真值,表明输入的实体是否与江苏省相邻;(5)额外的量词,包括 count, $\arg \max$, $\arg \min$ 等.

λ 演算是图灵完备的,具有很强的语义表达能力,但其本身较为繁琐,增加了构造难度,因此在其基础上改进,使得表述更紧凑的语义表征同样得到广泛使用,如 λ -DCS^[45] (Lambda Dependency-based Compositional Semantics). λ -DCS 隐去了 λ 演算中的变量,以上段问句为例,其 λ -DCS 表达式为 $\arg \max (\text{type. province} \sqcap \text{border. Jiangsu, size})$. 具体来说, λ -DCS 主要包括 2 个组成成分,即一元逻辑表达式 (unary) 和二元逻辑表达式 (binary). 前者表示的是实体或实体集合,如 Jiangsu, 对应的是实体江苏省;后者表示的是实体对集合,如 border, 对应的是接壤的实体对集合,其在 λ 演算中的对应表达式为 $\lambda x . \lambda y . \text{border}(x, y)$. 不同的逻辑表达式可以通过连接 (join)、交叉 (intersection)、求并集 (union)、否定 (negation)、高阶算子 (high-order function) 等操作符组合,例如一元逻辑表达式 Jiangsu 和二元逻辑表达式 border 通过连接操作可以得到一元逻辑表达式 border. Jiangsu, 对应与江苏省相邻的实体集合,其在 λ 演算中对应 $\lambda x . \text{border}(x, \text{Jiangsu})$.

3.2 图状逻辑表达式

随着图数据库成为知识库的主流,以图为形式的逻辑表达式也成为知识库问答中的主要语义表征. 在图状逻辑表达式中,节点通常指实体、事件、数值等常量或变量,连边代表节点之间的语义关系. 相较于 λ 演

算和 λ -DCS,图状逻辑表达式有更强的可读性,同时与问句本身的句法结构或者图数据库的拓扑结构具有一定的相似性,降低了构造难度.从问句本身的句法结构改造而来的图状逻辑表达式有语义图^[46](semantic graph)、语义查询图^[47](semantic query graph)、抽象语义表征^[48](Abstract Meaning Representation, AMR)等.从图数据库的子图改造而来的图状逻辑表达式有查询图^[3](query graph).

图4为语义图示例^[49].其中图4(a)为无根基的语义图(ungrounded semantic graph),图4(b)为有根基的语义图(grounded semantic graph).前者是将问句通过组合范畴文法^[50,51](Combinatory Categorical Grammar, CCG)解析或者依存句法解析(dependency parsing)得到

的 λ 演算表达式改造成图的形式,节点有单词、实体、事件、类别、函数等,连边主要为依存关系,且图上的节点与连边没有与特定知识库对应;后者将前者的节点链指到知识库中对应的实体,将实体节点间的连边对应到知识库中对应实体间的关系,并依据人工定义的操作进行结构修改,从而与特定知识库对应.语义查询图与之类似,是将问句的依存句法解析树通过预定义的操作改造而来.而抽象语义表征中,节点表示概念,连边表示概念之间的语义角色关系.其概念和语义角色继承并改进自PropBank^[52],图5给出了问句“Who is starring in Spanish movies produced by Benicio del Toro?”的抽象语义表征^[53],其中“unknown”节点被视为答案节点.

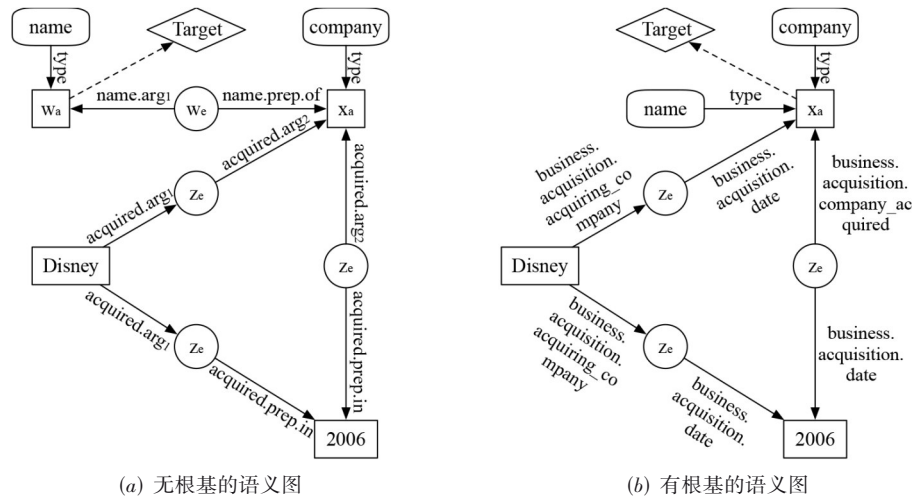


图4 问句“*What is the name of the company which Disney acquired in 2006?*”的语义图^[49]

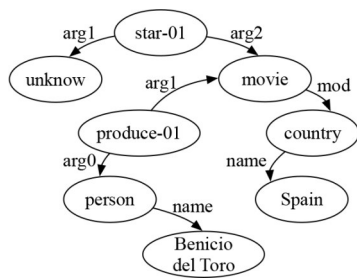


图5 问句“*Who is starring in Spanish movies produced by Benicio del Toro?*”的抽象语义表征^[53]

由于问句的句法结构与知识库子图的拓扑结构依旧存在差异,在根据解析得到的语义表征生成具体知识库查询时需要先通过图匹配算法找到与解析结果最相关的知识库子图.为了进一步降低构造难度,查询图直接在知识库子图上进行改造,得到的语义表征可以直接改写为知识库查询.以图6中的问句“Who first voiced Meg on Family Guy?”为例,查询图包括4类结点:常量节点(圆角矩形),即问句中出现的知识库实体、类

别、数值等,如图6中的“Family Guy”和“Meg Griffin”;变量节点(圆形),包括未被链指到知识库中的实体、事件、数值等节点,如图6中的“?y”; λ 变量(阴影圆形),表示问句的答案,即图6中的“?a”;聚合函数(菱形),作用在实体集合或者单个实体的数值属性上,如图6中的“argmin”.节点之间的连边均为知识库中的对应关系或属性.在查询图的常量节点中,其中一个节点作为根节点,对应问句的主题实体,而主题实体到答案节点的关系路径称为核心推理链(core inferential chain),其余常量节点和函数算子作为核心推理链上的分叉,对应问句中的约束条件.图6所示的查询图以“Family Guy”为主题实体,(Family Guy, cast, ?y, actor, ?a)为核心推理

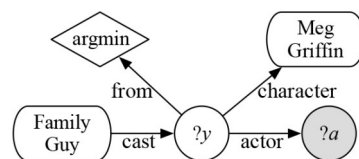


图6 问句“*Who first voiced Meg on Family Guy?*”的查询图^[3]

链, (?y, from, argmin) 和 (?y, character, Meg Griffin) 分别作为候选答案在配音工作开始时间和配音角色这两个知识库属性或关系上的约束条件。

3.3 程序语言

无论是 λ 演算、 λ -DCS 还是图状逻辑表达式, 最终均要转变为知识库查询语言, 因此一部分工作会直接采用查询语言作为语义表征, 如 SPARQL^[18]。与之类似地, 一部分工作会基于已有的或者自定义的解释器(interpreter)改造或设计相应的程序语言式语义表征, 典型的包括 Prolog^[54] 和 Lisp^[55] 等。

Prolog 主要基于一阶谓词逻辑(first-order logic), 主要是表达对象之间的关系, 并基于此来计算给定的查询。Prolog 的输入格式遵从霍恩子句(Horn clause)形式, 包含头部(head)和体部(body)两部分, 需要通过计算来返回输入的真值或满足输入为真的文字。例如问句“Which provinces border Jiangsu?”对应的 Prolog 表达式为 $? - \text{province}(X), \text{border}(X, \text{Jiangsu})$, 使该表达式为真值的 X 值即为问句的答案。同样基于一阶谓词逻辑的语义表征 FunQL^[56] (Functional Query Language) 是不含变量(variable-free)的, 这一点与 λ -DCS 类似。上述问句对应的 FunQL 表达式为 $\text{answer}(\text{province}(\text{border}(\text{Jiangsu})))$ 。

Lisp 以列表为主要数据结构, 列表内为一列 S-表达式(S-expression), 每个 S-表达式均为前缀表达式, 以左括号开始, 以右括号终止, 括号内为预定义的函数和函数论元。例如问句“Which is the largest city in China?”对应的 Lisp 表达式为 $[(\text{Hop China } !\text{CityIn}), (\text{Argmax } R1 \text{ Population})]$, 其中 Hop 和 Argmax 均为预定义的函数, ! 表示反关系, $(\text{Hop China } !\text{CityIn})$ 表示以实体 China 为尾实体, 返回与其以 CityIn 关系相连的头实体集合, 作为 $R1$, $(\text{Argmax } R1 \text{ Population})$ 表示对实体集合 $R1$ 按照 Population 这一属性值降序排序, 返回值最大的实体。部分工作^[43,57] 借鉴 Lisp, 专门设计了函数-论元形式的语义表征, 将自然语言问句的语义建模为一系列函数-论元表达式。

3.4 小结与讨论

综上所述, 语义表征的表达完备性决定了可覆盖问句类型的范围, 而构造便捷性影响了模型构造该语义表征的难易程度。在上述 3 类语义表征中, λ 演算及其衍生和程序语言均具有较强的语义表达能力, 其中函数-论元型的程序语言语义表征可通过更新自定义函数来增强其表达能力。而图状逻辑表达式受限于图结构, 表达完备性较弱于其他两类语义表征, 例如问句“Which province is larger, Jiangsu or Anhui?”所需的三目

运算符 $\text{size}(\text{Jiangsu}) > \text{size}(\text{Anhui})$? Jiangsu: Anhui 较难通过图状逻辑表达式呈现。

虽然表达能力较弱, 但图状逻辑表达式有更强的可读性, 并且与问句本身的句法结构或者图数据库的拓扑结构具有一定的相似性, 可以在问句的句法解析结果或者知识库子图上直接改造, 降低了语义表征的构造难度。而 λ 演算虽然表达能力强, 但是表述较为繁琐, 增加了构造难度, 因此 λ -DCS 在其基础上改进, 使得表述更紧凑。程序语言式语义表征通常以嵌套结构表示子式和语义成分之间的关系, 与问句句法结构及知识库拓扑结构差异较大, 无法像图状逻辑表达式一样利用问句结构或知识库拓扑结构的信息, 在生成嵌套结构时易出错。

因此, 平衡好表达完备性和构造便捷性, 是选择语义表征时需要考虑的主要问题, 在覆盖更多类型的问句的同时, 降低模型构造该语义表征的难度, 从而提升模型的解析效果。

4 基于枚举排序的语义解析方法

在选定了语义表征的具体形式后, 研究者设计相应的语义解析方法将非结构化的自然语言问句转变为结构化的语义表征。当前的语义解析方法可以分为基于枚举排序的方法和基于编码解码的方法。本节主要介绍基于枚举排序的方法。

如图 3 所示, 枚举排序类的方法先枚举所有可能的候选语义表征子式, 再利用打分模型对枚举的所有候选排序, 并输出高分候选。所有的候选均通过特定文法或者人工定义的处理步骤获得, 为了减少候选语义表征的搜索空间, 部分工作采用迭代式枚举排序, 即束搜索(beam search), 每轮迭代基于上一轮筛选得到的语义表征子式进行扩展, 再进入下一轮迭代, 从而完成对问句复杂语义的解析。在枚举产生的语义表征上, 一部分工作会直接构造与特定知识库对应的有根基的语义表征; 另一部分工作选择与特定知识库无关的无根基的语义表征, 最后将其映射到特定知识库上。表 3 将基于枚举排序的语义解析方法按照枚举方式进一步分类, 即依赖特定文法^[1,30,58-62] 和自定义处理步骤^[2,3,33,34,46,47,49,53,63-87] 列举了每种方式涉及的相关论文, 概括介绍其主要特点。下文分别介绍这几类方法, 并总结了枚举排序类方法的主要区别以及各自的优缺点。

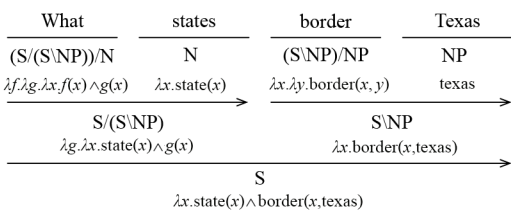
4.1 依赖特定文法

早期的工作在枚举语义表征阶段依赖特定文法中的规则, 包括问句词语与语义表征字段之间的表述映射、语义表征子式的组合规则; 在打分筛选阶段主要利用基于离散特征的打分函数。

表3 基于枚举排序的语义解析方法分类

枚举方式	相关文献	主要特点	
依赖特定文法	文献[1,30,58~62]	可控性强,枚举阶段依赖词汇映射与组合规则,具体应用时受限于规则,导致泛化性较差	
自定义处理步骤	基于线图	文献[2,63,64]	相邻子式组合方式由自定义操作控制,泛化性较好
	基于转移	文献[3,33,46,47,49,53,65~76]	基于问句的句法结构或问句相关的知识库子图,以当前语义表征构造情况为状态,通过自定义的处理步骤,迭代生成语义表征,处理步骤可调整,可干预性较强
	基于模板填充	文献[34,77,79~87]	根据问句的语义信息直接从知识库中抽取三元组或子图,填充到候选查询模板中构成候选知识库查询.模板严格限定了最终表征的结构,模板的覆盖率决定了可回答问句类型的上限

如图7所示,Zettlemoyer等人^[1]在枚举阶段采用组合范畴文法,首先通过词汇表将问句中的词语映射到所有可能的范畴,再根据句法类别和组合规则进行组合,得到候选的语义表征.例如图7的问句“What states border Texas?”中“states”既可能对应到名词“州”,也可能对应到动词“声明”,其名词范畴“ $N: \lambda x. state(x)$ ”中包含了该词的句法类别“N”,即名词,与语义表征 $\lambda x. state(x)$,其动词范畴“(SNP)/NP: $\lambda x. \lambda y. state(x, y)$ ”中句法类别“(SNP)/NP”包含了相应的组合规则,即“state”需要先向右结合一个名词短语(NP),再向左结合一个名词短语.正是由于一个单词可能存在多个语义表征字段与之对应,再加上组合顺序的不确定性,一个问句在组合范畴文法下可以枚举出多个候选语义表征.Zettlemoyer等人^[1]在打分阶段采用对数线性函数对候选语义表征进行筛选.

图7 组合范畴文法解析示意^[1]

后续工作主要从3个角度进行改进:(1)尝试词汇表的不同获取方式^[30,58~60],从而提升词汇表的扩展能力;(2)当前方法受限于文法规则,无法应对自然语言语序灵活多变、词语省略的现象,部分工作^[61]降低组合规则的限制;(3)当前方法从问句直接解析出有根基的 λ 演算的难度大,部分工作^[62]引入无根基的 λ 演算,在语义成分的表述上接近于输入问句,在结构和形式上接近于最终语义表征,从而降低解析难度.

4.2 自定义处理步骤

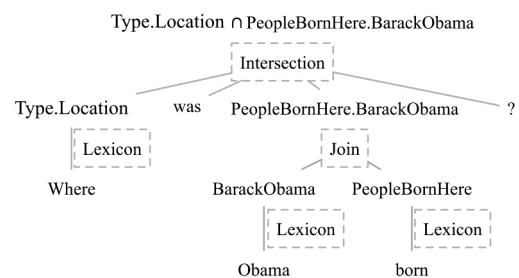
早期依赖特定文法的方法在枚举阶段由于从数据集中归纳出表述映射与组合规则,泛化性受限.因此,

近期的工作利用自定义的处理步骤枚举候选语义表征,解析过程简洁明了.这些工作可以进一步分为3类,分别为基于线图(chart)的、基于转移(transition)的和基于模板填充(template filling)的.

4.2.1 基于线图的方法

基于线图的方法在枚举阶段主要借助词汇映射与自定义操作,先对问句中的词语或短语进行词汇映射,从映射后获得的词级别子式出发,通过自定义操作进行组合,并利用打分函数进行筛选,迭代生成最终的语义表征.如图8所示,Berant等人^[2]以 λ -DCS作为目标语义表征,首先利用构造好的词汇映射表,将问句“Where was Obama born?”中的每个词语映射到所有可能的一元或二元逻辑表达式,如将“Obama”映射为一元逻辑表达式“BarackObama”,表示姓名为巴拉克·奥巴马的实体,将“born”映射到二元逻辑表达式“PeopleBornHere”,表示所有以出生地关系相连的实体对集合;再通过连接(join)、交叉(intersection)、桥接(bridging)等简单操作对任意两个不重复的相邻子逻辑表达式进行两两组合,生成新的候选子式,如“BarackObama”和“PeopleBornHere”可通过连接操作得到一元逻辑表达“PeopleBornHere.BarackObama”,表示巴拉克·奥巴马的出生地,同时利用对数线性函数对所有可能的组合进行打分筛选.

后续工作主要从2个角度进行改进:(1)当前方法的打分函数未考虑问句与语义表征的结构差异,部分

图8 问句“Where was Obama born?”的解析流程示意^[2]

工作^[63]改良打分函数,利用人工模板将候选语义表征复述为自然语言问句,以复述问句与输入问句的相似度为打分函数的特征;(2)为了减少当前方法按固定顺序自底向上构建语义表征时产生的干扰项,部分工作^[64]调整枚举阶段的子式组合顺序,利用打分函数从当前候选中选择子式置于线图中,同左右两边的子式组合为新的子式。

4.2.2 基于转移的方法

基于转移的方法定义了迭代过程中的不同状态,状态对应当前语义表征子式,当前状态下可执行的操作得到的所有候选语义表征通过打分筛选,形成新一轮的状态,从而完成状态转移。

部分方法^[46,47,49,53,65-68]采用语义分析(semantic analysis)的思路,先利用特定语法、依存解析工具、自定义操作等获取问句的中间形态,如无根基的图状逻辑表达式,再通过实体链接、关系映射、图匹配算法等操作,将中间形态对应到知识库子图上,得到最终语义表征,如有根基的图状逻辑表达式。因此,在这类工作中,可执行的操作包括特定文法解析、依存解析、自定义操作、实体链接、关系映射、图匹配等,由上述操作枚举候选语义表征,打分函数进行筛选,从而实现状态的转移。例如,Reddy 等人^[49]首先利用依存解析为问句生成了无根基的 λ 演算,并将其改造为无根基的语义图,如图4(a)所示,在此基础上,通过自定义的收缩(contract)操作与展开(expand)操作进一步改写无根基的语义图,从而减少结构差异;再通过实体链接、关系映射将无根基的语义图中连边上的标签变成知识库中的关系,类别结点变成实体类别,实体结点对应到知识库中的实体,从而得到有根基的语义图,如图4(b)所示。为了在依存解析得到的无根基语义表征中减少冗余成分,Hu 等人^[67]先通过节点识别找出问句中的实体、类别、字符串,然后根据问句的依存解析结果对识别出的节点进行连结(connect)、合并(merge)、展开(expand)、折叠(fold)得到语义查询图,再通过图匹配算法将其与知识库子图对应,并采用对数线性模型对其进行打分筛选。由于依存解析工具解析的结果与知识库查询存在较大差异,Sun 等人^[68]在依存句法的基础上设计了骨架文法(skeleton grammar)对问句进行解析,再通过自定义操作构造出语义查询图。该文法的解析结果为一棵有向树,其节点表示问句中的文本片段,连边表示文本片段之间的依存关系。

考虑到问句的句法结构与知识库子图的拓扑结构存在差异,为了进一步降低构造难度,Yih 等人^[3]直接通过知识库子图构建查询图作为问句的语义表征,并定义了图9所示的4种状态,分别是空图 ϕ 、只有主题实体的单节点图 S_e 、含有核心推理链的图 S_p 、含有额外约

束条件的复杂查询图 S_c ,针对这4种状态,Yih 等人^[3]设计了4种动作,包括主题实体链接 A_e 、确定核心推理链 A_p 、添加聚合函数 A_a 和约束条件 A_c 。其中, A_e 即利用实体链接模型识别出问句中的所有实体并链接到知识库上,作为候选主题实体; A_p 即在知识库中枚举以候选主题实体出发的一跳或者两跳关系路径,作为候选核心推理链,而核心推理链的末端节点作为候选答案; A_c 与 A_a 即依据启发式规则在候选核心推理链上添加约束条件或聚合函数。在打分筛选阶段,上述动作得到的候选均采用对数线性函数进行筛选。

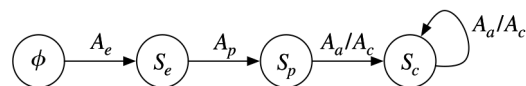


图9 基于状态转移的查询图生成流程示意^[3]

以图6所示问句“Who first voiced Meg on Family Guy?”为例,Yih 等人^[3]首先通过主题实体链接 A_e 识别出问句中的实体“Family Guy”和“Meg Griffin”,并以得分高的候选实体为主题实体,不妨假设“Family Guy”为主题实体,再通过枚举在知识库中以“Family Guy”为头实体的一跳或者两跳关系路径作为候选核心推理链,如“Family Guy, cast, ?y, actor, ?a), (Family Guy, genre, ?a)等,得分高的候选核心推理链进入下一轮处理,即添加聚合函数 A_a 和约束条件 A_c 。不妨以(Family Guy, cast, ?y, actor, ?a)例,问句中除了主题实体以外的其他实体作为实体约束,连向核心推理链上的变量节点或答案结点,因此(?y, character, Meg Griffin)作为候选答案在配音角色这个关系上的约束;同时由于问句中出现了“first”这一最高级词汇,触发了排序操作,因此(?y, from, argmin)作为候选答案在配音工作开始时间这个属性上的约束。由此,图6所示的查询图最终生成。

在迭代枚举阶段,后续工作主要从以下2个角度改进:(1)当前方法可处理问句约束类型较少,部分工作细化约束条件类型及其触发规则^[33];(2)当前方法限制核心推理链长度,无法回答多跳问句,部分工作调整处理步骤,设计不同的多跳推理机制,从而处理核心推理链更长的多跳关系问句^[69,70]。

在打分筛选阶段,后续工作^[70-76]对打分函数中间问句和核心推理链的向量表示计算方式进行改进,从而抓住问句与核心推理链或查询图在整体或局部上的关联性。

4.2.3 基于模板填充的方法

基于模板填充的方法根据问句的语义信息直接从知识库中抽取三元组或子图,填充到查询模板中构成知识库查询。这类工作大多需要先选择可能的查询模

板,再将候选实体和候选关系信息填充到特定模板中,最后用打分函数完成筛选。

不同方法的主要区别在于获取查询模板的方式上,部分方法从数据集中归纳总结,得到查询模板。如 Unger 等人^[77]先对问句进行词性标注和词汇映射,将映射后的所有词汇输入到已有解析工具 Pythia^[78],并基于总结出的人工规则将解析结果改写为 SPARQL 模板; Bast 等人^[79]从数据集中人工归纳出图 10 所示的 3 种查询模板作为限定,其中 e_1 和 e_2 均为问句中出现的实体的占位符, m 为变量占位符, t 为答案占位符, r_1, r_2, r_3 为关系占位符,例如问句“*What company did Henry Ford work for?*”可以将主题实体“Henry Ford”相关联的一跳关系填入第一个模板,如“(Henry Ford, birthdate, t)”,两跳关系填入第二个模板,如“(Henry Ford, employment, m , company, t)”,由于没有出现额外实体或实体类别,因此第三个模板未得到填充,填充后得到的所有候选查询通过逻辑回归(logistic regression)和随机森林(random forest)模型进行打分排序; Zheng 等人^[80]通过将多个数据集中的查询图的节点和连边的标签剔除,构造了 12 个结构化查询模板,并根据输入问句的句法解析结果,以深度模型筛选模板并进行模板填充。

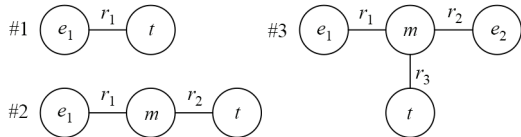


图 10 预定义的 3 种模板^[79]

后续工作^[34,81-84]从数据集中自动构建问句模板与查询模板的配对,当输入新的问句时,通过计算输入问句与已有的问句模板的相似度,如依存解析结果之间的最小树编辑距离等,找到合适的查询模板。而不同方法构造的问句模板在形式也不尽相同,例如, Abujabal 等人^[34,81]针对每个问句首先构造对应的查询模板,再依据查询模板,从问句的依存解析树中抽象出骨架作为问句模板,具体地, Abujabal 等人^[34,81]首先通过实体链接找到包含问句所有实体和答案的最小知识库子图,将答案和其他未提及实体替换为变量,构造为候选查询模板,再将问句中的短语和候选查询模板上的语义成分构造为二部图,通过求解整数线性规划问题,删除候选查询模板上有歧义的或者多余的边,并将所有的实体和关系替换为对应的占位符,从而得到查询模板;再对问句的依存树删除无法跟剪枝后的查询模板对应的成分,并将依存树中的复合短语替换为词性标签,将能对应到查询中的文本提及也替换为实体、关系、类别的标签,从而得到问句模板。而 Cui 等人^[84]直接将问答数据集中每一个样例问句的实体替换为实体类

别,构造出问句模板。相比之下, Cui 等人^[84]的模板更实例化,而 Abujabal 等人^[34,81]的模板有更强的泛化性。

然而从数据集中归纳出的模板有限,部分方法采用模型生成结构更多样的模板。例如 Bhutani 等人^[85]针对给定的问句,利用增强指针网络^[40](augmented pointer networks)生成查询模板,描述语义表征子式如何被连接和交叉操作组合为最终的语义表征,再通过 Yih 等人^[3]的方法来生成语义表征子式从而填充相应的查询模板。Ding 等人^[86]将标注数据中的查询图上的常量节点和部分谓词关系替换为占位符得到查询结构(query structure),并在查询结构上进一步分解得到查询子结构(query substructure),针对输入问句,先用神经网络模型判断需要哪些查询子结构,再判断将查询子结构按不同方式组合的查询模板与输入问句的匹配概率,最后将匹配得到的查询模板实例化。Chen 等人^[87]定义了查询模板的生成流程,包括向当前模板添加某一类节点、选择当前模板中的某个节点、为新添加的节点和被选择的节点之间添加某一类连边,并利用一个基于编码器-解码器的神经网络模型,从标注数据中学习如何迭代生成问句对应的查询模板,同时学习另一个打分函数对模板实例化后的候选查询进行排序筛选。

4.3 小结与讨论

综上所述,枚举排序这一类方法在应用时主要区别在两个方面:一个是特定文法、人工定义的处理步骤、或者模板库的不同,导致每轮迭代能枚举获得的候选语义表征子式不同,对方法本身的完备性也有不同的影响;另一个是打分函数设计多样,早期主要依赖以人工特征为输入的对数线性函数,随着深度学习的不断发展,卷积神经网络^[3]、循环神经网络^[71,72]、图神经网络^[73]、预训练语言模型^[70]等深度学习模型也得到广泛采用,打分函数考虑的特征信息愈发多样。

枚举排序这类方法的主要优点包括:(1) 可控性强,因为每轮迭代扩展得到的所有候选均符合规范,所以最终的语义表征是可执行的;(2) 可解释性强,因为解析的过程有确定性的语义表征子式或中间状态,所以每一步均可追溯。但不同的技术路线也展现出不同的特点。例如,基于模板填充的方法利用模板严格限定了最终表征的结构,模板的覆盖率决定了可回答问句类型的上限,使用固定模板库的方法泛化性较差,但增量更新时只需更新模板库,较其他方法更新组合规则、调整处理步骤会更容易;基于线图的方法通过词汇映射和组合规则约束生成子式,相比完全依赖文法规则的方法和使用固定模板库的方法,能灵活地应对语义组合复杂的问句;基于转移的方法在枚举阶段可调整处理步骤以覆盖更多类型的问句,可干预性强。

5 基于编码解码的语义解析方法

本节主要介绍基于编码解码的语义解析方法. 相比枚举排序类方法依赖特定文法或自定义处理步骤, 编码解码的方法采用基于编码器-解码器架构的模型直接学习自然语言到语义表征的映射, 其中编码器将自然语言问句转变为低维稠密的向量表示, 解码器根据编码器输出的问句向量输出符号化的语义表征. 如表 4

所示, 本文按照与知识库交互的方式将现有方法分为两类, 分别是单次交互^[43,88-103]与迭代交互^[57,104-106]. 前者在测试阶段解码时不与知识库进行交互, 只将最后的完整语义表征在知识库中执行, 这种方式需要在训练阶段以完整的语义表征为标注数据; 后者在解码阶段与知识库进行迭代交互, 执行子式获取中间结果, 并指导后续子式生成, 这种方式适用于训练阶段只有最终答案作为标注的弱监督(weak supervision)场景.

表 4 基于编码解码的语义解析方法分类

方法分类	相关文献	主要特点
单次交互	文献[43,88~97,99~103]	知识库查询开销小,不需要与知识库多次交互; 解码阶段无知识库指导,搜索空间大
迭代交互	文献[57,104~106]	知识库查询开销大,需要与知识库多次交互; 解码阶段与知识库交互缩减搜索空间,指导语义表征子式生成

5.1 单次交互

早期的工作以神经机器翻译的方式, 将语义表征视为输入问句在符号语义空间的译文, 只在输出完整语义表征后与知识库进行交互, 如 Jia 等人^[88]采用了基于注意力机制(attention mechanism)的编码器-解码器模型, 其中编码器与解码器均采用长短时记忆网络, 并且在解码阶段利用了基于注意力的复制机制(copy mechanism)将输入问句中的实体名等长尾字符串输出. 后续工作主要针对解码器架构、输出形式、解析流程这三方面进行了调整.

5.1.1 解码器架构

部分工作考虑到以长短时记忆网络为代表的传统解码器通常以序列化文本为输出, 难以捕捉语义表征的结构信息, 因此对解码器进行了适应性调整. 例如, Dong 等人^[89]利用分层树解码器(hierarchical tree decoder), 如图 11 所示, 以宽度优先的方式, 按层依次解码嵌套子树, 其中非终止符<n>作为嵌套树结构中的语义表征子式的占位符. 解码过程中, 解码器按层依次解码, 输出</s>时当前层的子树解码完毕, 并将当前层输出的<n>替换为非终止符, 即根据当时的解码器的隐状态向量解码嵌套子树, 由此完成自顶向下, 从左向右的语义解析过程.

Cheng 等人^[90,91]采用解码嵌套子树的方式生成目标语义表征 FunQL, 其解码器为栈-长短时记忆网络(stack-LSTM), 并且设置了缓冲区-栈(buffer-stack), 其中缓冲区存放输入的问词语, 栈里存放生成的语义表征字段. 图 12 展示了问句“Who is Barack Obama’s eldest daughter?”的树形式的 FunQL 语义表征, 树上的节点分为非终止符和终止符. 其中, 非终止符包括聚合函数, 如排序(argmax); 终止符包括实体占位符、关系占位符. 在解码时, 解码器以深度优先的方式迭代输出语

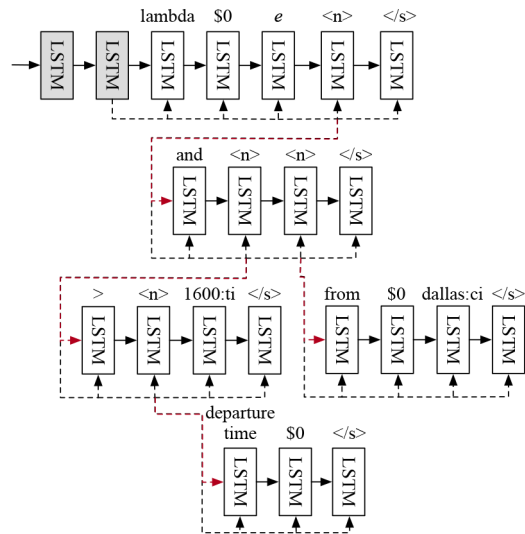


图 11 基于分层树解码器的模型解析问句^[89]

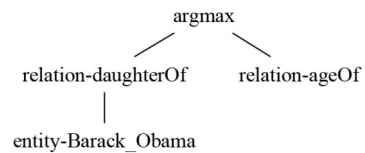


图 12 问句“Who is Barack Obama’s eldest daughter?”的语义表征^[90]

义表征字段, 候选字段包括 FunQL 中的聚合函数、实体占位符、关系占位符、代表子树完结的 Reduce 函数. 具体地, 若解码器输出聚合函数, 则直接填入对应的函数名作为非终止节点; 若输出关系或实体的占位符, 则需要进一步预测具体对应的知识库关系或实体; 若输出 Reduce 函数, 则将堆中的状态向量不断弹出, 直到得到一个非终止节点的状态向量, 再计算当前完成的子树的向量表示用于当前时刻解码器的状态更新. 考虑到语义表征中存在多种不同成分, 李等人^[92]将解码器分

为4个可递归调用的模块,分别对应查询语言中的不同成分.

5.1.2 输出形式

为了解码阶段引入针对语义表征的结构约束,保证生成的语义表征符合规范,部分工作不直接输出语义表征,而是输出与特定语义表征等价的派生序列,或者输出生成特定语义表征所需要的动作序列.此时可通过先验规则对解码阶段的候选字段或候选动作进行检查筛选,从而过滤掉不符合规范的候选项.

Xiao等人^[93]指出结构化语义表征可以等价转变为派生树(derivation tree),再通过前序遍历可得到其派生序列(derivation sequence),因此以派生序列作为解码器的目标输出,并在解码阶段引入结构约束,硬性剔除不符合相关限制的候选项.考虑到先验规则提供的结构约束可能存在纰漏,Xiao等人^[94]将先验规则提供的约束从直接保留或剔除变成先验概率打分,避免先验不够准确从而将部分正确的候选项直接剔除的情况,此时,即使先验概率打分较低,但如果解码器给出的分数足够高,解码器也能输出特定候选项.高等人^[95]将查询图切分为不同的语义块,将查询图生成过程形式化为语义块派生序列,解码器迭代输出语义块,从而生成完整的查询图.

为了解码器的输出形式更加紧凑,部分工作采用预定义的动作集合代替派生序列作为目标输出.例如,Chen等人^[96]定义了生成查询图的动作集合作为解码器的输出内容,具体包括添加变量节点、添加实体节点、添加类别节点、添加连边、添加函数操作,解码阶段下一个动作的选择与当前部分查询图强关联,不合规范的候选动作可以通过先验约束过滤.Yin等人^[97]将 λ 演算表示为抽象语法树(abstract syntax tree),并通过抽象语法描述语言文法(abstract syntax description language grammar)^[98]定义了生成抽象语法树的动作集合,从而得到问句对应的抽象语法树,其叶节点构成了最终语义表征,即 λ 演算,而候选动作的范围受到抽象语法描述语言文法的限制.Cao等人^[43]以程序语言KoPL为目标语义表征,该程序语言中的27种基本函数对应不同的动作,从而对多种知识元素进行操作并支持不同的查询意图.

5.1.3 解析流程

考虑到复杂问句对应的语义表征的语义成分丰富且结构复杂,将复杂问句直接解码为语义表征的难度大.为了降低直接解析的难度,部分工作调整了解析流程.提出多阶段的语义解析,将原有的端到端的解析流程拆分为多个子任务.

一类工作包含原始问句映射到中间形态,中间形态映射到最终语义表征这两个子任务,即先解码出问

句相关的中间形态(如无根基的语义表征、语义表征骨架等),再生成最终的语义表征.例如,Cheng等人^[99]先用解码器生成无根基的语义表征,再通过词汇映射将其转变为有根基的语义表征.后续工作的重点在于如何选择合适的中间形态.例如Dong等人^[100]以目标语义表征的粗粒度骨架作为中间形态,先通过一组基于长短时记忆网络的编码器-解码器生成目标语义表征的粗粒度骨架,再通过一组编码器-解码器,根据输入问句和粗粒度骨架,生成目标语义表征.这是由于粗粒度的骨架忽视了一些具体细节,如变量和实体名等,从而将高阶语义信息和低阶语义信息区分开,让解码器们能在不同的粒度去建模语义表征.相比无根基的语义表征和语义表征骨架这类结构化的中间形态,Cao等人^[101]利用半结构化的标准化问句(canonical question)作为中间形态,第一阶段通过编码器-解码器模型将自然语言问句复述为标准化问句,作为浅层语义解析,第二阶段通过另一个编码器-解码器模型将标准化问句映射为结构化的语义表征.

另一类工作包含原始问句分解为子问句、子问句解析、子问句解析结果组合为最终表征这三个子任务.例如,Zhang等人^[102]设计了三阶段的语义解析:第一阶段将输入的复杂问句解码为多个简单子问句;第二阶段根据原始问句与子问句的编码,解码出问句中所涉及的知识库关系;第三阶段基于原始问句、子问句、所有关系的编码,解码出最终的语义表征.Guo等人^[103]设计了问句切分器,迭代地从问句中切出子片段,交由解析器将该片段映射为一个部分语义表征;最后,所有的中间解析结果被组合为最终的语义表征,从而降低模型解析复杂问句的难度.

5.2 迭代交互

为了更好地利用知识库中的信息辅助解码,部分工作在解析过程中与知识库迭代交互.这类工作在解码阶段主要输出生成特定语义表征所需要的动作序列,并通过先验规则获取解码器当前符合规范的候选字段,但是和单次交互的工作相比,这部分工作能在与知识库迭代交互的过程中执行语义表征子式获取中间结果,从而指导后续子式生成,并进一步缩减搜索空间.这类工作往往需要额外缓存区保存中间结果.例如Liang等人^[104]提出神经符号机(Neural Symbolic Machine, NSM),在编码器-解码器模型的基础上,增加了键值记忆模块(key-variable memory)来保存实体链接结果以及输出的语义解析子式在知识库中执行的结果,作为后续子式的论元使用,同时在解码阶段引入语义表征的字段合法性检查,在每个时刻约束解码器的词典范围,从而确保输出的字段是符合规范且可执行的语义表征.

考虑到当前工作定义的函数操作有限,可覆盖问句类型受限,后续工作^[57,105,106]主要拓展了解码器可选择的函数操作,从而能够覆盖更多类型的问句,有效应对问句组合语义的复杂性.其中,Ansari等人^[105]定义了3种基本类型和5个函数操作用于生成程序语言式的语义表征:基本类型包括知识库事实(即实体、关系、类别)、基本数据类型(即整数值、布尔值、补全符)、组合数据类型(即集合,例如实体集合);5个函数操作分别作用于不同类型的论元上,得到集合、布尔值或整数值.Saha等人^[57]在此基础上丰富了函数操作类型与基本类型,提出20个函数操作符,3大类、9小类基本类型.在解析过程中,模型以解码器的隐藏状态作为当前程序的状态,并根据当前程序状态迭代地通过操作选择器、论元变量选择器、输出变量生成器构造程序字段,执行生成的程序子式,将中间结果写入到内存中,最终生成完整的语义表征.

为了更好地利用知识库的拓扑结构信息,Qiu等人^[106]以查询图为语义表征,将查询图中的三元组按照关系连边的类型分为5类,分别是知识库基本事实(basic)、并集事实(union)、数值比较(filter)、数值排序(ordinal)、聚合(aggregation),并在解析阶段采用层次化解码,顶层解码器基于编码后的问句和当前查询图决定待生成三元组的类别,底层解码器根据选定的类别,迭代解码出相应的三元组,即选择当前查询中的已有节点,选择一条连边,添加一个新的节点.

5.3 小结与讨论

综上所述,编码解码的方法利用模型直接建模语义表征的生成过程,这类方法的优点在于不依赖复杂的语法规则,候选语义表征的获取相对灵活.但语义解析需要保证生成的语义表征符合规范,因此这类方法往往需要在解码过程中加入多种不同约束或者先验信息,如:引入非终止符改造标注的语义表征,形成嵌套树结构,辅助解码器深入理解语义表征的结构信息;引入字段检查,为每一步解码时提供前后一致且符合规范的候选字段.

同时,与知识库的交互也影响语义解析的结果.单次交互的方法在测试阶段只在生成最终语义表征后与知识库交互一次,得到最终答案.虽然知识库查询开销小,但是这类方法在每一步解码时,需以知识库的全部实体或关系作为候选词表,在大规模知识库下面临搜索空间过大的问题,并且由于解码出的实体和关系,实体和实体之间可能不相连,导致最终语义表征的执行结果可能为空.而迭代交互的方法虽然知识库查询开销大,但是能从知识库中获取子式的中间结果,指导后续子式生成,降低了搜索空间.从适用场景角度分析,单次交互的方法相对更适用于小规模知识库问答,并

且需要以完整的语义表征作为标注,虽然标注成本较高,但模型在训练与测试阶段均比迭代交互的方法更高效;而迭代交互的方法相对更适用于大规模知识库问答,并且可以只以最终答案作为弱监督训练.

6 低资源场景下的数据利用

语义解析模型的训练往往依赖大量的标注数据,其中一部分工作需要完整的问句-语义表征对作为监督信号,另一部分工作允许只有问句的最终答案作为弱监督信号.但无论是何种监督方式,所需的数据标注都需要标注人员熟悉知识库的本体和实例,大量数据必然带来高昂的标注成本,如何在标注样本匮乏的低资源场景下利用现有数据,有效训练语义解析模型.也是近期受到关注的一个话题.本文在这一节分别从数据驱动和人机协同两个方面介绍近期工作的尝试.

6.1 数据驱动

当前数据驱动的方法主要分为3种:一种是通过问句扩充得到新的训练样例,即根据当前知识库人工构造查询,并通过归纳好的文法、模板或者半监督(semi-supervised)方式训练好的问句生成模型生成自然语言问句,为语义解析模型提供训练数据,或者利用无标注的问句提供给解析模型,利用人为设计的监督信号更新解析模型参数;第二种是通过重采样技术调整数据集分布,为出现次数较少的数据采样相似样本,弥补样本分布差异,从而缓解训练集中样本分布不均衡的现象;第三种是利用领域迁移技术,将其他领域的知识迁移至当前领域,从而完成低资源场景下语义解析模型的训练.

6.1.1 问句扩充

Jia等人^[88]从训练集中归纳出同步上下文无关文法(synchronous context-free grammar),利用归纳出的文法生成新的训练语料用于训练解析模型.但这种归纳认为问句中的实体表述或短语表述,同问句的其他表述在语义上条件独立,因此生成问句会出现搭配不一致等问题.为了获得更流畅通顺的问句,Wang等人^[107]借助非专家人群的众包协助,从模板问句复述出流畅自然的问句提供给解析模型.

随着深度学习在文本生成领域的广泛应用,后续工作^[108-111]利用神经网络训练问句生成模型,并将问句生成与问句解析组合,从而能根据两个任务各自的反馈为另一个任务提供监督信号.例如,Kočíský等人^[108]设计了两个基于长短时记忆网络的编码器-解码器叠加,第一个的目标是根据输入的语义表征解码出自然语言问句,第二个的目标是根据自然语言问句解码出语义表征,整个模型构成自编码(autoencoders)框架,并且将有标注的问句-语义表征对和人工构造的语义表征

同时喂给当前框架,针对有标注数据,第二个编码器的输入是真实问句,对应解码器的目标输出是标注的语义表征;针对人工构造的语义表征,第二个编码器的输入是第一个解码器的输出.为了增强问句生成与问句解析这两个任务之间的关联性,部分工作^[101,109-111]引入对偶学习框架,将任务的对偶性建模为损失函数中的正则项或者对偶任务的奖励信号.

不同于从人工构造的语义表征生成问句给解析模型训练,Yin等人^[97]直接利用无标注的问句提供给解析模型,并从两个角度评价解析结果,分别是训练好的问句生成模型能否从解析结果重构出原始问句,以及解析结果是否服从于语义表征的先验分布,该先验分布由基于长短时记忆网络的语言模型建模.

6.1.2 问句重采样

大数据集中的特定类型的样本也会面临低资源困境,即数据集中不同类型的问句存在分布差异,部分类型的样本在训练集中出现次数很少.

针对这一问题,部分工作^[112-114]提出问句重采样的方法,从数据集中找出与长尾样本或域外样本相似的数据训练语义解析模型,而非直接在原始数据集上随机采样进行训练,这是考虑到不同问句的特性差异与分布偏差导致很难用一个语义解析模型去适应不同类型的问句.例如,Lu等人^[112]维护了一个记忆缓存,用于保存具有代表性的有标注数据,在遇到之前训练没见过的特定类型样本时,语义解析器先从记忆缓存中找到相似的问句进行适应性训练,直到解析模型可以拟合该域外样本.为了在样本量较少的问句类型上也能取得稳定效果,Hua等人^[113,114]利用元学习(meta learning)为输入问句在数据集中寻找相似问句,针对性地改变数据集分布,调整模型参数.

6.1.3 领域迁移

考虑到特定领域的问答训练数据难以获取,部分工作提出采用领域迁移技术,将其他领域的知识迁移至当前领域,从而完成低资源场景下语义解析模型的训练.这类工作主要考虑到不同领域的问句仅在知识库相关的词汇(即实体、关系)上有区别,在问句主体结构上有共通之处,因此部分研究的关注点主要在如何迁移问句主体结构的知识.然而,虽然问句主体结构的知识可以迁移,但词汇映射依旧是获得最终语义表征的关键步骤,因此部分工作关注如何在低资源场景下完成知识库相关词汇的映射.

在迁移问句主体结构知识上,Herzig等人^[115]提出将问句中特定知识库相关的词汇替换为占位符,通过语义解析模型将替换后的问句解析为语义表征的抽象骨架,而语义解析模型可通过其他领域的知识库问答数据集进行训练,完成问句主体结构与语义表征抽

象骨架之间映射关系的知识迁移.

在知识库相关词汇映射上,Herzig等人^[115]将问句关联的知识库候选实体及候选关系填入到抽象骨架中对应的槽,再利用启发式设计的打分函数进行筛选.为了更有效地完成词汇映射表的扩充,Chen等人^[116]提出基于图的半监督学习框架的词汇映射学习方法,首先利用大量文本语料和词汇映射资源来学习词语之间的相似度,并构建用于图传播的图,接着使用图传播算法从少量标注的词汇中学习新的词汇.

6.2 人机协同

人机协同的方法^[117-119]针对解析模型在解析过程中的疑惑,由标注人员提供适当的反馈信号,辅助问句解析或更新模型参数,从而缓解低资源场景下缺乏有效标注的问题.这类工作主要关注两点:第一点是样本选择,即如何找到最值得提交给标注人员的样本,从而既能减少寻求协助的次数,又能有效提升模型性能;第二点是指导方式,即选择合适的指导方式,从而降低标注人员的指导成本.

在样本选择上,Iyer等人^[117]和Abujabal等人^[118]将在线上应用遇到的失败案例提交;Ni等人^[119]设计了打分函数,综合考虑了信息量、代表性、多样性这三个指标对数据集中的样本进行筛选.

在指导方式上,Iyer等人^[117]将线上应用阶段失败的案例收集后交给有经验的标注人员,直接提供完整的语义表征标注;Abujabal等人^[118]先利用之前基于模板的方法^[34]在小规模训练语料上训练得到一个问答模型,并维护一个问句-查询模板库,当有新的问句输入时,先对比该问句与库中问句的依存解析树,若未能匹配,则以一个相似度量函数从库中检索得到少量相似问句,并将对应的查询模板均实例化后查询知识库获得答案,将相似的问句与答案均返回给用户,并将用户满意的查询作为标注数据喂给模型进行新样例的学习,更新问句-查询模板库;Ni等人^[119]采用多轮迭代的方式,每一次迭代时,先训练语义解析模型至收敛;然后将打分函数挑选出的样本提交给标注人员,获得对应语义表征的骨架,从而降低直接标注具体语义表征的成本.

6.3 小结与讨论

低资源场景下的语义解析方法主要从数据驱动和人机协同这两个方向进行探索.在数据驱动方向上,现有工作分别从问句扩充、问句重采样以及领域迁移这三个角度出发.在问句扩充的角度下,现有方法利用归纳出的文法或训练出的问句生成模型,为解析模型提供更丰富且均衡的训练数据;部分工作亦通过众包的方式对人工构造的模板问句进行复述,从而获取自然语言问句,与当前数据集构造方式类似;在问句重采样

的角度下,现有方法通过元学习和重采样技术为当前样例调整训练集分布,从而缓解数据集中样本分布不均衡,特定类型样本面临低资源困境的情况;在领域迁移的角度下,现有方法既要考虑到如何迁移不同领域的共性知识,又要考虑如何低成本获取不能迁移的领域特定知识.而人机协同的方法将最能提升模型性能的样例提供给标注人员,根据反馈更新模型参数或模板库.

这两个方向各有优势,相比人机协同的方法,数据驱动的方法成本更低,例如通过在问句生成数据集上微调预训练语言模型,能迅速为解析模型提供大量训练样本,但提供的样本的质量不如人机协同的方法,效果提升可能没有人机协同的方法显著.因此,在模型初始表现较低时,可通过数据驱动的方法进行加强;在模型表现相对较高,更需要高质量数据或反馈时,可通过人机协同的方法进行加强.

7 未来展望

通过上述比较分析可以看出,面向知识库问答的语义解析已经取得一定进展,但仍然存在没有克服的问题和值得深入研究的方向.

7.1 多模态异构知识的统一语义表征

由于现实世界的知识模态多样、结构各异且更新迅速,单一知识库难以覆盖用户的信息需求,部分工作^[19,22,23]借助开放文本辅助知识库问答模型的决策,但依旧未能有效利用大量结构化、半结构化、非结构化知识以及多模态信息,如常识知识库、社区问答库、图片、视频等.将多模态异构的知识有效整合,并且以统一的语义表征进行知识表示,可以有效支撑在不同模态和结构的知识上进行协同查询推理.

7.2 多轮知识库问答中的语义解析

在实际场景中,用户往往围绕一个主题进行多次提问,在多轮问答场景中,单个问句常出现语义缺失,如实体指代不清、谓词缺失等,导致单轮语义解析模型无法准确理解问句的语义.现有工作^[120,121]主要利用对话历史中已生成的语义表征的信息对后续问句的表征进行补全,因此如何根据对话状态构建前后语义表征字段之间的共指关联成为研究重点.而在关联置信度低的情况下,需要探索如何根据缺失的语义信息生成相应问句提供给用户,通过用户反馈来明确查询意图.

7.3 知识库问答中语义解析模型的鲁棒性

近期的语义解析工作通常采用深度学习模型,但深度学习模型往往因为鲁棒性不足而被诟病.近年来,在文本阅读理解中,模型的鲁棒性开始得到关注,主要对答案来源的文本加入噪音,从而增强问答模型的抗

攻击能力^[122,123].而在知识库问答领域,除了考虑知识库本身的噪音信息外,还需要考虑问句中的噪音对语义解析的影响,主要包括一词多义、拼写错误、介词短语引起的歧义、句式倒装等.增强语义解析模型对问句和知识库中的噪音的抗干扰能力是实现知识库问答系统落地的关键.

7.4 知识库问答中语义解析模型的低资源训练

当前的语义解析方法往往是在训练数据较为充分的情况下进行训练,区别主要在标注数据是完整的语义表征还是只有最终的答案.但知识库问答系统落地时经常是面向特定领域的,缺乏标注数据.现有方法主要利用问句生成和重采样进行数据增强和分布调整、或者利用领域迁移或人机协同引入额外知识和反馈信号,尚缺乏对预训练-微调这一范式的探索,包括利用知识库本身信息、通用语言知识、跨领域语义解析模型,进行融入知识的语言模型预训练、低资源下目标领域的语义解析模型的预训练等.

8 总结

面向知识库问答的语义解析技术将自然语言问句映射为计算机可读的结构化语义表征,是实现人机交互的重要手段.这一技术当前主要面临3个挑战,分别是如何选择合适的语义表征形式以表达问句的语义,如何解析问句的复杂语义并输出相应的语义表征,以及如何应对特定领域中数据标注成本高昂、高质量数据匮乏的问题.本文分析了知识库问答中常用的语义表征的特点与不足,然后梳理现有方法并总结分析其如何应对问句的复杂语义,接着介绍了当前方法在标注数据匮乏的低资源场景下的尝试,最后展望并讨论了面向知识库问答的语义解析的未来发展方向.

参考文献

- [1] ZETTLEMOYER L S, COLLINS M. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars[C]//Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence. Edinburgh: AUAI Press, 2005: 658-666.
- [2] BERANT J, CHOU A, FROSTIG R, et al. Semantic parsing on freebase from question-answer pairs[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2013: 1533-1544.
- [3] YIH W T, CHANG M W, HE X D, et al. Semantic parsing via staged query graph generation: Question answering with knowledge base[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics

- and the 7th International Joint Conference on Natural Language Processing. Beijing: Association for Computational Linguistics, 2015: 1321-1331.
- [4] BORDES A, CHOPRA S, WESTON J. Question answering with subgraph embeddings[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). Doha: Association for Computational Linguistics, 2014: 615-620.
- [5] DONG L, WEI F R, ZHOU M, et al. Question answering over freebase with multi-column 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. Beijing: Association for Computational Linguistics, 2015: 260-269.
- [6] MILLER A, FISCH A, DODGE J, et al. Key-value memory networks for directly reading documents[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016: 1400-1409.
- [7] GREEN B F, WOLF A K, CHOMSKY C, et al. Baseball: An automatic question-answerer[C]//IRE-AIEE-ACM' 61 (Western): western joint IRE-AIEE-ACM computer conference. Los Angeles: ACM, 1961: 219-224.
- [8] HÖFFNER K, WALTER S, MARX E, et al. Survey on challenges of question answering in the semantic web[J]. *Semantic Web*, 2017, 8(6): 895-920.
- [9] DIEFENBACH D, LOPEZ V, SINGH K, et al. Core techniques of question answering systems over knowledge bases: A survey[J]. *Knowledge and Information Systems*, 2018, 55(3): 529-569.
- [10] WU P Y, ZHANG X W, FENG Z Y. A survey of question answering over knowledge base[C]//China Conference on Knowledge Graph and Semantic Computing. Hangzhou: Springer, 2019: 86-97.
- [11] LAN Y S, HE G L, JIANG J H, et al. A survey on complex knowledge base question answering: Methods, challenges and solutions[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence: Survey Track. Montreal: International Joint Conferences on Artificial Intelligence Organization, 2021: 4483-4491.
- [12] MILLER G A. WordNet: A lexical database for English [J]. *Communications of the ACM*, 1995, 38(11): 39-41.
- [13] AUER S, BIZER C, KOBILAROV G, et al. DBpedia: A Nucleus for a Web of Open Data[M]//The Semantic Web. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007: 722-735.
- [14] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: A collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data. Vancouver: ACM, 2008: 1247-1250.
- [15] VRANDEČIĆ D, KRÖTZSCH M. Wikidata: A free collaborative knowledgebase[J]. *Communications of the ACM*, 2014, 57(10): 78-85.
- [16] CHURCH A. A set of postulates for the foundation of logic[J]. *The Annals of Mathematics*, 1932, 33(2): 346-366.
- [17] CARPENTER B. *Type-logical Semantics*[M]. Cambridge: MIT Press, 1997.
- [18] HARRIS S, SEABORNE A, PRUD' HOMMEAU E. SPARQL 1.1 query language[J]. *W3C recommendation*, 2013, 21(10): 778.
- [19] XU K, REDDY S, FENG Y S, et al. Question answering on freebase via relation extraction and textual evidence [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016: 2326-2336.
- [20] HAO Y C, ZHANG Y Z, 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. Vancouver: Association for Computational Linguistics, 2017: 221-231.
- [21] YIN W, YU M, XIANG B, et al. Simple question answering by attentive convolutional neural network[C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. Osaka: The COLING 2016 Organizing Committee, 2016: 1746-1756.
- [22] SUN H T, DHINGRA B, ZAHEER M, et al. Open domain question answering using early fusion of knowledge bases and text[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018: 4231-4242.
- [23] SUN H T, BEDRAX-WEISS T, COHEN W. PullNet: Open domain question answering with iterative retrieval on knowledge bases and text[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019: 2380-2390.

- [24] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. *Neural Computation*, 1997, 9(8): 1735-1780.
- [25] ZHOU M, HUANG M, ZHU X. An interpretable reasoning network for multi-relation question answering[C]// *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe: Association for Computational Linguistics, 2018: 2010-2022.
- [26] QIU Y Q, WANG Y Z, JIN X L, et al. Stepwise reasoning for multi-relation question answering over knowledge graph with weak supervision[C]// *Proceedings of the 13th International Conference on Web Search and Data Mining*. Houston: ACM, 2020: 474-482.
- [27] HE G L, LAN Y S, JIANG J, et al. Improving multi-hop knowledge base question answering by learning intermediate supervision signals[C]// *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. New York: ACM, 2021: 553-561.
- [28] SHI J X, CAO S L, HOU L, et al. TransferNet: An effective and transparent framework for multi-hop question answering over relation graph[C]// *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Punta Cana: Association for Computational Linguistics, 2021: 4149-4158.
- [29] DAHL D A, BATES M, BROWN M, et al. Expanding the scope of the ATIS task: The ATIS-3 corpus[C]// *Proceedings of the workshop on Human Language Technology*. Plainsboro: ACM, 1994: 43-48.
- [30] CAI Q Q, YATES A. Large-scale semantic parsing via schema matching and lexicon extension[C]// *ACL 2013- Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Bulgaria: Association for Computational Linguistics, 2013: 423-433.
- [31] BORDES A, USUNIER N, CHOPRA S, et al. Large-scale simple question answering with memory networks [EB/OL]. (2015-06-05)[2022-03-01]. <https://arxiv.org/abs/1506.02075>.
- [32] YIH W T, RICHARDSON M, MEEK C, et al. The value of semantic parse labeling for knowledge base question answering[C]// *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin: Association for Computational Linguistics, 2016: 201-206.
- [33] BAO J, DUAN N, YAN Z, et al. Constraint-based question answering with knowledge graph[C]// *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka: The COLING 2016 Organizing Committee, 2016: 2503-2514.
- [34] ABUJABAL A, YAHYA M, RIEDEWALD M, et al. Automated template generation for question answering over knowledge graphs[C]// *Proceedings of the 26th International Conference on World Wide Web*. Perth: International World Wide Web Conferences Steering Committee, 2017: 1191-1200.
- [35] SAHA A, PAHUJA V, KHAPRA M, et al. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph[C]// *Proceedings of the AAAI Conference on Artificial Intelligence*. New Orleans: AAAI Press, 2018, 32(1): 705-713.
- [36] SU Y, SUN H, SADLER B, et al. On generating characteristic-rich question sets for QA evaluation[C]// *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin: Association for Computational Linguistics, 2016: 562-572.
- [37] SERBAN I V, GARCÍA-DURÁN A, GULCEHRE C, et al. Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus[C]// *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin: Association for Computational Linguistics, 2016: 588-598.
- [38] KEYSERS D, SCHÄRLI N, SCALES N, et al. Measuring compositional generalization: A comprehensive method on realistic data[EB/OL]. (2019-12-20)[2022-03-01]. <https://arxiv.org/abs/1912.09713>.
- [39] GU Y, KASE S E, VANNI M, et al. Beyond I.I.D.: Three levels of generalization for question answering on knowledge bases[C]// *Proceedings of the Web Conference 2021*. Ljubljana: ACM, 2021: 3477-3488.
- [40] TALMOR A, BERANT J. The web as a knowledge-base for answering complex questions[C]// *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans: Association for Computational Linguistics, 2018: 641-651.
- [41] TRIVEDI P, MAHESHWARI G, DUBEY M, et al. LC-QuAD: A corpus for complex question answering over knowledge graphs[C]// *International Semantic Web Conference*. Vienna: Springer, 2017: 210-218.
- [42] DUBEY M, BANERJEE D, ABDELKAWI A, et al. LC-QuAD 2.0: A Large dataset for complex question answering over Wikidata and DBpedia[C]// *International Semantic Web Conference*. Auckland: Springer, 2019: 69-78.

- [43] CAO S L, SHI J X, PAN L M, et al. KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: Association for Computational Linguistics, 2022: 6101-6119.
- [44] ZHANG Y Y, DAI H J, KOZAREVA Z, et al. Variational reasoning for question answering with knowledge graph [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018: 6069-6076.
- [45] LIANG P. Lambda dependency-based compositional semantics[EB/OL]. (2013-09-17)[2022-03-01]. arXiv: 1309.4408. <https://arxiv.org/abs/1309.4408>.
- [46] REDDY S, LAPATA M, STEEDMAN M. Large-scale semantic parsing without question-answer pairs[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 377-392.
- [47] ZOU L, HUANG R Z, WANG H X, et al. Natural language question answering over RDF: A graph data driven approach[C]//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. Snowbird: ACM, 2014: 313-324.
- [48] BANARESCU L, BONIAL C, CAI S, et al. Abstract meaning representation for sembanking[C]//Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. Sofia: Association for Computational Linguistics, 2020: 178-186.
- [49] REDDY S, TÄCKSTRÖM O, COLLINS M, et al. Transforming dependency structures to logical forms for semantic parsing[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 127-140.
- [50] STEEDMAN M. Surface Structure and Interpretation[M]. Cambridge: MIT Press, 1996.
- [51] STEEDMAN M. The Syntactic Process[M]. Cambridge: MIT Press, 2000.
- [52] PALMER M, GILDEA D, KINGSBURY P. The proposition bank: An annotated corpus of semantic roles[J]. Computational Linguistics, 2005, 31(1): 71-106.
- [53] MIHINDUKULASOORIYA N, ROSSIELLO G, KAPANIPATHI P, et al. Leveraging semantic parsing for relation linking over knowledge bases[C]//Lecture Notes in Computer Science. Athens: Springer, 2020: 402-419.
- [54] COLMERAUER A, ROUSSEL P. The birth of prolog [C]//The second ACM SIGPLAN conference on History of programming languages. Cambridge: ACM, 1993: 37-52.
- [55] MCCARTHY J. Recursive functions of symbolic expressions and their computation by machine, Part I[J]. Communications of the ACM, 1960, 3(4): 184-195.
- [56] KATE R J, WAH Y, RAYMOND W, et al. Learning to transform natural to formal languages[C]//Proceedings of AAAI Conference on Artificial Intelligence. Pittsburgh: AAAI Press, 2005: 1062-1068.
- [57] SAHA A, ANSARI G A, LADDHA A, et al. Complex program induction for querying knowledge bases in the absence of gold programs[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 185-200.
- [58] KWIATKOWSKI T, ZETTLEMOYER L, GOLDWATER S, et al. Inducing probabilistic CCG grammars from logical form with higher-order unification[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge: Association for Computational Linguistics, 2010: 1223-1233.
- [59] KWIATKOWSKI T, ZETTLEMOYER L, GOLDWATER S, et al. Lexical generalization in CCG grammar induction for semantic parsing[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh: Association for Computational Linguistics, 2011: 1512-1523.
- [60] WANG A, KWIATKOWSKI T, ZETTLEMOYER L. Morpho-syntactic lexical generalization for CCG semantic parsing[C]//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014: 1284-1295.
- [61] ZETTLEMOYER L S, COLLINS M. Online learning of relaxed CCG grammars for parsing to logical form[C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL). Prague: Association for Computational Linguistics, 2007: 678-687.
- [62] KWIATKOWSKI T, CHOI E, ARTZI Y, et al. Scaling semantic parsers with on-the-fly ontology matching[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: Association for Computational Linguistics, 2013: 1545-1556.
- [63] BERANT J, LIANG P. Semantic parsing via paraphrasing [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: Association for Computational Linguistics, 2014: 1415-1425.
- [64] BERANT J, LIANG P. Imitation learning of agenda-

- based semantic parsers[J]. *Transactions of the Association for Computational Linguistics*, 2015, 3: 545-558.
- [65] KAPANIPATHI P, ABDELAZIZ I, RAVISHANKAR S, et al. Leveraging abstract meaning representation for knowledge base question answering[C]//*Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Punta Cana: Association for Computational Linguistics, 2021: 3884-3894.
- [66] HU S, ZOU L, YU J X, et al. Answering natural language questions by subgraph matching over knowledge graphs [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 30(5): 824-837.
- [67] HU S, ZOU L, ZHANG X B. A state-transition framework to answer complex questions over knowledge base [C]//*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics, 2018: 2098-2108.
- [68] SUN Y W, ZHANG L L, CHENG G, et al. SPARQA: Skeleton-based semantic parsing for complex questions over knowledge bases[C]//*Proceedings of the AAAI Conference on Artificial Intelligence*. New York: AAAI Press, 2020, 34(5): 8952-8959.
- [69] CHEN Z Y, CHANG C H, CHEN Y P, et al. UHop: an unrestricted-hop relation extraction framework for knowledge-based question answering[C]//*Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis: Association for Computational Linguistics, 2019: 345-356.
- [70] LAN Y S, JIANG J. Query graph generation for answering multi-hop complex questions from knowledge bases [C]//*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, 2020: 969-974.
- [71] YU M, YIN W P, HASAN K S, et al. Improved neural relation detection for knowledge base question answering [C]//*Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver: Association for Computational Linguistics, 2017: 571-581.
- [72] LUO K Q, LIN F L, LUO X S, et al. Knowledge base question answering via encoding of complex query graphs [C]//*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics, 2018: 2185-2194.
- [73] SOROKIN D, GUREVYCH I. Modeling semantics with gated graph neural networks for knowledge base question answering[C]//*Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe: Association for Computational Linguistics, 2018: 3306-3317.
- [74] ZAFAR H, NAPOLITANO G, LEHMANN J. Formal query generation for question answering over knowledge bases[C]//*European Semantic Web Conference*. Heraklion: Springer, 2018: 714-728.
- [75] MAHESHWARI G, TRIVEDI P, LUKOVNIKOV D, et al. Learning to rank query graphs for complex question answering over knowledge graphs[C]//*International Semantic Web Conference*. Auckland: Springer, 2019: 487-504.
- [76] ZHU S G, CHENG X, SU S. Knowledge-based question answering by tree-to-sequence learning[J]. *Neurocomputing*, 2020, 372: 64-72.
- [77] UNGER C, BÜHMANN L, LEHMANN J, et al. Template-based question answering over RDF data[C]//*Proceedings of the 21st International Conference on World Wide Web*. Lyon: ACM, 2012: 639-648.
- [78] UNGER C, CIMIANO P. Pythia: Compositional meaning construction for ontology-based question answering on the semantic web[C]//*International Conference on Application of Natural Language to Information Systems*. Alicante: Springer, 2011: 153-160.
- [79] BAST H, HAUSSMANN E. More accurate question answering on freebase[C]//*Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. Melbourne: ACM, 2015: 1431-1440.
- [80] ZHENG W G, ZHANG M. Automated query graph generation for querying knowledge graphs[C]//*Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. New York: ACM, 2021: 2698-2707.
- [81] ABUJABAL A, SAHA ROY R, YAHYA M, et al. QUINT: Interpretable question answering over knowledge bases[C]//*Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Copenhagen: Association for Computational Linguistics, 2017: 61-66.
- [82] ZHENG W G, ZOU L, LIAN X, et al. How to build templates for RDF question/answering: An uncertain graph similarity join approach[C]//*Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. Melbourne: ACM, 2015: 1809-1824.
- [83] ZHENG W G, ZOU L, PENG W, et al. Semantic SPARQL similarity search over RDF knowledge graphs[J]. *Pro-*

- ceedings of the VLDB Endowment, 2016, 9(11): 840-851.
- [84] CUI W Y, XIAO Y H, WANG H X, et al. KBQA: Learning question answering over QA corpora and knowledge bases[J]. Proceedings of the VLDB Endowment, 2017, 10(5): 565-576.
- [85] BHUTANI N, ZHENG X Y, JAGADISH H V. Learning to answer complex questions over knowledge bases with query composition[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing: ACM, 2019: 739-748.
- [86] DING J W, HU W, XU Q X, et al. Leveraging frequent query substructures to generate formal queries for complex question answering[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019: 2614-2622.
- [87] CHEN Y R, LI H Y, HUA Y C, et al. Formal query building with query structure prediction for complex question answering over knowledge base[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Yokohama: ACM, 2021: 3751-3758.
- [88] JIA R, LIANG P. Data recombination for neural semantic parsing[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016: 12-22.
- [89] DONG L, LAPATA M. Language to logical form with neural attention[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016: 33-43.
- [90] CHENG J P, LAPATA M. Weakly-supervised neural semantic parsing with a generative ranker[C]//Proceedings of the 22nd Conference on Computational Natural Language Learning. Brussels: Association for Computational Linguistics, 2018: 356-367.
- [91] CHENG J P, REDDY S, SARASWAT V, et al. Learning an executable neural semantic parser[EB/OL]. (2017-11-14)[2022-03-01]. <https://arxiv.org/abs/1711.05066>.
- [92] 李青, 钟将, 李立力, 等. 一种按需聚合的语义解析图查询模型[J]. 电子学报, 2020, 48(4): 763-771.
- LI Q, ZHONG J, LI L L, et al. Semantic parsing graph query model for on-demand aggregation[J]. Acta Electronica Sinica, 2020, 48(4): 763-771. (in Chinese)
- [93] XIAO C Y, DYMETMAN M, GARDENT C. Sequence-based structured prediction for semantic parsing[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016: 1341-1350.
- [94] XIAO C Y, DYMETMAN M, GARDENT C. Symbolic priors for RNN-based semantic parsing[C]//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne: International Joint Conferences on Artificial Intelligence Organization, 2017: 4186-4192.
- [95] 高留杰, 赵文, 张君福, 等. G2S: 基于语义块的知识图谱问答语义解析[J]. 电子学报, 2021, 49(6): 1132-1141.
- GAO L J, ZHAO W, ZHANG J F, et al. G2S: Semantic segment based semantic parsing for question answering over knowledge graph[J]. Acta Electronica Sinica, 2021, 49(6): 1132-1141. (in Chinese)
- [96] CHEN B, SUN L, HAN X P. Sequence-to-action: End-to-end semantic graph generation for semantic parsing[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 766-777.
- [97] YIN P C, ZHOU C T, HE J X, et al. StructVAE: Tree-structured latent variable models for semi-supervised semantic parsing[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 754-765.
- [98] WANG D C, APPEL A W, KORN J L, et al. The Zephyr abstract syntax description language[C]//Proceedings of the Conference on Domain-Specific Languages on Conference on Domain-Specific Languages(DSL). Santa Barbara California: USENIX Association, 1997: 17.
- [99] CHENG J P, REDDY S, SARASWAT V, et al. Learning structured natural language representations for semantic parsing[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017: 44-55.
- [100] DONG L, LAPATA M. Coarse-to-fine decoding for neural semantic parsing[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 731-742.
- [101] CAO R S, ZHU S, YANG C Y, et al. Unsupervised dual paraphrasing for two-stage semantic parsing[C]//Pro-

- ceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 6806-6817.
- [102] ZHANG H Y, CAI J J, XU J J, et al. Complex question decomposition for semantic parsing[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 4477-4486.
- [103] GUO Y, LIN Z, LOU J G, et al. Iterative utterance segmentation for neural semantic parsing[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2021, 35(14): 12937-12945.
- [104] LIANG C, BERANT J, LE Q, et al. Neural symbolic machines: Learning semantic parsers on freebase with weak supervision[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017: 23-33.
- [105] ANSARI G A, SAHA A, KUMAR V, et al. Neural program induction for KBQA without gold programs or query annotations[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao: International Joint Conferences on Artificial Intelligence Organization, 2019: 4890-4896.
- [106] QIU Y Q, ZHANG K, WANG Y Z, et al. Hierarchical query graph generation for complex question answering over knowledge graph[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020: 1285-1294.
- [107] WANG Y S, BERANT J, LIANG P. Building a semantic parser overnight[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: Association for Computational Linguistics, 2015: 1332-1342.
- [108] KOČISKÝ T, MELIS G, GREFFENSTETTE E, et al. Semantic parsing with semi-supervised sequential autoencoders[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016: 1078-1087.
- [109] HU S, ZOU L, ZHU Z X. How question generation can help question answering over knowledge base[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Dunhuang: Springer, 2019: 80-92.
- [110] CAO R S, ZHU S, LIU C, et al. Semantic parsing with dual learning[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 51-64.
- [111] YE H, LI W J, WANG L. Jointly learning semantic parser and natural language generator via dual information maximization[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 2090-2101.
- [112] LU Z C, ARABSHAHI F, LABUTOV I, et al. Look-up and adapt: A one-shot semantic parser[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019: 1129-1139.
- [113] HUA Y C, LI Y F, HAFFARI G, et al. Retrieve, program, repeat: Complex knowledge base question answering via alternate meta-learning[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Yokohama: International Joint Conferences on Artificial Intelligence Organization, 2021: 3679-3686.
- [114] HUA Y C, LI Y F, HAFFARI G, et al. Few-shot complex knowledge base question answering via meta reinforcement learning[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP). Menlo Park: Association for Computational Linguistics, 2020: 5827-5837.
- [115] HERZIG J, BERANT J. Decoupling structure and lexicon for zero-shot semantic parsing[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018: 1619-1629.
- [116] CHEN B, AN B, SUN L, et al. Semi-supervised lexicon learning for wide-coverage semantic parsing[C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018: 892-904.
- [117] IYER S, KONSTAS I, CHEUNG A, et al. Learning a neural semantic parser from user feedback[C]//Proceed-

ings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017: 963-973.

- [118] ABUJABAL A, ROY R S, YAHYA M, et al. Never-ending learning for open-domain question answering over knowledge bases[C]//Proceedings of the 2018 World Wide Web Conference. Lyon: ACM, 2018: 1053-1062.
- [119] NI A S, YIN P C, NEUBIG G. Merging weak and active supervision for semantic parsing[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020, 34(5): 8536-8543.
- [120] SUHR A, IYER S, ARTZI Y. Learning to map context-dependent sentences to executable formal queries[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018: 2238-2249.
- [121] GUO D Y, TANG D Y, DUAN N, et al. Dialog-to-action: Conversational question answering over a large-scale knowledge base[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018: 2946-2955.
- [122] JIA R, LIANG P. Adversarial examples for evaluating reading comprehension systems[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 2021-2031.
- [123] LIU K, LIU X, YANG A, et al. A robust adversarial training approach to machine reading comprehension [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020, 34(5): 8392-8400.



王元卓(通讯作者) 男,1978年出生. 博士. 中国科学院计算技术研究所研究员,博士生导师,中科大数据研究院院长. 主要研究方向为网络大数据分析、开放知识计算、社交网络演化计算.

E-mail: wangyuanzhuo@ict.ac.cn



白 龙 男,1993年出生. 中国科学院计算技术研究所博士研究生. 主要研究方向为知识图谱、事件预测.

E-mail: bailong18b@ict.ac.cn



尹芷仪 女,1982年出生. 博士. 中国科学院计算技术研究所高级工程师,硕士生导师. 主要研究方向为社会计算、网络空间安全.

E-mail: yinzhiji@ict.ac.cn



沈华伟 男,1982年出生. 博士. 中国科学院计算技术研究所研究员,博士生导师,数据智能系统研究中心主任. 主要研究方向为网络数据挖掘、社交网络分析、图神经网络.

E-mail: shenhuawei@ict.ac.cn



白 硕 男,1956年出生. 博士. 中国科学院计算技术研究所研究员,博士生导师,恒生电子股份有限公司首席科学家.

E-mail: bshuo@sina.cn

作者简介



仇 轶 琦 男,1994年出生. 中国科学院计算技术研究所博士研究生. 主要研究方向为知识库问答、语义解析.

E-mail: qiuyunqi19b@ict.ac.cn