软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(1):236−265 [doi: 10.13328/j.cnki.jos.006913]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

常识问答研究综述*

范怡帆[1], 邹博伟[2], 徐庆婷[1], 李志峰[1], 洪 宇[1]

[1](苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

[2](Infocomm Research Institute of Singapore, Singapore 138635, Singapore)

通信作者: 洪宇, E-mail: tianxianer@gmail.com

摘 要: 常识问答是一项重要的自然语言理解任务, 旨在利用常识知识对自然语言问句进行自动求解, 以得到准确答案. 常识问答在虚拟助手或社交聊天机器人等领域有着广泛的应用前景, 且其蕴涵了知识挖掘与表示、语言理解与计算、答案推理和生成等关键科学问题, 因而受到工业界和学术界的广泛关注. 首先介绍常识问答领域的主要数据集; 其次, 归纳不同常识知识源在构建方式、常识来源和表现形式上的区别; 同时, 重点分析并对比前沿常识问答模型, 以及融合常识知识的特色方法. 特别地, 根据不同问答任务场景中常识知识的共性和特性, 建立包含属性、语义、因果、语境、抽象和意图 6 大类的知识分类体系. 以此为支撑, 针对常识知识数据集建设, 感知知识融合和预训练语言模型的协作机制, 以及在此基础上的常识知识预分类技术, 进行前瞻性的研究, 并具体报告上述模型在跨数据集迁移场景下的性能变化, 及其在常识答案推理中的潜在贡献. 总体上, 包含对现有数据和前沿技术的回顾, 也包含面向跨数据知识体系建设、技术迁移与通用化的预研内容, 借以在汇报领域技术积累的前提下, 为其理论和技术的进一步发展提供参考意见.

关键词: 常识问答; 常识知识源; 常识知识类型

中图法分类号: TP391

中文引用格式: 范怡帆, 邹博伟, 徐庆婷, 李志峰, 洪宇. 常识问答研究综述. 软件学报, 2024, 35(1): 236–265. http://www.jos.org.cn/1000-9825/6913.htm

英文引用格式: Fan YF, Zou BW, Xu QT, Li ZF, Hong Y. Survey on Commonsense Question Answering. Ruan Jian Xue Bao/Journal of Software, 2024, 35(1): 236–265 (in Chinese). http://www.jos.org.cn/1000-9825/6913.htm

Survey on Commonsense Question Answering

FAN Yi-Fan[1], ZOU Bo-Wei[2], XU Qing-Ting[1], LI Zhi-Feng[1], HONG Yu[1]

[1](School of Computer Science & Technology, Soochow University, Suzhou 215006, China)

[2](Infocomm Research Institute of Singapore, Singapore 138635, Singapore)

Abstract: Commonsense question answering is an essential natural language understanding task that aims to solve natural language questions automatically by using commonsense knowledge to obtain accurate answers. It has a broad application prospect in areas such as virtual assistants or social chatbots and contains crucial scientific issues such as knowledge mining and representation, language understanding and computation, and answer reasoning and generation. Therefore, it has received wide attention from industry and academia. This study first introduces the main datasets in commonsense question answering. Secondly, it summarizes the distinctions between different sources of commonsense knowledge in terms of construction methods, knowledge sources, and presentation forms. Meanwhile, the study focuses on the analysis and comparison of the state-of-the-art commonsense question answering models, as well as the characteristic methods fusing commonsense knowledge. Particularly, based on the commonalities and characteristics of commonsense knowledge in different question answering task scenarios, this study establishes a commonsense knowledge classification system containing attribute, semantic, causal, context, abstract, and intention. On this basis, it conducts prospective research on the construction of

* 基金项目: 国家重点研发计划 (2020YFB1313601); 国家自然科学基金 (62076174, 61836007)
收稿时间: 2022-10-18; 修改时间: 2022-12-29; 采用时间: 2023-02-03; jos 在线出版时间: 2023-08-09
CNKI 网络首发时间: 2023-08-10

commonsense knowledge datasets, the collaboration mechanism of perceptual knowledge fusion and pre-trained language models, and corresponding commonsense knowledge pre-classification techniques. Furthermore, the study reports specifically on the performance changes in the above models under cross-dataset migration scenarios and their potential contributions in commonsense answer reasoning. On the whole, this study gives a comprehensive review of existing data and state-of-the-art technologies, as well as a pre-research for the construction of cross-data knowledge systems, technology migration, and generalization, so as to provide references for the further development of theories and technologies while reporting on the existing technologies in the field.

Key words: commonsense question answering; common sense knowledge source; common sense knowledge type

“感知智能”向“认知智能”转化是人工智能最新的发展趋势. “感知智能”是指机器具备视觉、听觉和触觉等感知与加工能力, 比如人脸识别、语音识别等[1]. 相比而言, “认知智能”是从类脑研究和认知科学的角度出发, 结合跨领域的知识图谱[2], 因果推理[3]和主动学习[4]等技术, 赋予机器类似人类的思维逻辑和认知能力, 尤其是理解、归纳和应用知识的能力. 其中, 智能问答是“认知智能”的典型案例之一, 而常识问答 (commonsense question answering, CQA) 是以常识知识为认知基础的智能问答关键技术. 相关技术产出已在苹果 Siri 语音助手、谷歌智能助理、阿里小蜜和微软小冰社交机器人等工业产品中得以应用.

CQA 的任务定义是: 给定特定自然语言问句, 机器结合已有常识知识或其自助挖掘技术, 实现答案求解. 求解过程可为判别式, 也可为生成式. 其中, 判别式 CQA 进一步细分为多项选择和正误判断. 前者旨在基于问题理解和语段阅读理解, 结合常识知识, 从包含正确答案的选项集合中选择“符合答案特性”的正确答案; 后者基于给定文本的理解以及与其相关的常识知识, 判断该文本表述的内容是否正确. 生成式 CQA 则不依赖上下文, 实现答案文字片段的自动生成. 表 1 与表 2 分别给出了判别式 CQA 和生成式 CQA 的样例.

表 1　判别式常识问答的样例

分类	属性	样例
多项选择	数据集	Social IQA
	上下文	Sasha spent time with their kids and they played video games all day long. (萨沙和孩子们一起玩了一整天的电子游戏.)
	问题	How would Sasha feel afterwards? (萨沙之后会有什么感觉?)
	选项	A) bored (无聊的) B) happy (高兴的) C) conflicted (矛盾的)
	答案	B) happy (高兴的)
正误判断	数据集	CommonsenseQA2.0
	问题	The end of a baseball bat is larger than the handle. (棒球棒的末端比把手大.)
	答案	Yes (是的)

表 2　生成式常识问答的样例

数据集	ProtoQA
问题	Name something that an athlete would not keep in her refrigerator. (说出运动员不会放在冰箱里的东西.)
答案	unhealthy food (不健康的食物) (36): chocolate (巧克力), junk food (垃圾食品),… unhealthy drinks (不健康的饮料) (24): coke (可口可乐), alcohol (酒),… clothing/shoes (衣服/鞋子) (24): gloves (手套), clothes (衣服), shoe (鞋子),… accessories (配件) (7): handbag (手提包), medal (奖牌), tennis (网球),…

CQA 与一般的自动问答系统 (如开放域自动问答、知识库自动问答、社区自动问答) 的区别是答案来源不同. 前者的答案来源通常是常识知识库, 而后者的答案来源于互联网资料、知识库或历史问答对数据. 其共同点是模型均需要对给定问题以及答案来源之间建立推理机制, 以求解正确答案. 特别地, 知识库问答与 CQA 存在较大区别. 首先, 知识库问答的答案来自知识库, 而 CQA 的答案需要对常识知识库的信息做深层推理, 同时, 后者涉及的知识库往往更为抽象, 如一种表示概念关系的图谱 ConceptNet[5]; 其次, 知识库问答研究的问题一般针对知识库

中已有的实体和关系, 而 CQA 涉及的问题通常更为开放, 无法仅依赖模式相对固定的知识库来求解.

目前, CQA 已经获得了广泛的研究, 在数据建设、任务设置与更新、关键技术突破方面, 都取得了重要成果. 在数据建设方面, 现有 CQA 的权威数据集数量达 12 种, 数据来源涉及 9 个领域, 包括社交媒体、自然科学、日常生活等. 在任务设置的多样性方面, 现有 CQA 研究方向可细分为常识知识源构建、常识知识获取、知识融合推理和可解释性生成共计 4 个主干子方向. 在关键技术突破方面, 相关研究已从传统的基于规则和统计的方法, 以及前期利用循环神经网络 (RNN)[6]、长短期记忆网络 (LSTM)[7]和注意力机制[8]的中小型神经网络常识知识问答模型, 过渡到近期基于预训练语言模型 (如 BERT[9]、RoBERTa[10]、BART[11]、GPT3[12]和 T5[13]) 的大型神经 CQA 技术, 以及一系列结合经验发现和认知原理的特色技术[14,15]. 相关工作在深度语义理解、知识挖掘与应用、问答关系线索感知, 以及智能答案推理与生成等关键问题上, 形成了一批出色的技术产出[16−19].

本文综述了上述技术发展现状, 并对现有研究热度较高的权威数据集, 以及相对应的 CQA 任务特色 (构建方式、知识源和表现形式) 和常识知识类型进行了介绍, 以此推动学术界和工业界同行进行“精准”的课题定位与技术实践. 特别地, 本文提供了一项小型的专题介绍, 围绕该领域的技术攻坚重点, 深入探讨“基于大型预训练语言模型进行常识知识融合”的 CQA 技术, 对融合方法、推理机制、知识源、知识挖掘线索的内在关系进行了详细分析. 在此基础上, 本文通过系统性的实验, 分析验证了上述知识融合技术和预训练语言模型的适应性, 特别是对不同 CQA 数据源和常识知识类型的适用性, 以此为未来相关研究提供基线.

本文第 1 节介绍 CQA 任务的研究现状, 包含了早期基于规则和统计技术的传统方法, 以及近期利用神经网络架构和预训练语言模型的前沿方法. 第 2 节描述了近 5 年 CQA 任务常用的 9 套数据集, 包括 Commonsense QA[20]、Openbook QA[21]、ARC[22]、Social IQA[23]、Cosmos QA[24]、MCScript[25]、MCScript2.0[26]、ReCoRD[27]和 ProtoQA[28]. 第 3 节从构建方式、知识来源和表现形式这 3 个方面对不同的常识知识资源进行归纳与对比, 并将 CQA 数据集所需的常识知识分为属性、语义、因果、语境、抽象和意图这 6 种类型. 第 4 节为小型专题, 其从方法设计角度, 对目前大型 CQA 模型与常识知识融合方法进行分析和对比. 第 5 节是对第 3 节和第 4 节内容的扩展, 侧重提供实验和量化数据, 借以反映现有主要 CQA 技术在不同数据集和常识知识类型上的适用性, 纳入实验的 CQA 技术具有一种感知常识知识的语言模型框架, 其在现阶段具有较高的代表性和前沿性. 第 6 节在总结全文的基础上, 剖析了目前 CQA 任务存在的难点、发展趋势和未来的挑战.

1 CQA 技术现状回顾

传统问答任务主要测试模型的语义理解和推理能力, 通常根据给定的上下文寻找问题的答案. 当给出的问题超出模型的认知范畴时, 其难以预测出正确答案. 因此, 考虑外部知识或世界知识的 CQA 任务逐渐引起了学者的广泛关注. 目前, CQA 研究已历经两个技术发展阶段, 早期研究主要围绕基于规则和特征工程的方法展开, 初步结合外部知识库或互联网知识挖掘技术, 通过常识知识提取和简易推理手段, 实现答案推荐或判定. 从 2018 年开始, 随着神经网络模型的进一步发展和预训练语言模型的提出, 研究者们广泛地将神经网络和预训练语言模型应用于 CQA 任务, 形成了基于深度语义理解与表示, 以及知识结构和关系建模的神经 CQA 研究流派. 下面分别对两个研究阶段中的代表性工作给与回顾 (各项技术[29−37]的详细回顾见表 3).

表 3 CQA 技术现状回顾

模型	设计特色	外部知识来源	优点	数据集
Feature-based[29]	建立8类特征工程	FrameNet	基于特征工程, 挖掘指代词与问题先行词的相关性	WSC
Rule-based[30]	设定统一的句法学习模式、设计启发式规则	WordNet	引入WordNet, 通过启发式规则获取符合关系元组模式的常识知识	WSC
Attention-based[31]	引入循环神经网络	无	以候选答案的浅层特征表示为基准, 计算上文的注意力分布	SCT
Self-talk[32]	引入GPT与GPT-2	ConceptNet、Google Ngrams、COMET	经过常识知识库的预训练, 为常识性问题提供可解释性线索	COPA、CSQA、SIQA、WSC

表 3 CQA 技术现状回顾 (续)

模型	设计特色	外部知识来源	优点	数据集
Ernie 3.0[33]	基于知识图谱设计预训练任务	知识图谱、百科知识	具备通用性, 在多个自然语言理解与生成任务上均适用	WSC
Knowledge driven-based[34]	实现面向外部常识知识源的封闭域预训练	Atomic、ConceptNet、WordNet、Visual Genome和Wikidata	通过数据增强, 提升预训练语言模型的常识知识感知能力	CSQA、SIQA
KagNet[35]	引入N-gram、长短期记忆网络	ConceptNet	采用直接“知识灌输”模式, 形成综合常识知识与问答语义的联合表示	CSQA
Heterogeneity feature-based[36]	加载异构知识源	ConceptNet、Wikipedia	产生结构化知识与非结构化知识的融合特征表示	CSQA
Headhunter[37]	为常识知识分配不同的注意力分数	OMCS	提供高质量的常识知识表示	CSQA

注: “数据集”所在列仅罗列本文调研范畴内的常识问答数据集

1.1 基于规则和特征工程的传统 CQA 方法

智能问答系统的兴起, 使得研究者们开始关注机器是否具备常识推理能力. Roemmele 等人[38]于 2011 年设计了 COPA (choice of plausible alternatives) 数据集, 它是第一个与 CQA 任务相关的数据集 (早期并没有专门针对 CQA 任务的研究). 该数据集提出了一项因果推理任务, 其每个样本均包含一个前提句子和两个候选答案. 构建该任务的模型需要利用因果推理技术方案, 选择正确的答案. 随后, Levesque 等人[39]在 2012 年建设了 WSC (Winograd schema challenge) 数据集, 该数据集主要检验 Winograd 模式[40]的常识指代消解问题. 其中, Winograd 模式的强混淆性使得 WSC 一度成为图灵测试的替代方案.

因此, 基于 WSC 的早期研究较为密集. 其中, 代表性工作来自 Rahman 等人[29]. 其指出 WSC 的高混淆性主要体现在指代词与问题先行词之间的模糊关系, 并由此专门建立了 8 类语言学特征类 (框架语义、常识归属、情感极性等) 的识别和表示方法, 以及基于支持向量机的二元相关性分类 (指代词与问题先行词高相关和低相关的二元性), 从而支撑了针对高混淆性测试样本的有效处理. Rahman 等人的 8 类特征工程中, 常识归属是一类表示实体和属性关系的特征. Rahman 等人建立了严密的查询构造方法和启发式关系诊断规则, 从而实现了基于网络数据的常识归属信息挖掘与表示.

Emami 等人[30]认为 WSC 的实例均具有相似的句法结构, 可通过设定统一的句法学习模式, 检测其中蕴涵的常识知识, 进而推理出目标代词与正确先行词之间的关系. Emami 等人基于 CoreNLP 分析句法结构, 利用通用的结构化模板将问题实例转换成关系元组模式 (两个候选先行词、上下文谓词、目标代词与查询谓词), 并采用基于词典 WordNet 的查询构造和扩展方法, 支持常识知识的实时检索, 特别是获取和推荐符合关系元组模式的外部数据. 这一工作采用了启发式规则, 侧重利用常识知识的元素和结构进行文本匹配、评分和关系推理, 以此支持问答过程中的指代消解和混淆消减.

上述方法具有高效率且复现简单的优势, 但由于人工制定的规则模式较单一, 且特征工程对专家经验有着较高依赖性, 使得这类方法无法在更大规模和问题多样性更宽泛的 CQA 数据上表现出较高的鲁棒性和适应性.

1.2 基于神经网络的 CQA 技术

面向 CQA 投入神经网络和深度表示学习的工作, 在最初阶段面临标记数据缺失的窘境. 相关研究仅在任务较为类似的其他数据集上予以展开. 比如, Mostafazadeh 等人[41]基于“故事”文本建立的选择填空测试集 SCT. 该数据集考察模型的语义理解以及推理能力, 测试样本围绕日常生活故事展开, 推理目标是在给定故事的开头句子和两个选项的情况下, 选择其中一个作为故事的结局. 针对这一数据集, 研究人员将问答架构解释为一种条件 (前文) 和关联性更强的答案 (结局) 推理, 近似于当前 CQA 的任务模式. 由此, 研究焦点集中于条件的语义表示, 以及二选一的答案预测方法. 相应地, 研究人员将循环神经网络[6]、长短期记忆网络[7]以及注意力机制[8]引入表示学习,

且利用非线性解码层进行答案预测. 特别地, Cai 等人[31]以候选答案的浅层特征表示为参照, 计算上文的注意力分布, 借以重塑条件和候选答案的联合表示, 以此评估问答信息的关联强度, 以及正解的预测.

2018 年是 CQA 技术的重要分水岭, 一系列精细加工且领域丰富的数据集陆续释出, 如 SWAG[42]. SWAG 侧重检验常识推理能力, 问答样本以“符合自然逻辑的常识关系”为根本, 以单项选择或多项选择题的形式予以呈现, 推理目标是正确推断符合常识关系的选项. 此外, 知识结构更为健全或领域特色更为鲜明的其他数据集也相继出现, 包括 CommonsenseQA[20]、Openbook QA[21]和 Social IQA[23]等. 这类数据集引入的 CQA 任务更富于挑战 (第 2 节将对上述 CQA 数据及其特色予以详细介绍). 这类挑战主要体现于对“广域常识知识的认知能力”和“适用于多样语用形式的统一语义理解模式”的高要求. 也因此, 预训练语言模型逐步成为突破 CQA 技术壁垒的重要手段, 其在知识储量丰富且语用多样性较高的大规模通用数据上进行预训练, 由此得到的感知、表示和生成能力, 对提升 CQA 的性能水平有着重要的支撑作用[9].

近期, 基于 Transformer[43]架构的 BERT、GPT[44]或 XLNet[45]等预训练语言模型在自然语言处理领域取得了广泛应用. CQA 领域也随之迎来了基于大模型的迁移研究时期. 尽管如此, 最初的研究往往简单地对预训练模型进行直接迁移和参数微调, 并未结合 CQA 的任务特色和数据分布实现模型重构和优化, 特别是对常识知识理解与关系识别, 以及基于常识的推理和计算机制, 都没有形成任务相关的代表性工作. 为解决该问题, 大量研究开始探讨更加符合 CQA 任务需求的研究方案, 尤其集中研究了如何在充分利用预训练语言模型的通用语义理解和表示能力的基础上, 结合常识知识推理正确答案的方法. 相关研究可分为两类, 分别是感知常识知识和加载常识知识的方法, 下面分别予以介绍.

(1) 感知常识知识的方法

该类方法的特点是, 通过面向常识知识源的机器学习, 促使 CQA 模型从根本上具备感知语言中常识知识的能力. 具体地, Shwartz 等人[32]认为预训练语言模型能够对大规模数据中蕴涵的通用知识进行表示学习, 从而对其获得感知能力. 因此, Shwartz 等人将预训练语言模型引入 CQA 领域, 并基于无监督的方式提出了 self-talk 框架. 该框架利用“自问自答机制”实现可解释性常识线索的感知, 并将这类线索的表示作为推理最终答案的依据. 基于预训练语言模型 GPT 和 GPT-2, Shwartz 等人开发了“自问自答”的问题生成器, 借以引导常识线索的感知, 同时利用 GPT 和 GPT-2 构建了表示模式适配的 CQA 模型, 使之依照常识线索进行答案的求解.

相对地, Sun 等人[33]认为预训练语言模型 GPT 及其变体 (GPT3) 并未完全具备感知常识的能力, 其从大规模预训练数据中获取的认知能力, 更多地集中在通用句法和语义知识, 以及普及率较高的常见知识, 欠缺对特定领域知识图谱中知识结构和关系特征的模式识别与处理. 为此, Sun 等人将知识图谱和百科知识纳入预训练模型的二次学习过程, 即在一种知识关系和表述重塑的任务框架下, 借助掩码语言模型的重新训练, 专门性地锤炼神经网络的常识知识认知能力, 使之对常识知识的高注意力关系特征和语用特征具备识别和表示能力.

Ma 等人[34]则利用多个外部知识源的信息, 以数据增强为主要优化手段, 提升了预训练语言模型的常识知识感知能力. 其关键贡献是建立并利用通用的启发式规则, 在常识知识的三元组 (头实体、尾实体和关系) 与问答样本 (问题、答案和关系) 之间建立了映射关系, 从而支撑了“从外部知识源向 CQA 问答数据集”的迁移学习方法. 具体地, Ma 等人引入了 Atomic[46]、ConceptNet[5]、WordNet[47]、Visual Genome[48]和 Wikidata[49]多种常识知识源, 并建立上述映射体系和实例, 且利用随机抽样和生成模型批量地制造干扰项. 在此基础上, Ma 等分别对基于 RoBERTa 或 GPT-2 的 CQA 模型进行迁移学习, 实现了面向外部常识知识源的封闭域预训练, 以及面向 CQA 问答样本的二次微调, 逐层地增强了 CQA 模型的任务依赖的学习力度与深度, 从而为其感知常识和正确推理提供了辅助.

(2) 加载常识知识的方法

该类方法的特点是面向外部或任务相关的常识知识源 (如 ConceptNet 中蕴涵的知识三元组) 进行独立的表示学习, 并在此基础上, 将知识表示纳入 CQA, 使之与 CQA 的编码结果进行合并、融合或扩展, 从而形成蕴涵常识信息的编码表示, 由此支撑 CQA 的答案解码过程. 具体地, Lin 等人[35]利用 ConceptNet 设计实现了基于知识图谱的 CQA 模型 KagNet. 该模型通过 N 元语法 (N-gram) 提取问答文本中的实体, 并在 ConceptNet 上搜索

与该实体信息相关, 同时距离不超过 4 跳的知识路径 (含节点和边). 在此基础上, Lin 等人借助卷积神经网络直接对知识路径上的实体知识进行编码形成知识结构的特征表示, 并通过拼接 LSTM 产出的问题和候选答案的文本特征表示, 获得综合了常识知识和问答语义的联合表示. 实验验证, 这种直接的“知识灌输”模式能够对 CQA 性能产生优化.

此外, Lv 等人[36]发现特定常识概念 (如实体) 在知识图谱中的拓扑结构特征, 及其在文本 (如蕴涵答案的语段) 中的上下文语义特征, 都有助于面向常识概念的深度理解, 也显然能够支撑 CQA 的问题理解和答案求解. 因此, Lv 等人提出一种异构信息源 (知识图结构和自由文本) 的特征结合方法. 具体地, 基于给定的问答文本, Lv 等人从 ConceptNet[5]中提取常识知识路径, 并通过语义角色标注从 Wikipedia 句子中提取三元组, 以此构造一个融合多种常识知识源信息的子图. 同时, Lv 等人构建了一个图推理模型, 其包含基于上述子图的上下文表示学习和基于图的异构信息融合. 前者利用子图结构重新定义问答样本中单词之间的距离, 以便更好地学习上下文表示; 后者采用图卷积网络将邻居信息与节点的表示融合. 由此形成的最终表示, 将被用于支撑解码模型的答案推理. Li 等人[37]发现将大量常识知识引入 CQA 模型时, 模型无法克服噪声的干扰, 对相关知识的甄别与理解出现偏差. 为此, Li 等人设计了 Headhunter 模块, 旨在基于知识检索模块提取的多条常识知识, 过滤其中包含的噪声, 提高 CQA 模型加载的常识知识的质量. 其中, Headhunter 采用自注意力直接编码 ElasticSearch 检索的结果 (即 OMCS [50]语料库中与问答文本相关的常识知识)、问题以及候选答案, 由此获得蕴涵常识信息的分布式表示, 并引入注意力池化层, 对多条知识分配不同的注意力分数, 从而强化相关知识的表示, 过滤无关的噪声. 结果表明, Headhunter 模块为 CQA 模型提供了高质量的常识知识表示, 增强了该模型的常识推理能力.

上述介绍中, 采用感知常识知识的方法, 充分“挖掘”自然语言文本中隐含的常识性语言现象, 使得 CQA 模型具备较强的常识感知与表示能力; 加载常识知识的方法, 直接在语义表示中融入常识信息, 增加了 CQA 模型表示的知识量, 为答案求解过程提供了额外可解释的参照信息. 尽管如此, 现有前沿 CQA 研究仍然欠缺通用的知识获取和理解方法, 针对这一问题 (通用常识获取为前提的高适应性 CQA) 的讨论将在第 4 节给与深入讲解.

1.3 CQA 技术面临的挑战

上述相关研究已经极大推动了 CQA 领域的技术进步, 并在各类任务场景中取得了显著的性能优化. 尽管如此, 形成高可靠性的 CQA 系统仍面临诸多挑战, 包括如下几个方面.

(1) 常识知识源通常以自然文本或三元组的形式呈现, 其来源具有多样性, CQA 系统获取常识知识仍具有难度. 由于常识知识具有隐含性、多样性和开放性, CQA 模型基于给定的问题, 从常识知识源中精确且全面地获取常识知识仍十分困难.

(2) 虽然利用预训练语言模型的语义感知和泛化学习能力, 有助于融合常识知识, 并借此提升答案推理性能. 但是, 相应技术仍难以理解深藏于知识结构之内的关系语义. 特别是在异构常识知识源与自然语言的对齐问题上, 以及利用常识知识源的完备结构方面, 现有研究仍然面临挑战.

(3) GPT3[12]或 T5[13]等参数量达千万级甚至亿级以上的预训练模型, 在 CQA 任务上获得的性能仍无法达到实用水平. 该现象尚未推动关于“极深语义计算是否能够主导常识知识理解和答案推理”的研究与讨论. 此外, 在利用单一评价标准 (如准确率) 进行可靠性界定的情况下, CQA 领域事实上欠缺一种知识理解过程的可解释性评价标准.

(4) 最后, 常识知识库的容量虽然得到快速扩容 (比如, ConceptNet[5]包含数百万个常识知识或概念), 但配套的挖掘技术并未随之增强. 如何面向特定常识问题, 精准且快速地检索相关常识知识与关系, 恰是当下有待快速解决的关键问题之一.

总体上, 围绕预训练语言模型展开 CQA 研究已成为基本范式. 不同预训练语言模型在训练阶段感知的信息存在差异. 同时, 其自身的常识推理能力仍处于探索阶段. 将此类模型应用于 CQA 任务上时, 尽管获得了较为显著的性能优化, 但其推理过程的可解释性通常并不“完全透彻”难以进一步分析该模型如何利用常识知识 (自身蕴涵或源于外部常识知识源) 推理出正确答案的根本原理. 本文认为, 对于 CQA 任务, 显式地分析常识知识与给定问题的关系至关重要, 其常识知识的选取, 以及常识知识与问答文本的联合特征表示是直接影响模型常识感知与推理

的重要因素.

2 面向常识知识的问答数据

为了探索 CQA 数据集如何考察模型的常识推理能力, 本文对该类数据集进行了广泛调研. 我们从数据集、任务形式、选项数量、背景知识和数据规模 5 个方面对 9 套 CQA 数据集进行统计和分析, 具体如表 4 所示.

表 4 常识问答数据集

数据集	任务形式	选项数量	背景知识	数据规模
CommonsenseQA[20]	多项选择	5	无	12 247个常识问题
Openbook QA[21]	多项选择	4	无	5 957个初级学科问题, 1 326个核心科学事实
ARC[22]	多项选择	4	无	挑战集(2 590个), 简单集(5 197个), 14M个与任务相关的科学语料
Social IQA[23]	多项选择	3	有	37 588个问题
Cosmos QA[24]	多项选择	4	有	35 600个基于常识的阅读理解的大型数据集
MCScript[25]	多项选择	2	有	110个场景, 2 119个文本, 14 000个问题
MCScript2.0[26]	多项选择	2	有	200个场景, 3 487条文本, 19 821个问题
ReCoRD[27]	抽取式	—	有	120 000个示例
ProtoQA[28]	生成式	—	无	9 762个问题

CQA 数据集分为多项选择、抽取式和生成式 3 种任务形式. 采用不同任务形式的 CQA 数据集旨在从多个角度测试模型的常识推理能力. 例如, 采用多项选择形式的 CQA 数据集侧重考察模型分析问题和候选答案之间关系的能力. 模型依据分析结果获取常识知识, 结合问答文本的特征推理出正确答案. 抽取和生成式的 CQA 数据集主要考验模型依据问题或上下文获取相关常识知识的能力. 其中, 抽取式任务 (如 ReCoRd) 要求 CQA 模型理解给定问题和上下文, 并通过常识推理模块从上下文中抽取实体信息作为答案. 生成式 CQA 任务 (如 ProtoQA) 要求 CQA 模型以问题为中心, 结合外部常识知识源, 有序生成多个答案.

从选项数量的角度分析, 除 ReCoRd 和 ProtoQA 外, 其余 7 套 CQA 数据集均包含数量不一致的选项. 选项数量的多少与该任务的推理难度相关. 以 CommonsenseQA 与 MCScript2.0 为例, 前者的选项数量为 5 个, 后者的选项数量为 2 个. 从或然率的角度思考, 正确回答 MCScript2.0 的概率比 CommonsenseQA 高出 30%; 其次, 面向 CommonsenseQA 数据进行处理时, CQA 模型需要分析问题与 5 个候选答案之间的关系, 对 MCScript2.0 而言, 模型仅需考虑问题与 2 个候选答案之间的关联, 复杂度并不一致.

以信息量为视角进行分析, 如果数据集未提供背景知识或禁用背景知识, 则 CQA 模型推敲答案的难度较高. 换言之, 背景未知与已知情况下, CQA 面临的挑战截然不同. 背景知识在现有 CQA 任务中多表现为上下文, 辅助 CQA 模型深度理解问题的语义, 为其提供获取常识知识的线索. 特别地, Openbook QA 与 ARC 均提供与自身任务相关的科学事实或语料库, 用于降低 CQA 模型获取常识知识的难度.

值得赘述的是, 现阶段的主要工作密集地应用神经网络搭建 CQA 模型. 然而, 由于该类模型复杂度较高, 待学习的参数量较多, 从而对数据集的规模有要求 (数据量不足将直接导致低资源 CQA 场景). 对比表 4 中 9 套 CQA 数据集的规模, 均达到 10k 以上 (除 ProtoQA 外), 可满足深度神经网络模型训练的需求. 这些 CQA 数据集均从不同角度评估模型是否具备常识推理能力, 以下根据其各自特点分别进行介绍 (读者可根据经验有选择地跳过熟识的数据集进行阅读).

(1) CommonsenseQA: 真实场景中, 人们回答问题往往需要结合已习得的知识, 建立推理依据, 如空间关系、因果关系、科学事实、社会习俗的常识或特定领域的背景知识. 为了研究基于先验知识的问答, Talmor 等人[20]提出了 CommonsenseQA (CSQA). 在 CSQA 的构建过程中, 标注人员基于 ConceptNet 自由构造大规模常识性问题. 每个问题都需要将 1 个目标概念 (正确答案) 和相关于该目标概念的其他 3 个概念 (候选答案) 区分开. 除此之外, 标注人员自行添加一个与问题相关, 但人们容易忽略的干扰项. 这使得面向 CSQA 的问答模型, 需基于常识知识

分析概念之间的关联性, 以此推敲更为贴切的答案.

(2) Openbook QA: 基于文档或知识库建立的问答数据集, 通常用于验证模型的语言理解能力, 而未研究模型在理解问题语义的基础上, 推理该信息的能力. 为此, Mihaylov 等人[21]参考开卷考试的考核目标, 提出了 Openbook QA (OBQA). 特别地, 回答该数据集的问题需结合物理、化学、生物、天文学等学科类常识知识. 同时, OBQA 提供了一个包含 1 326 个科学事实的语料库, 以测试模型根据科学事实进行推理的能力. Mihaylov 等人希望学者们在研究中考虑纳入常识知识源, 而不仅仅依赖于给定的语料库展开推理. 于是, 如何结合外部常识知识源是深入研究该数据集的主要难点之一.

(3) ARC: 一部分问答数据集集中于检索式任务, 模型仅从给定的上下文与问题的语义匹配可得出答案, 并未引入推理、常识知识或其他深入理解文本语义的方法. 为此, Clark 等人[22]提出了 ARC, 其旨在测试 CQA 模型深入理解问题并进行常识推理的能力. 该数据集来自中小学生考试中的科学问题, 分为 ARC 挑战集 (ARC-Challenge) 和 ARC 简单集 (ARC-Easy). 其中, ARC-Challenge 主要包含基准模型回答不正确的问题 (基准模型为基于检索算法和单词共现算法的模型). 此外, ARC 提供与问题相关的科学事实, 用于辅助 CQA 模型结合事实建立答案预测机制. 其中, 95% 的科学事实与 ARC-Challenge 的问题相关, 但不足以回答 ARC 中所有的问题. 值得注意的是, ARC 与 OBQA 存在类似的部分, 即两者均包含以事实类常识为中心进行答案求解的样本. ARC 与 OBQA 不同的不同之处在于, 前者的知识来源涵盖的领域不如 OBQA 广泛.

(4) Social IQA: 社交常识能够帮助人们推测他人的精神状态或可能的行为. 为了研究社交情景中蕴涵的常识推理, Sap 等人[23]提出一个大规模社交情景推理数据集 Social IQA (SIQA). 该数据集旨在测试 CQA 模型的社交和情绪感知能力, 并对人物隐含的情绪或行为进行推理. 该数据集主要的难点包括两个方面. 其一, 如何利用给定的上下文, 深入分析其与问题之间的联系; 其二, 依据问题蕴涵的线索, 如何获取与之相关的社交类常识知识, 并设计常识推理模型. 此外, 该数据集中, 候选答案之间的语义混淆性, 以及不同问题之间的相似性, 进一步增加了研究该数据集的难度. 目前, 该数据集刚刚释出不久, 其常识知识的选择另辟蹊径, 且符合舆情分析和电子商务, 具有科学研究和实际应用的广泛价值. 因此, 本文推荐科技人员对其予以研究和攻关.

(5) Cosmos QA: Cosmos QA[24]侧重在人们的日常叙述文本中, 给出有关事件的原因或其造成影响的问题, 考察 CQA 模型对跨越上下文的文本片段进行推理的能力, 即跨篇章推理的能力. Cosmos QA 的上下文来源于 Spinn3r 博客数据集[51]的个人叙述样本, 其收集了多种日常生活类的常见事件描述. 该数据集的样本中包含大量“What might be the possible reason of …? (译文: 可能的原因是什么?)”和“What would have happen if…? (译文: 如果…会发生什么?)”类型的问题表述. 回答这类问题需要 CQA 模型跨越不同上下文片段, 定位离散分布的推理线索, 且全部环节 (线索挖掘, 线索关联性感知和答案推理) 都对常识知识具有较强依赖性.

(6) MCScript: 解决自然语言的歧义问题对于机器深度理解文本语义至关重要. 基于这一思考, Ostermann 等人[25]提出一个利用脚本知识评估机器理解能力的 MCScript 数据集. 其中, 脚本知识被定义为关于日常活动的知识, 也被称为场景. 场景是描述典型人类活动的事件, 对其内容的理解有助于 CQA 模型参透口语化叙述中指代不明的问题. MCScript 的构建来源于 InScript 语料库[52](围绕日常情境). 该数据集包含基于文本与脚本知识两类问题, 前者主要考察 CQA 模型对文本与问题的语义理解能力, 后者在前者基础上, 测试 CQA 模型对问题蕴涵脚本知识的理解以及推理能力.

(7) MCScript2.0: 该数据集在上述 MCScript 的基础上增加了 90 个场景, 并扩展了已有场景的问题, 重点考察 CQA 模型对脚本事件 (日常活动中典型的时间顺序) 和参与者 (日常活动中发挥作用的人或对象) 的理解以及推理能力[26]. MCScript2.0 中, 52% 的问题需要 CQA 模型结合日常活动中的常识知识辅助推理, 38% 的问题仅依赖于 CQA 模型对叙事文本的语义理解, 10% 的问题侧重考验 CQA 模型是否具备同时理解文本的语义, 以及对日常活动推理的能力. 类似地, MCScript2.0 也包含基于文本与脚本知识两类问题, 在基于脚本知识类的问题占比方面, MCScript2.0 高于 MCScript.

(8) ReCoRD: 现有数据集缺乏多语句依赖的 CQA 样本, 即推理线索附着于多条语句 (而非一条语句或局部片

段) 的样本, 从而无法考察 CQA 模型是否能够在理解多语句整体含义的基础上捕获答案. 基于此, Zhang 等人[27]提出了 ReCoRd 数据集. 该数据集中, 每个问题由上下文实际支持的语句构成, 语句中的 X 表示缺少的命名实体. ReCoRd 要求 CQA 模型深度理解上下文的语义, 结合常识知识从上下文中推理出最适合 X 的实体作为问题的答案. 特别地, 考察该模型基于上下文的常识推理能力的样本占 75%, 仅有 10% 的样本可通过上下文自身的线索给出答案. 该数据集将机器阅读理解与常识推理结合, 相关的 CQA 模型首要攻克的难点是深度理解上下文以及问题的语义, 并结合常识知识以及多句推理, 从上下文中抽取回答问题的实体信息.

(9) ProtoQA: 目前, 大多常识问答数据集 (如 CSQA) 采用多项选择的任务形式, 这些候选答案为 CQA 模型提供了部分的推理信息, 模型可通过对选项进行排序来预测答案. 为排除候选答案对 CQA 模型推理能力的影响, Boratko 等人[28]提出生成式 CQA 数据集 ProtoQA. 该数据集要求模型根据问题生成多个有序排列的正确答案. 例如, 对于问题“说出人们在离开家去工作之前通常做什么?”由于每个人的生活习惯不同, 此问题包含多个正确答案 (如关闭电源、装好钥匙或锁门). 该数据集旨在测试模型在理解问题语义的情况下, 结合常识知识提供多个有效答案.

3 常识知识来源与类型

常识知识被定义为大多数人共享的世界知识, 在人们的日常交流中, 其作为背景信息, 用于填补自然语言中的“留白”, 帮助人们在交流过程中达成共识[53]. 由于常识知识的隐含性与多样性, 研究者们将常识知识源定义为一种辅助常识提取的多模式存储库[54]. 常识知识来源有多种形式, 并涵盖了不同类型的知识. 本节统计了一组具有代表性的常识知识源 (包括涵盖常识知识的知识库或词典). 此外, 本节探讨了 CQA 数据集中, 问题与答案之间所需的常识知识类型, 辅助未来研究选择合适的常识知识源进行研究.

3.1 常识知识来源

现有工作研究了多种类型的常识知识源, 本节在 Ilievski 等人[54]工作的基础上, 选取 21 个与自然语言理解领域相关的常识知识源进行对比分析, 并对其中蕴涵的关系类型数量、表示形式、知识来源以及构建方式进行归纳. 各个常识知识源[5,11,46,48,49,55−72]的统计信息如表 5 所示.

表 5 常识知识源统计

常识知识源	描述	关系类型数量	表示形式	知识来源	构建方式
Cyc[55]	由术语和断言组成, 术语包含概念、关系和实体的定义, 断言用来建立术语之间的关系	50万条术语、700万条断言	形式化知识	维基百科数据	知识工程、人工
ConceptNet[5]	三元组形式的关系型知识, 表现为数据或事物之间的链接	36	常识知识图谱	OMCS语料库、WordNet	众包
Atomic[46]	词汇形式表示事件及其参与者的前后状态	9	常识知识图谱	多种语料库(书籍、维基词典)	众包
Glucose[56]	包含事件、状态、动机和情绪的因果知识	10	常识知识图谱	ROCStories(儿童故事为主)	半自动模板, 众包
WebChild[57]	节点和关系通过消歧作为WordNet表示	20	常识知识图谱	Web信息	自动提取, 人工规范化
Quasimodo[58]	关于物体属性、人类行为和一般概念的常识知识	78 636	常识知识图谱	搜索日志和论坛	自动提取
SenticNet[59]	概念知识和情感知识的知识库[59]	1	常识知识图谱	文本	自动提取, 自动聚合
HasPartKB[60]	一个hasPart语句的知识图谱	1	常识知识图谱	语料库	自动提取, 自动凝练
Probase[61]	IsA语句的概率分类	1	常识知识图谱	语料库	自动提取
IsaCore[62]	ConceptNet和Probase中选择的IsA的知识分类	1	常识知识图谱	ConceptNet、Probase	自动提取

表 5 常识知识源统计 (续)

常识知识源	描述	关系类型数量	表示形式	知识来源	构建方式
Wikidata[49]	一个通用领域的知识图谱	6.7k	通用知识图谱	Wikipedia	自动, 众包
YAGO[63]	一个通用的知识图谱, 节点和关系是消歧的实体	116	通用知识图谱	Schema.org[64]、Wikipedia	自动提取, 自动整合
DOLCE[65]	通过消除概念和关系捕获自然语言和人类常识基础上的高级本体	1	通用知识图谱	—	专家手工创建
SUMO[66]	消除概念及其关系歧义的上层本体	1 614	通用知识图谱	—	专家手工创建
WordNet[48]	基于同义性和反义性来描述词语和概念之间的语义关系类型的词典	10	词汇源	—	专家手工创建
Roget[67]	一个英语单词包含同义词和反义词的同义词典	2	词汇源	—	专家手工创建
FrameNet[68]	形式化框架语义学理论的词汇资源	8	词汇源	语义学理论的词汇资源	专家手工创建
MetaNet[69]	一个概念框架的仓库, 通常采用隐喻来表示关系	14	词汇源	—	人工手动创建
VerbNet[70]	一个描述动词的语法和语义模式的资源	36	词汇源	—	人工手动创建
GenericsKB[71]	以自然发生的句子表示的自包含的通用事实	—	语料库、语言模型	Waterloo、SimpleWiki、ARC corpus	手写过滤+基于BERT的分类器
Language Models	BART[11]和GPT-2[72]等预训练模型	—	语料库、语言模型	—	预训练任务

注: “—”表示某些常识知识源暂未提供具体的关系类型数量以及知识的来源

常识知识源主要用于描述不同对象、实体或事件之间的关系. 表 5 中罗列的常识知识源涵盖的关系类型大致分为实体关系、事件和情感变化、因果关系和词汇关系. 这些关系类型从不同角度细化了常识知识的范畴. 例如, ConceptNet 关注实体概念之间的关系, Atomic 侧重于事件之间的假设推理关系, Roget、FrameNet 和 VerbNet 主要关注词汇级的关系 (如同义词或反义词). 其次, 关系类型数量表示常识知识源中所能提供常识知识的数量. 由表 5 可知, 不同常识知识源涵盖的常识知识数量 (关系类型或以自然语言文本呈现的句子数量) 存在差异. 表 5 中“描述”与“关系类型数量”两个维度的统计结果, 分别提供了不同常识知识源的特点, 以及涵盖常识知识的信息. 相关研究可根据这类信息, 结合 CQA 任务自身的特点, 选择合适的常识知识源, 设计相应的知识融合和机器学习方法, 以此提高 CQA 模型的常识推理能力.

从表 5 中“表示形式”这一列分析, 常识知识源主要分为形式化知识、常识知识图谱、通用知识图谱、词汇源、语料库和语言模型 6 类. 下文将按类别归纳每种类型常识知识源的特点与不足.

(1) 形式化知识类: 基于形式化的谓词逻辑刻画知识. 形式化的优势是可支持复杂的推理. 但是, 过于形式化的知识结构和表示, 容易导致知识源的扩展性低与应用灵活性不足.

(2) 知识图谱类: 知识图谱类的常识知识源通常由蕴涵实体或事件关系的结构化三元组构成. 这类知识源可为 CQA 模型提供相关于特定问题的结构化信息. 目前, 面向知识图谱的普遍应用方式是, 利用图神经网络将结构化三元组转换为图表示, 再与预训练语言模型产生的表示进行融合, 借以提升模型的常识推理能力. 但 CQA 任务均为文本形式的数据集, 使得模型在引入该类常识知识时, 需进一步考虑异构数据的融合推理.

(3) 词汇源或语料库类: 词汇源或语料库类的常识知识源多以文本表示. CQA 模型融合该类常识知识源时, 通常利用文本中的语义信息, 实现语义层面的推理, 而难以直接获取其中蕴涵的结构化信息. 同时, 词汇源以及语料库的来源较为多样, 包含冗余信息, 其质量无法得到保证.

(4) 语言模型类: 语言模型在大规模的开放域文本数据上进行预训练, 可为 CQA 任务提供蕴涵通用知识的上下文表示, 模型可利用该表示展开推理. 然而, 这类模型无法对答案的推理过程提供合理的解释. 此外, 这类模型是否自身蕴涵常识知识, 且是否具备常识推理能力, 目前仍然备受研究人员的质疑. 从而, 大量工作考虑将常识知识源与预

训练生成模型结合, 以探索预训练语言模型是否具备推理新常识知识的能力. 例如, Bosselut 等人[73]提出自适应常识知识生成框架 COMET. 该框架基于 ConceptNet 和 Atomic 常识知识源训练语言生成模型, 在此基础上构建常识知识库. 其中, ConceptNet 和 Atomic 为 COMET 提供可学习的知识库结构和关系, COMET 可动态调整预训练阶段学习的知识表示, 为常识知识图添加新的节点和边, 以此作为新的常识知识. 因此, 语言模型类是一种隐式且不可解释的知识源, 本身不具备特定的结构约束、可见样本和存储模式, 知识蕴藏于预训练语言模型的参数空间和计算模式之中, 呈现为一种极为独特且难以解释的来源. 其优点仅限于动态可更新性, 其缺点是异同的模型结构和数据领域分布, 将形成不可预期的知识源表示形式. 从而, 面向特定任务的迁移学习和微调, 是实践中不可或缺的环节.

此外, 由于常识知识来源广泛, 不同常识知识源在构建方式上存在差异. 由表 5 中“构建方式”可知, 除语言模型外, 其他的常识知识源在构建过程中, 均有人工制定规范或存在常识知识筛选步骤, 说明目前机器在常识知识的自动识别与质量评估方面仍需进一步探索.

3.2 常识知识维度

常识知识维度的划分可辅助研究者们分析不同常识知识源对常识知识的覆盖范围, 以选择合适的常识知识源, 为 CQA 模型提供给定问题所需的常识知识. Ilievski 等人[54]在统计常识知识来源的基础上, 从知识维度层面剖析常识知识源在推理阶段可提供的线索类型, 将其蕴涵的常识知识分为 13 个维度. 表 6 以 13 个知识维度为基础, 归纳每种知识维度的含义以及涵盖该维度的常识知识源.

表 6 常识知识源的知识维度

知识维度	描述	常识知识源
Lexical	名词的复数形式、动词的过去式以及概念与语言表达之间的形式化表述	ConceptNet, Language Models, WordNet
Similarity	表述之间的相似性, 包括同义词和语义相似的表述或概念	ConceptNet, Roget, Wikidata, WebChild, WordNet
Distinctness	可区分性, 比如反义词或内在不相容的关系	ConceptNet, FrameNet, Wikidata, Roget
Taxonomic	一种排列分类, 在此分类中, 一些对象被放置到具有继承关系或更具体的分组中	IsaCore, Probase, Wikidata, WordNet
Part-whole	部分与整体的关系	COMET, ConceptNet, HasPartKB, Wikidata,
Spatial	描述与空间有关的表述	ConceptNet, WebChild, Wikidata,
Creation	描述某物存在的过程	COMET, ConceptNet, Wikidata
Utility	物体在某些目的上的适用性或实用性的概念	COMET, ConceptNet, Wikidata
Desire or goal	关于主体的动机或目标	COMET, ConceptNet, Atomic
Quality	描述主题的属性或与对象相关的性质或功能	Atomic, COMET, ConceptNet, SenticNet, WebChild, Wikidata
Comparative	根据对象属性的相对值与其进行比较	WebChild
Temporal	通过消除概念和关系捕获自然语言和人类常识基础上的高级本体, 比如时间关系	Atomic, COMET, ConceptNet, WebChild, Wikidata
Relational-other	不同概念间的关系或与上下文相关的关系	ConceptNet, Wikidata

上述 13 个知识维度之间具有可区分性, 可为 CQA 模型提供粗粒度的推理线索. 以“Part-whole”和“Spatial”为例, 前者表示部分与整体的关系, 后者表示空间关系. 将这些知识维度映射于不同的常识知识源中, 其表达方式存在差异. 比如, 知识维度为“Part-whole”的数据, 在 ConceptNet 中对应的关系表示为“PartOf”或“HasA”. 而在 Wikidata 中, 与该知识维度映射的关系表示为“Has part”或“Member of”. 虽然同一类关系表示的多样性会增加模型对该类关系的表征能力, 但是当模型引入多种常识知识源时, 如何处理相同关系的不同表示, 以弥合问题与答案之间的关联, 仍是目前 CQA 任务面临的难点. 相比于其他知识维度, Relational-other 知识维度侧重关注上下文之间的关联, 可辅助 CQA 模型深度理解上下文, 展开推理. 目前只有 ConceptNet 和 Wikidata 包含该知识维度.

此外, 在表 6 的“常识知识源”中, 不同的常识知识源涵盖的知识维度差异较大. 比如, ConceptNet 和 Wikidata

包含的知识维度最多达 11 种; 其次, COMET、WebChild、WordNet 和 Atomic 囊括的知识维度不多于 7 种; Roget、FrameNet、Probase、IsaCore、HasPartKB 和 SenticNet 仅涉及一个或两个知识维度. 以上统计信息说明, 不同常识知识源之间在知识维度划分上虽然存在交叉 (如 ConceptNet 和 WebChild 均包含知识维度“Temporal”) 或互补 (如知识维度“Comparative”仅存在于 WebChild 中), 但常识知识源在完备性上却各有不同. 对于 ConceptNet 和 Wikidata 而言, 前者源自 OMCS 语料库和 WordNet, 涉及 36 种关系, 后者源自 Wikipedia, 包含 6.7k 种关系. 这两种常识知识源成为研究者们首选的外部知识库, 相关研究往往受益于它们涵盖的丰富常识知识维度. 特别地, 语言模型类的常识知识源, 其预训练阶段通过学习语料中词汇之间的关系, 蕴涵对词汇级知识维度 (如 Lexical 和 Similarity 知识维度) 的表示和感知能力. 通过预训练的方式是目前研究者们使用 CQA 模型感知常识知识并提升其推理能力的一种常见做法.

3.3 常识知识类型

第 3.2 节的知识维度侧重于围绕某一个独特的自然属性进行常识知识分类, 比如时间 (temporal) 或空间 (spatial) 维度. 本节提及的常识知识类型在上述知识维度划分的基础上, 采用聚合的思想将具有相同特点的常识知识归为一种类型. 例如, lexical (词形)、similarity (相似语义) 和 distinctness (互斥语义) 均从词汇的维度对常识知识进行分类. 本节将三者统一归为“语义”类常识知识. 基于第 2 节对 CQA 数据集的分析, 本节将数据集中常识知识划分为 6 种类型, 分别是属性、语义、因果、语境、抽象和意图. 表 7 展示数据集中问题与正确答案之间所需常识知识类型的人工分析过程. 其中, 实例分别来自 5 套 CQA 数据集 (MCS2.0、CSQA、ARC-C、OBQA 以及 SIQA). 为简化描述, 实例部分只展示问题与候选答案, 不提供样本中提及的上下文、事实或者外部语料, 加粗斜体表示正确答案. 具体地, 本节采用定义与实例分析相结合的方式阐述 6 种常识知识类型. 针对常识知识类的分析如下所示.

表 7 常识问答数据集中实例分析常识知识类型

实例	常识知识类型分析过程	常识知识类型	来源
What is around the plant? A) soil B) water	在种花的场景下, 考虑空间信息, 植物的周围通常存在土壤	属性	MCS2.0
Google Maps and other highway and street GPS services have replaced what? A) united states B) mexico C) countryside D) atlas E) oceans	“Google Maps、GPS”和“Atlas”之间的语义相似性	语义、属性	CSQA
Which factor will most likely cause a person to develop a fever? A) a leg muscle relaxing after exercise B) a bacterial population in the bloodstream C) several viral particles on the skin D) carbohydrates being digested in the stomach	发烧由什么原因导致, 由常识知识可知有可能是因为细菌感染	因果、属性	ARC-C
When did they have their own place and space? A) When in high school B) When the person moved out of the parent’s house	根据上下文可知, 存在时序关系, 由常识知识可得出答案	语境	MCS2.0
Poison causes harm to which of the following? A) a tree B) a robot C) a house D) a car	答案是外部事实中“living things”的实例化. 由常识可知“a tree”属于有生命的物体, 而其他的3个选项均不属于该范畴	抽象、属性	OBQA
What will Quinn want to do next? A) steal a cone B) buy a new pet C) eat some ice cream	根据上下文可知在冰激凌店门口, 由常识知识得知其蕴涵想吃冰激凌的意图	语境、意图	SIQA

(1) 属性: 主要指对象 (比如实体或事件) 的特性, 语言形式多为形容词或名词. 进一步地, 属性也包括对象的用途或其造成的影响. 相应地, 时间或空间位置关系从时空维度刻画了对象的特性, 本文将其归为属性类型. 例如, 在表 7 的 MCS2.0 中 (第 1 个实例), 问题中的“around (译文: 周围)”表示答案与空间位置相关, 结合“种花”的场景信息, 可根据常识知识“在种花的场景下, 植物在土壤里生长”推理出答案“A) soil (译文: 土壤)”. 具体地, “植物在土壤

里生长”可理解为植物周围是土壤, 两者之间可看成一种空间关系的映射.

(2) 语义: 指问题与正确答案的语义信息. 比如, 问题与正确答案之间的同义词共现、反义词映射以及语义相似性. 例如, 在表 7 的 CSQA 中 (第 2 个实例), 问题为“Google Maps and GPS services have replaced what? (译文: 地图和 GPS 服务已经取代了什么?)”, “replaced”表明该问题考察 CQA 模型对问题的语义理解, 以及同义关系的推理能力, 结合候选答案“D) atlas (译文: 地图集)”, 与问题中“Google Maps”和“GPS”之间存在 语义上相似性, 同时三者均具有提供“导航信息”的特性. 这一例证显示了语言学的语义类知识在 CQA 研究中的重要价值.

(3) 因果: 体现事件和状态发生变化的条件和因由的一类知识, 其往往有助于置信推理、驳斥推理, 以及因或果的回溯与预测. 当缺乏这类知识的时候, CQA 模型难以从给定的上下文、问题与候选答案的浅层语义层面出发, 进行答案的准确求解. 例如, 在表 7 的 ARC-C 数据源中 (第 3 条示例), 针对问题“Which factor will most likely cause a person to develop a fever? (译文: 哪种因素最有可能导致一个人发烧?)”, CQA 模型需要根据常识知识“导致发烧的常见原因之一, 即细菌或真菌感染”, 并结合可训练和可运算的常识推理模块, 才能求解答案为“B) a bacterial population in the bloodstream (译文: 血液中的细菌群)”.

(4) 语境: 蕴涵问题与答案关系的常见情景知识, 在文字层面表现为一种能够解释问题与答案逻辑关系的完整上下文语境. 推敲答案的过程中, 场景知识是一种重要的约束条件. 在去除这一约束的情况下, 问题的答案具有较高的不确定性. 场景类知识的运用, 往往对完成多项选择形式的 CQA 任务有着极高的价值. 例如, 在表 7 的 MCS2.0 数据源中 (第 4 个实例), 对问题“When did they have their own place and space? (译文: 他们什么时候有了自己的地方和空间?)”进行求解, 需认知“符合成年标准的人会离开家独自生活”的场景, 当这一场景存在于上下文语境时, 答案为“B) When the person moved out of the parent’s house (译文: 当这个人搬出父母家的时候)”. 相对地, 如果忽略场景的约束, 候选答案“A) When in high school (译文: 在高中的时候)” 也可作为正确答案.

(5) 抽象与实例: 概念与具象化对象的附庸关系知识. 例如, 在表 7 的 OBQA 中 (第 5 个实例), 问题为“Poison causes harm to which of the following? (译文: 抑制剂会对下列哪项造成伤害?)”, 结合该示例对应的外部事实“Poison causes harm to living things. (译文: 抑制剂会对生物造成伤害.)”, 可知具有“living things (译文: 生物)”的特性为正确答案, 由常识知识“生物具备生命的特征”, 可推理出正确答案为“A) a tree (译文: 树)”. 其中, “a tree”是对外部事实中“living things (译文: 生物)”的具象化表述.

(6) 意图: 表征诉求的一类常识知识. 例如, 在表 7 的 SIQA 中 (第 6 个实例), 问题中“What will Quinn want to donext? (译文: 奎恩接下来想做什么?)”, 是对奎恩意图的提问, 在给定上下文“停留在冰激凌店门口”的情况下, CQA 模型需要揣摩意图与“冰激凌店”的关系, 从而才能从候选答案“A) steal a cone (译文: 偷一个蛋筒)、B) buy a new pet (译文: 买一个新宠物) 和 C) eat some ice cream (译文: 吃冰激凌)”中, 有效选择正解. 这类意图知识或意图的感知能力, 往往需要预先建立机器学习过程, 才能有效完善 CQA 模型的推理能力.

由上述分析可知, 了解不同 CQA 任务对知识类型的需求, 有助于科研人员有针对性地设计常识知识挖掘和应用技术. 值得进一步说明的是, 不同数据集中的常识知识类型分布迥异, 分析类型分布有助于建立“数据驱动”的高兼容性 CQA 建模思维. 具体地, 本文基于第 2 节提及的 9 套 CQA 数据集中的 8 套 (即排除 ReCoRd), 进行知识分布的统计分析, 平均从每套数据集中随机采样 100 个样本, 分别面向属性、语义、因果、语境、抽象和意图 6 个类型, 统计上述样本在各个类型上的数量分布情况. 结果如图 1 所示. 其中, ARC 仅统计挑战集 (简写为 ARC-C), MCScript 和 MCS2.0 只统计问题属于常识范畴的样本. 由于 ReCoRd 主要面向 MRC 领域, 涵盖的常识问答样本不多, 因此图 1 未对 ReCoRd 进行归纳. 需要注意的是, 根据人工归纳统计的过程发现, 针对某些数据集的随机采样存在“未登录项” (即样本不对应任何本文定义的知识类型), 在汇报这类未登录项的同时, 本文直接将该部分样本不纳入知识类型统计中, 因此, 某些数据集对应的样本数量实际上不足 100 条. 此外, 针对某些数据集的随机样本对应多个知识类型 (如表 7 的第 2 个实例), 本文将其分别累计至对应的知识类型统计中, 因此, 某些数据集各类别样本数量和大于 100 条.

根据图 1 中的分布状况可知, CQA 数据集中所需的常识知识类型的占比有着较大的差异, 这一差异反映了数据集及其任务的知识关注焦点不尽相同. 以属性类型为例, 除了 SIQA 和 Cosmos QA, 其余 6 套 CQA 数据集中

50% 的样本需要依赖属性类知识探寻可靠的答案. 这一特点在 CSQA、OBQA、ARC-C 和 ProtoQA 数据集中尤为突出. 比如, OBQA 考察的对象往往集中在物理、生物、化学或地理等科普类的属性知识. 此外, 语义类问答样本在 CSQA、OBQA、ARC-C、SIQA 和 Cosmos QA 中均占比 (如图 1 所示) 25% 左右, 其主要考察 CQA 模型捕获和理解问题中语义知识 (如语义相似性的表述) 的能力. 值得指出的是, 对 CSQA 中的问题求解, CQA 模型需要同时具备理解属性类和语义类知识的能力.

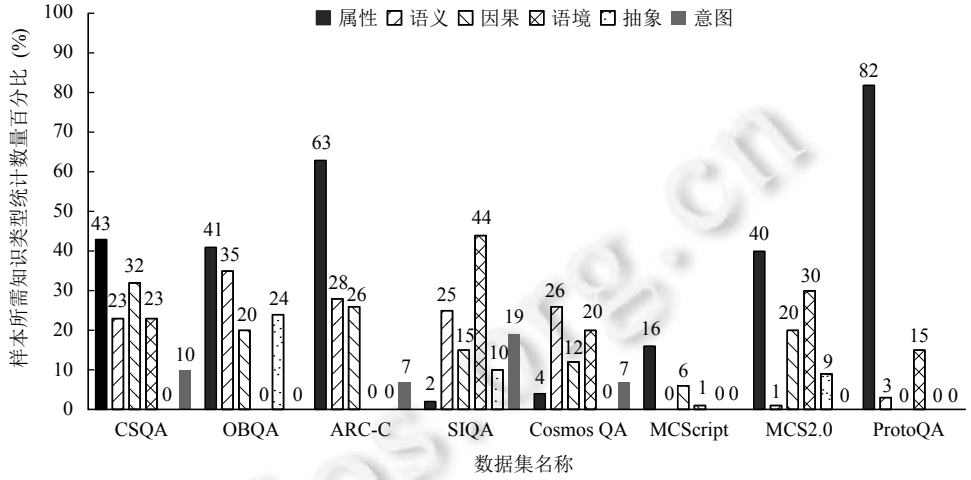

图 1 CQA 数据集中所需知识类型占比图

进一步分析发现, 除 ProtoQA 外的其他数据集均包含“依赖因果知识进行求解”的问答样本, 且占比 (如图 1 所示) 多高于 10% (最高占比接近 32%). 针对这类样本, CQA 模型需要善于溯因或求果. 利用少量训练数据往往难以达到理想的效果, 从而, 引入因果类外部知识或预训练相应的感知能力, 是求解这类问题的重要手段.

语境类型常识知识有助于 CQA 模型深层次推敲问题与答案之间的关系. 其前提是, CQA 模型对语境的范畴和高注意力线索具备感知能力. 在本节给出的 8 套数据集中, 仅有 OBQA 和 ARC-C 未设计语境类知识, 反映了语境理解问题的普遍性 (如图 1 所示, 其中“0”表示某些数据集随机样本中没有需要对应知识类型的案例). 其中较有特色的是, 在本文考察的 SIQA 样本中, 约 44% 的案例涉及语境理解问题, 特别是, 这类语境往往映衬了人物类实体中隐含的情绪或行为. 这类数据可以支撑基于语境的实体画像, 以及在此基础上的问题求解.

针对抽象知识的认知需求, 在上述数据集中的分布并不广泛, 仅在 OBQA、SIQA 和 MCS2.0 这 3 套数据集中存在相应样本, 占比 (如图 1 所示) 也相对较低. 然而, 针对这类样本的研究却极富挑战和趣味性, 对于反映 CQA 模型可解释性的认知水平, 有着重要的意义. 具体地, 在问题中的抽象感念与具象化的答案之间, 蕴藏着晦涩的关联性或对应性, 其对知识理解的广度和深度都有着较高的要求, 且对知识表示的可解释性具有不容忽视的依赖性. 相比而言, 蕴涵意图类常识知识的样本, 在不同数据集中的分布也较低, 仅有 CSQA、ARC-C、SIQA 和 Cosmos QA 蕴涵相应案例, 且占比同样较低. 此外, 意图的学习与认知难度低于抽象类知识, 且前者在推荐系统、个性化检索和观点挖掘等领域都具有同类或近似的任务, 相关研究较为充分. 特别是早期面向问题意图的研究, 已经在技术水平上达到了一定高度, 仅在相关技术的鲁棒性和通用性方面, 仍存在持续探索的空间.

总体上, 常识知识源是研究 CQA 任务的重要组成部分. 针对多源多类的常识知识进行系统研究, 有助于探寻问题的深度理解和可解释性的表示学习方法. 特别是在感知问题与答案的多样性关联, 以及面向关联性的线索挖掘方面, 深度剖析知识源和知识类型的独特性质及内在规律, 都对 CQA 的技术发展有着极为重要的科学价值.

4 专题 1: 结合常识知识源的预训练 CQA 技术

如前所述 (第 1 节技术概述), 围绕 Transformer 架构的 CQA 技术可基本划分为感知常识知识与加载常识知识两大类研究. 在第 1 节, 本文概述了其中较为有代表性的前人工作. 本节将以 BERT、GPT-2 和 T5 等预训练语言

模型为中心, 系统地回顾和分析结合常识知识的预训练 CQA 技术. 本节涉及的技术主要来自 ACL、AAAI、EMNLP、NAACL 和 COLING 等自然语言处理权威国际会议. 回顾和分析的重点集中在 CQA 模型如何获取以及利用常识知识的技术细节, 借以辅助读者全面了解 CQA 研究的各项前沿技术 (各项技术的详细对比见表 8).

表 8 前沿 CQA 技术的对比

模型	设计特色	外部常识来源	融合方式	获取常识的依据	数据集
HyKAS[16]	选项比较网络、多知识源	ConceptNet、Atomic、OMCS	加载、感知	问答样本的实体	CSQA
PEAR[14]	知识路径的筛选	ConceptNet	加载	问答样本的实体	CSQA
DEKCOR[74]	细粒度理解知识三元组	ConceptNet、Wiktionary	加载	问答样本的实体	CSQA、OBQA
MHGRN[75]	多跳关系图推理	ConceptNet	加载	问答样本的实体	CSQA、OBQA
QA-GNN[76]	异构表示的推理	ConceptNet	加载	问答样本的实体	CSQA、OBQA
JointLK[77]	增强版异构表示推理	ConceptNet	加载	问答样本的实体	CSQA、OBQA
RaB-PR[78]	关系感知网络	ConceptNet、 Wikipedia	加载	问答样本的实体	CSQA、OBQA
HGN[79]	混合图网络	ConceptNet	加载	问答样本的语义	CSQA
MSKF[17]	知识过滤、多知识源融合	OMCS、 Wiktionary	加载	问答样本的语义	CSQA
ACP Graph[15]	语义角色图推理	ConceptNet	加载	问题的语义	CSQA
AMS[80]	构建常识类数据集	ConceptNet、Wikipedia	感知	无	CSQA
K-Adapter[18]	分布式常识知识的预训练	Wikidata、Book corpus	感知	无	CSQA、OBQA
Path Generator[19]	常识知识路径的生成	ConceptNet	加载	问答样本的实体	CSQA、OBQA
SEQA[81]	答案生成、基于语义评分	无	感知	无	SIQA、Cosmos QA
Unicorn[82]	基于多任务的迁移学习	Rainbow	感知	与常识相关的语料	CSQA、SIQA、Cosmos QA
UnifiedQA[83]	混合数据集格式的训练	无	感知	无	CSQA、SIQA、OBQA、ARC-C、ProtoQA

注: 数据集该列的统计信息主要来自CQA数据集的Leadboard或模型对应论文中用于实验阶段的CQA数据集

4.1 基于 BERT 的 CQA 技术 (面向知识运用)

针对 CQA 任务, 前人研究通常将预训练语言模型 (以 BERT 及其改进型 RoBERTa、ALBERT 和 ELECTRA[84] 等) 视为一种语言理解的表示工具, 并在此基础上, 引入外部常识知识源, 设计专门的常识推理网络. 这类工作按照获取常识知识的依据不同, 可分为围绕问题样本的实体和语义两类.

(1) 以问题样本中的实体为依据

Ma 等人[16]认为向 CQA 模型提供特定任务所需类型的常识知识, 可有针对性地设计常识推理网络. 为此, Ma 等人围绕 ConceptNet、Atomic 和 OMCS, 采用基于注意力机制和预训练任务两种方式, 分析不同的常识知识源以及常识知识的加载方式对常识推理过程的影响. 具体地, Ma 等人设计了 HyKAS 模型, 其基于词性标注以及精确匹配方式, 从问答样本和 ConceptNet 中 (头实体、关系、尾实体) 提取实体词 (问题中的实体与头实体相匹配, 答案中的实体与尾实体相匹配), 以及包含该实体词的知识三元组, 并采用注意力机制获取知识三元组与问答样本的注意力权重. 在此基础上, HyKAS 结合基于选项比较网络的 CQA 模型, 进行答案解码. 由于 OMCS 与 CSQA 具有相同的实体概念, 并存在域重叠现象, HyKAS 分别引入 OMCS 和 Atomic, 借助掩码语言模型的重新训练, 使之对常识知识的关系特征具备感知和表示能力. 结果表明了基于注意力加载常识知识的有效性, 且证明了常识知识源和数据集的域重叠特性, 适于常识知识的加载和 CQA 模型知识感知.

Yang 等人[14]将实体词作为关联 CSQA 与 ConceptNet 的重要依据, 其通过外部知识查找模块, 从 ConceptNet 中定位与问答样本相关的知识路径, 并计算路径中出现在问题和选项中单词的百分比, 借以筛选出前高相关性的多条路径. 同时, 该方法借助人工设置的模板以及 RoBERTa 的表示学习能力, 选择模式匹配度较高的候选句子作为推理依据. 较为特别的是, 该方法利用 GRU 识别每个词元的上下文信息, 通过注意力机制对多条推理证据分配

不同的注意力分数, 以获得高注意力多源依据的特征表示.

Xu 等人[74]指出, 虽然现有知识图谱能向 CQA 模型提供丰富的知识结构信息, 但它缺乏对概念更精细化的理解. 从而, 基于知识图谱的 CQA 模型欠缺抽象概念和实例关系的理解能力. 针对这一问题, Xu 等人在引入 ConceptNet 的基础上, 利用 Wiktionary 扩展 ConceptNet 概念 (实体) 的上下文. 具体地, Xu 等人将问题和选项中包含的概念 (实体) 作为检索对象, 从 ConceptNet 提取同时包含该概念的三元组, 并通过人工定义的文本匹配准则, 从 Wiktionary 中获取与该概念相关的定义. 该方法最终将最高匹配度的定义表述视作相关上下文, 并将其作为 ALBERT 的输入, 增强词嵌入表示的确切性, 借以弥合 CQA 模型对三元组的理解偏差.

Feng 等人[75]认为目前加载外部知识源的 CQA 模型, 无法提供可解释的预测, 且难以有效建模问题与答案中的实体多跳关系. 因此, Feng 等人提出了多跳图关系网络 (multi-hop graph relation network, MHGRN). 其主要贡献是提出了一个整合图神经网络和关系网络的新型架构, 并将其用于编码多关系图模型, 具备较好的可解释性和可扩展性. 从建模角度而言, MHGRN 通过识别问答样本中的实体, 并将其链接到 ConceptNet 中的对应实体, 构成初始化的实体节点集合, 同时将与该实体节点相关的任意两跳以内的实体添加到该集合中, 形成与问答样本相关的关系图. 在此基础上, MHGRN 采用图神经网络对关系图进行建模, 允许每个节点与多跳邻居之间进行消息传递, 赋予图神经网络直接建模路径的能力, 实现多跳关系推理. 结果表明, MHGRN 通过消息传递, 实现了图模型表示的可伸缩性, 且其引入的结构化关系和相应注意机制确保了信息传递方向的可解释性.

Yasunaga 等人[76]的研究显示, 从知识图谱中获取主题实体 (问题和答案中的实体) 及指定跳数范围内的子图, 会引入语义无关的实体节点. 此外, 问答文本和子图的独立编码, 无法支持两者语义感知模型的统一更新, 从而限制了结构化信息对 CQA 模型的辅助效果. 为此, Yasunaga 等人提出了 QA-GNN (question answering-graph neural network), 其代表性贡献分为两部分. 其一, 利用预训练语言模型计算实体与问答文本的相似性, 借以对子图上的实体进行评分, 从而支持基于评分的关联实体选择; 其二, 将问答文本的表示视为一种附加的节点, 将其链接到子图的主题实体上构成综合子图, 并引入文本特征与实体表示的相关性分数, 形成节点表示的新增特征. 在模型架构上, QA-GNN 采用基于注意力的 GNN 模块, 在综合子图上获取消息可达的局部特征, 且同时更新子图的实体以及问答文本节点的表示, 以弥合两者之间的差距.

Sun 等人[77]认为, 上述 QA-GNN 将问题的上下文表示汇聚到一个节点的方式, 实际上限制了文本表示形式的更新, 也限制了语言模型和 GNN 之间的细粒度信息交互. 为此, Sun 等人提出了 JointLK (joint reasoning with language models and knowledge graphs) 模型. 该模型通过语言模型和 GNN 的联合推理以及动态的子图剪枝机制, 解决了上述限制. 具体地, JointLK 与 QA-GNN 类似, 都以问答样本为查询依据, 从 ConceptNet 中获取与之相关的子图. 但是, JointLK 通过一个密集双向注意力模块在问答样本表示和子图表示之间建立联合推理机制. 此外, JointLK 通过多步交互, 对结构化的知识图和非结构化的问题样本进行信息融合和更新, 并利用动态剪枝模块, 递归地剪枝与问答样本无关的子图节点, 以获得与问答样本密切相关的知识子图.

现有工作未直接对链接问题与答案的实体关系进行建模, 因此, Wang 等人[78]提出了一种关系感知的推理方法, 其对应于表 8 的 RaB-PR 模型. Wang 等人的主要贡献是, 使用关系感知图网络组合学习多关系图中的节点和关系, 为双向推理模块提供增强的关系表示. 此外, 为了构成多常识知识源的子图, RaB-PR 从 ConceptNet 中提取相关知识的关系路径, 借以实现概念知识子图的融合, 同时采用 ElasticSearch 从 Wikipedia 中获取语义相关的句子, 并利用 Standford OpenIE5 提取句子中的三元组, 形成兼容于上述 ConceptNet 定义结构的知识子图和关系匹配机制. 在此基础上, RaB-PR 利用关系感知图网络模块对多源子图进行编码, 全面捕获多源外部常识知识中的丰富常识信息. 此外, RaB-PR 利用双向路径推理方法和实体表示的注意力加权聚类方法, 为 CQA 模型的预测提供明确的关系路径.

(2) 以问题样本的语义为依据

Yan 等人[79]发现, 从知识图谱中提取的子图可能产生稀疏边以及噪声边的问题, 其不利于 CQA 模型获取与问答样本高度相关的知识. 为此, Yan 等人提出了混合图网络 (hybrid graph network, HGN), 其主要关注知识图谱中边与上下文特征的信息, 借以优化子图的知识表示. 具体地, Yan 等人基于问答样本中的实体, 从 ConceptNet 中

提取与之相关的知识三元组构成子图. 其利用提取的子图以及上下文特征, 且训练 HGN 生成子图的缺失边. 由此形成的“混合图”, 支持图神经网络实现知识关系的重新加权 (降低不相关边的权重) 和清除消息传递中的干扰节点 (减弱无关边的影响). 这一方法聚焦语义的表示, 进行相关知识的获取与运用, 实现了知识图谱增强 CQA 模型的目的.

李志峰等人[17]发现当前的 CQA 模型往往受到常识知识源中的噪声知识干扰, 且对多知识源的融合与利用不够充分. 为此, 李志峰等人提出了一种基于多知识源融合 (multi-source knowledge fusion, MSKF) 的方法. 其采用贪心筛选策略对噪声知识进行过滤, 并应用多通道知识融合方法, 汇聚各知识源不同类型的知识. 具体地, MSKF 采用规则以及正向匹配算法, 从 Wiktionary 中匹配与问题概念或者选项相关的释义, 且利用 ElasticSearch 工具以“问题概念+选项”为关键字对 OMCS 进行检索, 基于语义相似度的结果, 提取可靠的常识知识. 在此基础上, MSKF 利用基于贪心策略的知识排序方法, 选择与问题以及选项高度相关的常识知识以及释义, 并将其直接加载至基于 ALBERT 构建的知识编码通道, 以此为答案解码提供知识. 结果表明, MSKF 中知识筛选方法有助于过滤噪声知识 (如与问答样本无关的常识知识或释义).

Lim 等人[15]观察到, 加载知识图谱的 CQA 模型, 虽然借助知识节点嵌入和样本表示, 实现了知识驱动的答案注意力重估. 但是, 其并未强化问题语义表示, 及其在预测答案过程中的作用. 为此, Lim 等人提出了抽象语义表示的概念剪枝图方法 (abstract meaning representation-ConceptNet-pruned, ACP). 其主要贡献是关注问题中隐含的语义信息, 通过抽象语义表示 (abstract meaning representation, AMR), 构建语义角色图, 联合 ConceptNet 建模问题到答案的推理过程. 具体地, ACP 使用 AMR 解析器生成问题的 AMR 图, 并引入与之相关的 ConceptNet 知识三元组及其关系. 特别地, 其根据常识关系对该图进行剪枝, 从而形成抽象的 ACP 图. 在此基础上, ACP 利用图路径推理框架增强 AMR 与 ConceptNet 概念的交互信息, 然后才将知识结构信息和文本语义信息进行融合, 形成可供答案推理的特征表示. 结果表明, ACP 图可辅助 CQA 模型实现语义强相关的路径识别, 其提高了推理过程的可解释性.

上述工作, 均以问题样本为依据, 从常识知识源中获取与之相关的常识知识, 并将其直接加载至 CQA 模型中. 此外, 有一些特色研究根据答案不直接来源于上下文的特点, 采用预训练的方式, 增强 CQA 模型感知常识知识的能力. 代表工作之一来自 Ye 等人[80]. 其提出一种将常识知识整合到语言模型中的预训练方法 AMS (align、mask、select). 该方法在保证预训练语言模型的表示能力不降低的前提下, 增强该模型的常识知识感知与推理能力. 其关键贡献是借助远程监督的思想, 探寻存在实例化关系的实体和概念, 及其句子级语境. 在此基础上, 其通过对齐 ConceptNet 和 Wikipedia 中实体或概念, 自动构建常识依赖的多项选择问答数据集, 并用于预训练语言模型 (BERT).

Wang 等人[18]发现, 现有工作通过设置预训练任务加载多种常识知识时, 语言模型的原始往往出现灾难性的原始感知遗忘 (参数被过度篡改). 为了解决这一问题, Wang 等人提出了 K-Adapter 框架, 它在固定预训练语言模型原始参数的基础上, 支持持续的知识加载. Wang 等人以 RoBERTa 为骨干模型, 将 K-Adapter 设计成一种连接到 RoBERTa 的插件, 用于对特定知识进行预训练. 该框架设计了事实知识适配器以及语言知识适配器, 前者加载 Wikipedia 和 Wikidata, 通过实体词进行文本三元组的自动对齐; 后者基于依存分析从 Book corpus 中获得的句法和语义等语言知识, 并对编码表示进行加载与扩展, 以此构建相关于事实和语言多项预训练任务 (关系预测和依存判断), 从而增强了适配器对不同类型常识知识的感知和表示能力. 此外, 不同的适配器以分布式的形式进行训练, 这种灵活性使 K-Adapter 能够有效且独立地加载不同类型的常识知识, 而不会对先前加载的常识知识造成损失.

4.2 基于 GPT-2 的 CQA 技术 (面向数据稀疏)

考虑到知识图谱的不完备性问题, 以及采用检索的方式会引入噪声信息的客观现状, Wang 等人[19]选择容错性较高且知识感知范畴更大的预训练语言模型 GPT-2 进行 CQA 的设计, 其主要方式是基于领域数据对 CQA 进行微调. 较为特别的是, Wang 等人利用 GPT-2 进行知识路径的生成, 并将其作为答案解码的重要线索. 该方法对应于表 8 的 Path Generator. 从技术角度而言, Path Generator 通过随机游走在 ConceptNet 上采样了不同跳数的知

识路径, 以这类知识为二次预训练任务对 GPT-2 进行微调, 目的是将丰富的结构知识编码在 GPT-2 中, 提高该模型的常识知识感知与泛化能力. 在此基础上, Path Generator 构成一个链接源概念 (问题) 与目标概念 (答案) 的路径生成器, 并联合问答样本的表示, 训练 CQA 模型选择高相关度的知识路径和推理高注意力的答案. 结果表明, 通过丰富的结构化知识进行微调的预训练生成模型, 可产出与下游任务相关的常识知识. 此外, 利用预训练模型本身编码到的语言现象, 生成一些本不存在于知识图谱中的知识路径, 以缓解知识图谱的稀疏性问题.

考虑从无监督的角度解决 CQA 任务, 以此缓解标记数据稀疏情境下 CQA 训练不足的问题. 基于此, Niu 等人[81]在 GPT-2 建模架构的基础上, 提出了基于语义的问答方法 (semantic based question answering, SEQA). 其主要贡献是跳出显式知识源加载或感知训练的固有思路, 利用预训练生成模型 GPT-2 本身的知识感知与生成能力, 在结合问题语义表示的前提下, 生成一组面向候选答案的“投票”. 在此基础上, Niu 等人以基于人工规则重写的问题为条件, 利用 SRoBERTa 分别获得选项以及“投票”的语义特征, 从而计算每个“投票”与选项的语义相似性, 并选择具有最大语义相关度分数的候选作为输出. 结果表明, 预训练生成模型虽然无法直接生成正确答案, 但是其在语义层面与正确答案具备一定程度的语义相似性, 可为 CQA 模型提供推理信息. 上述基于 GPT-2 的常识知识编码、吸收和融合方式, 以及实验环境如表 8 所示.

4.3 基于 T5 的 CQA 技术 (面向泛化性)

面向不同 CQA 数据集, Lourie 等人[82]希望建立一种泛化于多个 CQA 数据集的技术. 其首先提出了两种评估模型泛化水平的方法, 对应于表 8 的 Unicorn. 其中, 多任务基准 Rainbow 用于检验面向不同 CQA 任务的泛化性水平; 此外, 成本等效曲线用于检验上游模型 (预训练和迁移学习) 的适应性, 以及数据处理效率. 在此基础上, Lourie 等人对基于 T5 构建的 CQA 模型进行了迁移学习, 其学习方式采用多任务训练、顺序训练以及多任务微调 3 种. 结果表明, 较大的模型从迁移学习中受益更多, 只需使用少量的示例, 即可达到等同于基线模型的性能. 同时, Lourie 等人利用成本等效曲线, 检验了不同单任务模型在数据处理过程中产生的等效成本, 其有助于提升模型的计算效率, 且不损害其性能.

Khashabi 等人[83]认为 CQA 的任务形式和所需要的知识, 与其使用的数据集有着必然联系. 可以认为这种联系实际上是一种隐含的约束. 但是, 现有评估 CQA 模型的标准往往一致, 从而导致现有建模方案存在学习目标的盲目性. 针对这一问题, Khashabi 等人提出一种跨数据集格式的混合训练方案 Unified QA. 其采用文本范式对不同数据集的格式进行统一, 从而有效构造了混合训练池, 实现了跨格式跨数据集的监督学习. 在此基础上, Unified QA 利用 T5 或 BART 自身具备的知识感知与表示能力, 学习不同数据集的观测样本, 借此形成综合的感知模型和推理模型. 结果表明, Unified QA 模型表现出较强的泛化性, 可作为其他 CQA 任务展开研究的起点.

由上述枚举的研究可知, CQA 技术的核心是如何获取常识知识以及依据具体任务的特点, 设计推理机制和优化方法 (抗噪、低数据依赖和泛化). 总体上, 这类研究应用常识知识的方法可归纳为如下两大类. 其一, CQA 模型依据问答样本, 设计匹配或检索方法, 直接从外部常识知识源 (如 ConceptNet、Wikipedia) 中加载常识知识; 其二, 借助外部知识源设计预训练任务, 促进 CQA 模型增量数据中蕴涵的常识知识进行表示学习, 使之具备感知常识知识的表示能力. 同时, 推理机制因定位常识知识的方法而有所差异. 以问答样本为依据的常识推理方法主要围绕注意力机制和图神经网络等技术展开, 其源于加载的常识知识源多以图状或层级的形式组织. 此外, 相关研究深挖知识源中实体的内在联系, 借以支持文本表示与知识表示交互的语义计算方法. 需要指出的是, 特色技术研究也逐步展开, 在抗噪、可解释性和低数据依赖方面, 不断设计出各类优化方法. 最后, 基于预训练任务的常识推理方法, 专注于模型本身的表示学习能力, 通过在蕴涵常识知识的语料上进行二次训练, 增强其对特定领域或数据集中常识知识的感知与表示能力, 具备较好的泛化性和通用性.

5 专题 2: CQA 与知识类型体系的关系分析

目前, 针对不同知识类型进行考察的相关工作较少. 此外, 面向不同知识类型, 检验现有 CQA 技术适应性和通用性的专门研究也乏善可陈. 为此, 本节开展了一项预研性的验证工作, 即 CQA 技术与知识类型的适用关系研究,

以此为未来相关技术的优化提供参考. 这一初步研究纳入了第 3 节建立的常识知识分类体系, 并考察了一套基于预训练语言模型的 CQA 技术. 考察点设定为 CQA 技术在不同知识类型上的应用效果.

5.1 实验设置

5.1.1 数据集

第 2 节面向 9 套 CQA 数据集建立了分类体系. 然而, 不同数据集上的 CQA 任务形式略有不同. 为保证实验的公平性和可靠性, 本文仅选择了其中 5 套数据集进行测试, 包括 CSQA、OBQA、SIQA、ARC-C 和 MCS2.0. 实验未选择简化版数据、混杂选项数据和生成类 CQA 数据 (ARC-Easy、Cosmos QA、MCScript 和 ProtoQA). 具体理由如下.

(1) ARC-C 相比于 ARC-Easy, 前者包含的常识样本更多, 同时其问题均来自更具挑战且易错的样例, 任务难度高于 ARC-Easy, 所以本节未选择 ARC-Easy 作为分析对象.

(2) Cosmos QA 中存在一个问题指向两个不同的答案 (随机抽样 100 个样本中该案例占比 15%), 形成了单选和多选混杂的 CQA 模式, 与其他数据集上的任务模式异同, 本节实验并未纳入考虑范围.

(3) MCScript 中常识性问题较为稀疏, 占比不足 30%, 而 MCS2.0 数据集与其相比包含的常识问题更多, 且 MCScript 与 MCS2.0 的常识知识类型基本相似, 因此, 实验未纳入 MCScript.

(4) ReCoRd 主要面向机器阅读理解任务进行数据建设, 任务考察点侧重抽取式的答案定位和上下文理解, 对于常识知识的加载和感知, 并未形成有针对性的验证方法. 为此, 本节未将 ReCoRd 纳入实验验证范畴.

(5) ProtoQA 是生成式 CQA 数据集. 考虑到实验验证对模型架构的统一性要求, 本节不对其展开分析.

表 9 罗列了 5 个 CQA 数据集的样本形式与划分方式, 其中, MCS2.0 只包含常识问题的样本. 由表 9 中“样本形式”可知, 5 个数据集在样本形式的组成元素上存在区别. 比如, CSQA 中每个样本包括问题和候选答案; 在 CSQA 的基础上, OBQA 与 ARC-C 均包含基于科学事实的语料库; SIQA 和 MCS2.0 包含与问题相关的上下文. 需要指出的是, 基于科学事实的语料库为 CQA 模型预测正确答案提供关键的推理证据, 上下文则能够提供背景信息, 辅助 CQA 模型深入理解问题的语义. 两者均为模型预测正确答案提供了额外的参考信息源. 为了保证实验验证的公平性, 本文在数据预处理阶段禁用了 5 个 CQA 数据集中的常识知识文本或参考数据库. 最终, 每个数据集的样本形式在组成上只包含问题与候选答案.

表 9 实验数据

数据集	样本形式	数据集规模		
		训练集	验证集	测试集
CSQA[20]	问题, 候选答案	9741	1221	1221
OBQA[21]	问题, 候选答案, 语料库	4957	500	500
ARC-C[22]	问题, 候选答案, 语料库	1119	299	1172
SIQA[23]	上下文, 问题, 候选答案	33410	1954	2224
MCS2.0[26]	上下文, 问题, 候选答案	7091	966	1898

5.1.2 实验设置 (基于知识感知的 CQA)

在 CQA 模型中融入常识知识是本节实验验证最重要的步骤. 本节实验仿照 Ma 等人[34]的基于数据增强和知识感知 CQA 方法, 建立实验模型. 具体地, 本文基于常识知识源自动构建常识性问答数据集 (外部数据集), 并在该数据集上微调预训练语言模型及其上层 CQA 解码器, 形成迁移学习后的增强 CQA 模型. 实验将该模型投入前述 5 套数据集进行实验, 并观测该模型在不同数据集和知识类型上的表现. 上述 CQA 技术的形成过程共计如下 3 个步骤 (熟悉这一工作的读者可越过这一部分进行阅读).

(1) 常识知识类型与关系归纳

常识知识类型通常基于常识知识源中的关系进行划分. 按照 Ma 等人[34]提出的模型设置, 本节实验将基于 ConceptNet、WordNet、Wikidata 和 VisualGenome (CWWV) 中包含的关系, 通过 CSKG 框架[85], 将多个常识知识

源合并成一个集成的常识知识图谱. 在此基础上, 本文形成兼容于 36 种 ConceptNet 概念关系的匹配机制, 并根据关系的具体含义将其归纳于 6 种常识知识类型. 以表 10 中属性类型常识知识为例, 其主要关注实体或事件具有的特性. 因而, 本节将 ConceptNet 中表示该特性的关系 (如“UsedFor 和 SymbolOf”) 划分到表示属性类常识知识的关系集合.

表 10 常识知识类型与关系集合统计

常识知识类型	关系	数量
属性	UsedFor、CapableOf、SymbolOf、MadeOf、LocatedNear、HasProperty、PartOf、FormOf、AtLocation、HasA、DistinctFrom、CreatedBy、NotCapableOf、NotHasProperty、ReceivesAction	15
语义	Synonym、Antonym、SimilarTo	3
因果	Causes、HasPrerequisite	2
语境	HasLastSubevent、HasFirstSubevent、HasContext、RelatedTo、EtymologicallyRelatedTo、MannerOf、HasSubevent、Entails	8
抽象	IsA、DefinedAs、EtymologicallyDerivedFrom、InstanceOf	4
意图	CausesDesire、Desires、MotivatedByGoal、NotDesires	4

(2) 常识知识数据集的构建与作用

利用常识知识数据集对预训练语言模型进行迁移和微调, 可驱动该模型学习多源知识. 为此, 本文利用上述 CWWV 知识源, 构建了蕴涵问答模式的多项选择式常识知识数据集. 具体的构建方式与 Ma 等人[34]的工作类似, 即采用通用的启发式规则, 在常识知识的三元组 (头实体、尾实体和关系) 与问答样本 (问题、答案和关系) 之间建立映射关系, 且利用随机抽样和生成模型批量地制造干扰项. 其中, 映射的关系集合来源于步骤 (1) 参照的 36 种 ConceptNet 关系.

根据常识知识类型不同, 本文将常识知识数据集分为 6 种. 表 11 展示了基于 CWWV 构建的常识知识数据集示例, 其以“UsedFor”关系为例, 描述常识知识数据集的构建过程. 具体地, 本文构建模板“UsedFor: $node1 can be used for [MASK].”; 其次, 基于 CSKG 框架, 本文从 CWWV 中检索具有“UsedFor”关系的三元组“<alarm clock, UsedFor, awaken>”, 并结合模板将三元组的头实体“Alarm clock”和关系“UsedFor”转换成一个完整的自然语言句子“Alarm clock is used for [MASK].”, 同时将尾实体“awaken”作为正确答案. 其余的两个干扰答案均采取同样的方式从 CWWV 中随机抽取, 其抽取条件需要满足关系属于“UsedFor”的条件, 且符合头实体与“alarm clock”无相同字符的要求. 依照这类三元组, 本文提取其尾实体作为干扰答案. 值得注意的是, 本文通过限制检索知识三元组的关系类型, 使得问题与正确答案和干扰答案之间蕴涵的常识关系保持一致.

表 11 常识知识数据集示例

常识知识类型	关系	示例
属性	UsedFor	Question: Alarm clock is used for [MASK]. A) warm **B) awaken** C) removing chlorine used in pools
语义	Desires	Question: Humans like to [MASK]. A) conferences **B) awaken** C) information
因果	Causes	Question: Sometimes bringing home some fish causes [MASK]. A) possible jealousy B) or to get angry **C) spouse to bring out cookbook**
语境	RelatedTo	Question: Oil is related to [MASK]. A) dom **B) car fluid** C) protective gear
抽象	IsA	Question: A question is [MASK]. A) Gospels B) transferral **C) query**
意图	Motivated ByGoal	Question: You would go to the mall because [MASK]. A) tone up **B) bored** C) feel needed

注: 表中罗列了6种单一类型的常识知识数据集示例, 其示例所在列中, 加粗为正确答案

(3) 任务模式、CQA 模型与实验训练方式

实验以多项选择为目标任务, 构建了一个基于 BERT 的通用 CQA 模型. 实验涉及的数据集除了对应 6 种单一知识类型的常识知识数据集之外, 增设了一套混合类的常识知识数据集, 其由 6 种常识知识类型分布均匀的常识知识数据集合并而成, 并采用抽样选择的方式, 保证了其样本总量与单一类型常识知识数据集的一致性.

本文采用上述 7 种常识知识数据集 (6 个单一类型和一个混合类数据集) 独立地对 CQA 模型进行预训练 (参数微调), 促使模型学习不同类型的常识知识, 使之具备感知常识知识的能力. 在此基础上, 实验采用零样本 (zero-shot, ZS) 和微调 (fine-tuning, FT) 两种训练方式, 分别调优 BERT. ZS 训练特指基于上述常识知识数据集的一次性微调, 其不经过领域数据集的二次微调, 直接投入测试过程. FT 训练则不同, 除了基于常识知识数据集的第一次微调, 其额外利用领域数据集的训练集, 对 BERT 进行第 2 次微调. 本文实验采用的领域数据集, 为第 5.1.1 节给出的 5 套 Benchmark CQA 数据集. 实验采用的性能评估方法为准确率 (accuracy).

5.2 实验结果与分析

为便于理解, 本文在后续分析过程中对 CQA 模型进行了不同命名. 其中, 基于 ZS 设置进行训练的模型称为 ZS-CQA, 而在 FT 设置下进行训练的模型称为 FT-CQA.

(1) ZS 实验及其结果分析

实验结果如表 12 所示. 其中, 基于混合类知识形成的 ZS-CQA 模型为基线, 在属性类、语义类、因果类、语境类、抽象类和意图类知识上分别形成的 ZS-CQA 模型为对照模型. 验证焦点集中在“等量混合知识”与“等量同类知识”对于 CQA 模型优化的确切影响.

表 12 ZS 实验结果 (评估测度为准确率 accuracy) (%)

常识知识数据集	CQA数据集				
	CSQA	OBQA	SIQA	ARC-C	MCS2.0
混合类 (基准)	41.36	27.00	34.85	23.73	65.56
属性类	39.97 (↓1.39)	24.40 (↓2.60)	35.57 (↑0.72)	24.07 (↑0.34)	67.08 (↑1.52)
语义类	38.65 (↓2.71)	18.40 (↓8.60)	34.80 (↓0.05)	21.35 (↓2.38)	65.21 (↓0.35)
因果类	41.76 (↑0.40)	27.40 (↑0.40)	33.62 (↓1.23)	23.86 (↑0.13)	65.79 (↑0.23)
语境类	34.89 (↓6.47)	23.40 (↓3.60)	33.72 (↓1.13)	22.03 (↓1.70)	65.84 (↑0.28)
抽象类	37.23 (↓4.13)	22.80 (↓4.20)	35.47 (↑0.63)	23.05 (↓0.68)	65.42 (↓0.14)
意图类	39.56 (↓1.80)	25.20 (↓1.80)	33.93 (↓0.92)	25.42 (↑1.69)	65.18 (↓0.38)

注: 表中准确率的增长 (如符号“↑”所示)代表正面影响, 准确率的下降 (如符号“↓”所示)代表负面影响

由表 12 可知, 基于任意一种知识类进行独立学习和一次性微调的 ZS-CQA 模型, 在不同 CQA 数据集上的性能有较大差异. 该结果与 CQA 数据集本身的难度以及常识知识数据集中蕴涵同一常识知识类型的样本数量有关. 与混合类 ZS-CQA 模型相比, 属性类 ZS-CQA 模型 (即使用属性类知识单独训练所得的 ZS-CQA 模型) 在 SIQA、ARC-C 和 MCS2.0 上取得的性能均得到了提升, 尤其在 MCS2.0 上, 其提升幅度高达 1.52%. 同时, 该模型在 CSQA 与 OBQA 上的性能均呈现下降趋势, 下降的幅度分别为 1.39% 和 2.60%. 由分析可知, 混合类与属性类数据集在常识知识分布上存在差异, 前者包含 6 类分布均匀的常识知识类型 (属性类数据占总样本量的 1/6), 而后者仅包含单一类型的常识知识 (样本全部为属性类合成数据). 因此, 由两者分别预训练的 CQA 模型在不同的数据集上测试的性能变化具有较为明显的差异. 其结果说明, 依据常识知识类型划分的数据集, 模型可通过预训练的方式从中隐式地学习多源常识知识. 同时, 常识知识数据集中常识知识类型的划分与蕴涵同一常识知识类型样本数量的占比, 直接影响 ZS-CQA 模型在下游任务上的性能. 类似地, 其余 5 种单一类型 ZS-CQA 模型在不同 CQA 数据集上的测试结果, 同理可得以上结论.

单独观察每个 CQA 数据集的评估结果, 不同类型 ZS-CQA 模型所取得的性能有较大差异. 对于 CSQA, 因果类 ZS-CQA 模型在该验证集上的准确率提升了 0.40%, 其余 5 类单一类型 ZS-CQA 模型在该验证集上的准确率均呈现不同程度的下降趋势. 尤其是语境类 ZS-CQA 模型, 与混合类 ZS-CQA 模型相比, 其在 CSQA 验证集上的准

确率下降幅度达 6.47%. 该结果表明, 语境类常识知识数据集与 CSQA 数据集的分布存在较大差异, 预训练阶段使用该类型常识知识数据集, 相当于引入了噪声, 从而影响 ZS-CQA 模型在下游任务中的常识知识感知与表示能力. 类似地, 对于 SIQA 而言, 属性类和抽象类 ZS-CQA 模型在该验证集的准确率均取得提升, 提升幅度分别为 0.72% 和 0.63%, 其余 4 种单一类型 ZS-CQA 模型所取得性能均呈现下降趋势. 其中, 基于语境类 ZS-CQA 模型在 SIQA 上取得的性能最低, 为 33.72%. 以上结果表明, 语言模型采用预训练任务方式, 可隐式地学习常识知识源中蕴涵的常识知识, 以增强该模型的常识知识感知与表示能力. 同时, 由不同类型的常识知识数据集分别预训练的 CQA 模型, 其在同一个数据集上取得的性能呈现差异性的性能提升或下降, 说明该模型针对不同类型的常识知识, 其感知与表示能力存在差别.

(2) FT 实验及其结果分析

FT 实验如表 13 所示. 值得赘述的是, FT 实验中 CQA 模型都经历了二次微调, 此次微调使用的数据皆来自领域数据集的训练集 (即 CSQA、OBQA、SIQA、ARC-C 和 MCS2.0 的训练集), 而测试过程则利用了上述领域数据集的验证集. FZ 实验的考察点是, 在独立知识类型数据集上进行第 1 次微调后, 在混杂不同类型的 benchmarks 上进行二次微调, 是否有助于提高多类型知识问答的适应性. 同样地, 准确率用于评估性能的变化.

表 13　FT 实验结果 (评估测度为准确率 accuracy) (%)

训练设置		CQA数据集				
		CSQA	OBQA	SIQA	ARC-C	MCS2.0
无预训练 (基准)		53.64	53.20	51.20	33.22	70.60
有预训练	混合类	54.15 (↑0.51)	54.40 (↑1.20)	51.79 (↑0.59)	34.86 (↑1.64)	75.05 (↑4.45)
	属性类	55.38 (↑1.74)	58.00 (↑4.80)	53.84 (↑2.64)	35.59 (↑2.37)	76.09 (↑5.49)
	语义类	55.28 (↑1.64)	55.20 (↑2.00)	52.46 (↑1.26)	34.91 (↑1.69)	74.43 (↑3.83)
	因果类	54.95 (↑1.31)	53.00 (↓0.20)	53.32 (↑2.12)	34.58 (↑1.36)	75.56 (↑4.96)
	语境类	54.86 (↑1.22)	56.00 (↑2.80)	53.43 (↑2.23)	32.88 (↓0.66)	75.35 (↑4.75)
	抽象类	51.43 (↓2.21)	55.00 (↑1.80)	52.81 (↑1.61)	31.53 (↓1.69)	75.26 (↑4.66)
	意图类	54.79 (↑1.15)	52.10 (↓1.10)	52.96 (↑1.76)	36.27 (↑3.05)	74.74 (↑4.14)

注: 表中准确率的增长 (如符号“↑”所示)代表正面影响, 准确率的下降 (如符号“↓”所示)代表负面影响

由表 13 可知, 相比于无预训练任务, 混合类 FT-CQA 模型在不同数据集上均有不同程度的提升. 其中, 在 MCS2.0 上的性能优化高达 4.45%, 而幅度最小的优化也达到了接近 1.00% 的水平.

相比于混合类 FT-CQA 模型, 单一类型的 FT-CQA 模型在 5 套通用数据集上取得的性能呈现上升或下降的趋势. 以属性类数据集为例, 相比于混合类数据集, 前者对应的 FT-CQA 模型在 5 个 CQA 数据集上均取得不同程度的提升. 结合图 1 中常识知识类型的统计结果 (除 SIQA 外, 其余 4 个 CQA 数据集所需属性类常识知识的样本均占据 40% 以上的比例), 该模型的预测结果 (属性类 FT-CQA 与基准模型相比有所提升) 与人工观测的常识知识类型统计结果大体相符, 表明本文构建的常识知识分类体系具有可行性.

此外, 采用预训练任务的方式进行二次微调, 逐层增强了 CQA 模型对常识知识类型的学习力度与深度, 从而为其求解答案提供辅助. 需要说明的是, 对于 SIQA 而言, 其所需常识知识类型的统计结果与属性类 FT-CQA 模型在相应验证集上的实验结果存在偏差 (属性类常识知识的随机问答样本占比较少, 但是属性类 FT-CQA 模型与基准模型相比取得了 2.64% 的提升), 其原因极大可能是在 SIQA 的问答样本中, 同一个问题所需的常识知识归属于多个常识知识类型, 且人工观测未能给出一个精确的分类, 从而 CQA 模型学习单一类型的常识知识时, 未具备较好的常识知识感知与表示能力. 其他 5 种单一类型的 FT-CQA 模型, 在 5 套权威的 CQA 数据集得出的实验结果, 与图 1 中人工观测的常识知识类型分布相符, 细微的偏差来源于部分样本的常识类型难以明确分类 (即类型歧义). 以上仍可说明常识知识类型的预分类, 有助于 CQA 模型有针对性地增强其自身对不同类型的常识知识感知与表示能力.

单独观察每个 CQA 数据集的评估结果, 不同类型的 FT-CQA 模型在相应数据集的验证集上表现有较大差异.

以 SIQA 为例, 6 类单一类型 FT-CQA 模型在其验证集上的性能均取得了不同程度的提升. 其中, 提升幅度最大为语境类 FT-CQA 模型, 达到 2.23%. 相比于混合类 FT-CQA 模型, 其优势达到 1.64%. 由上文 (第 2 节的问答数据集构建) 可知, SIQA 着重测试模型对他人隐含的情绪或行为进行推理的能力. 此外, 根据图 1 显示的常识知识类型的统计结果, 可发现 SIQA 所需语境类常识知识的样本占比高达 44%. 上述两点特性, 促进了 CQA 模型学习同构常识知识类型的效果, 从而优化答案求解的精确性. 其余 5 种单一类型 FT-CQA 模型在 SIQA 验证集上的实验结果, 与图 1 中对 SIQA 所需的常识知识类型的统计结果相符. 进一步验证了常识知识分类体系的实用性, 以及借助知识类型调优 CQA 学习过程的可行性.

类似地, 对于 MCS2.0 而言, 相比于基准模型, 6 种单一类型的 FT-CQA 模型在该数据集上的性能均取得了不同程度的提高. 但与基于混合类 FT-CQA 模型相比, 两者的实验结果存在细微的差异. 具体地, 属性类 FT-CQA 在该数据集的验证集上取得了最优性能, 达到 76.10%. 与混合类 FT-CQA 模型相比, 其优势达到 1.04%. 其余 5 类单一类型的 FT-CQA 模型, 与之相比, 上升幅度在 0.60% 以内 (因果类、语境类和抽象类常识知识数据集). 语义类和意图类 FT-CQA 模型在该数据集上的性能均呈现下降趋势. 这一结果与图 1 对 MCS2.0 中所需常识知识类型的统计结果相比, 存在细微偏差. 主要体现在图 1 中, 人工统计 MCS2.0 中所需语境类常识知识的样本占比达 30%, 仅次于属性类. 但是, 语境类 FT-CQA 模型在 MCS2.0 的验证集上提升幅度次于由其他两种单一类型 (如属性类和因果类) 的 FT-CQA 模型. 其原因在于, MCS2.0 中同一个样本往往对应多个异构的知识类型. 从而, 单一类型 FT-CQA 难以利用本身具有的表示能力对多种类型的常识知识进行感知, 负面影响了答案求解效果.

由上述分析可知, 基于 FT 设置 (二次微调) 的 CQA 模型在多个领域数据集上都取得了不同程度的提升, 且实验结果与图 1 中 CQA 数据集中所需知识类型的统计结果大体一致.

(3) ZS 与 FT 对照实验

下面对照分析了 ZS 与 FT 设置下的 CQA 性能. 实验考察了领域数据集 CSQA 和 SIQA 上的验证结果. 图 2 和图 3 分别显示了 CQA 模型在 CSQA 与 SIQA 验证集上的结果. 其中, 横坐标表示 ZS 或 FT 设置下的不同训练数据集, 纵坐标表示 CQA 模型获得的准确率. 此外, 图中 ZS-Base 表示混合类型 ZS-CQA 模型 (基线模型), ZS 表示 6 种单一类型的 ZS-CQA 模型, FT-base 表示混合类型 FT-CQA 模型 (基线模型). FT 表示 6 种单一类型的 FT-CQA 模型.

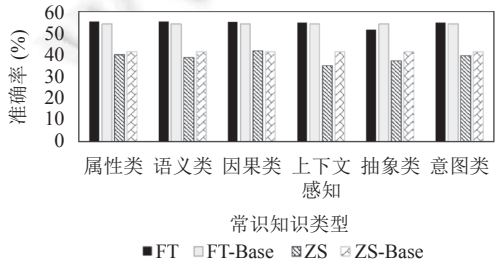

图 2 模型在 CSQA 上的实验结果

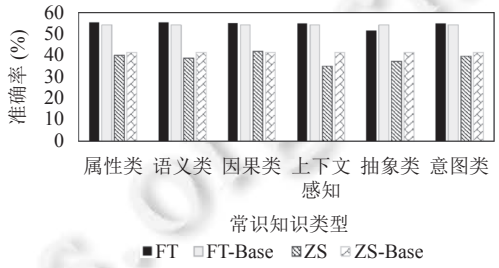

图 3 模型在 SIQA 上的实验结果

由图 2 可知, ZS 实验中, 只有因果类 ZS-CQA 模型在 CSQA 验证集上获得提升 (相比于基线). 但是, 实验采用 FT 实验设置时, 除抽象类 FT-CQA 模型外, 其余 5 种单一类型 FT-CQA 模型皆在 CSQA 上获得性能提升. 该现象说明, FT-CQA 系列模型具有较为普遍的优势. 其优势体现在两个方面. 其一是, 针对特定知识类的第一次微调增强了 CQA 模型的机器学习专注度; 其二是, 利用混合了异构类型知识 (领域数据训练集) 的样本进行二次微调, 拓展了 CQA 模型的机器学习视角. 这种先专注后发散的训练机制, 能够较为有效地提高问答精确性, 同时保证面向多类型数据的适应性.

由图 3 可知, 采用 ZS 实验设置, 单一类型 ZS-CQA 模型在 SIQA 验证集上取得不同程度的提升 (相比于基线), 其中, 利用属性类和抽象类知识的独立训练, 获得了较为显著的性能. 进一步地, 单一类型 FT-CQA 模型在其验证集上的性能均得以提升, 且与图 1 中人工观测的知识类型需求统计近似吻合. 以上结果说明, 先聚焦后发散的

知识学习方式具有普遍的优势. 同时说明, 对知识进行分类, 并选择合适的知识学习过程, 有助于完成大型 CQA 模型的逐步微调与优化.

6 总结与展望

CQA 是机器认知和理解常识知识, 并结合计算语言学形成智能运算和处理的重要任务. 从理论和关键技术层面开展 CQA 研究, 对问答领域的科技发展有着重要的意义. 本文回顾了现阶段 CQA 的主流研究趋势和代表性技术成果, 并在数据构建和知识分类上给与了详细介绍. 特别地, 本文借助两个专题介绍, 分别透视了基于预训练语言模型的各类 CQA 技术细节, 并验证了知识分类体系在优化现有 CQA 技术中的积极作用.

总体上, CQA 的发展状况可以总结为如下几个方面. 其一, 现有 CQA 数据集已从不同维度考察模型运用常识知识的能力. 其二, 以 Transformer 为基础的预训练语言模型, 从表示学习的角度推动了一系列新颖的 CQA 求解策略和设计思想. 其三, 以预训练语言模型为核心架构, 利用常识知识源进行知识加载和感知的技术路线, 已经产生较多关键技术成果. 其四, 本文综合前人研究证明了, CQA 数据集蕴涵的常识知识类型, 及其与常识知识源的匹配程度, 直接影响模型的知识感知和问题求解性能. 尽管如此, 目前 CQA 任务仍面临如下挑战.

(1) 常识知识源的完备性不足: 现有的常识知识源 (如 ConceptNet) 通过人工标注和规则方法生成, 具有比较高的质量和一定的数据规模. 但是, 全局常识知识数量庞大, 且具有动态发展和变化的特性. 因此, 现有常识知识源的完备性并不完善. 其导致的结果是, CQA 模型不足以回答实用阶段的所有问题. 扩展或补充现有知识源是一种潜在的优化手段, 但人工标注的成本昂贵且效率不高, 其面对不断更新的常识知识, 难以形成同步的更新.

(2) CQA 模型缺少深层次理解知识的能力: 现有实验表明, 预训练语言模型通过微调或采用与常识知识源融合的方式, 在一些 CQA 任务上取得较好的结果. 原因在于预训练语言模型本身的文本理解能力, 以及常识知识增强模型对知识的感知与浅层推理能力. 特别是注意力机制能够帮助模型着重关注常识知识或问题中重要的词元信息. 此外, 利用图结构的 CQA 模型, 能够在一定程度上获取结构化的关联线索, 形成逻辑可解释的推理过程. 然而, Kavumba 等人[86]发现 BERT 等预训练语言模型的表现会受到浅层线索 (比如问答文本的语义信息) 的影响, 对于 CQA 任务而言, 浅层线索往往不足以辅助模型推理出正确答案. 如何有效利用问题与常识知识源的信息, 并对其展开深层次的理解与推理, 仍是解决 CQA 任务的关键环节.

(3) 模型的鲁棒性和泛化能力不足: 近年来, 许多 CQA 研究集中在如何捕捉任务本身与常识知识之间的语言相关性, 以此作为求解答案的基础. Da 等人[87]研究显示 BERT 在时序和感知类知识问答上的性能偏低; Petroni 等人[88]和 Poerner 等人[89]的研究表明, 预训练语言模型仍在保留常识知识蕴涵的信息上仍有不足. 以上研究进一步说明, 目前的 CQA 模型仍存在泛化能力不足的问题. 此外, CQA 数据存在表述偏差 (即由于常识知识的隐含性, 知识、问题和答案的文字描述与实际现实之间存在的潜在差异), 由此导致的结果是, CQA 模型过度收敛于开发集的表述模式, 在测试集和实用过程中的真实性能往往存在较为显著的落差. 其体现了现有 CQA 模型的低鲁棒性问题.

(4) 缺乏对常识知识的评估: 目前, CQA 模型仅通过下游任务的表现, 反推常识知识运用的合理性, 并不具备透视推理内核的评估手段. 完善的评估标准可辅助 CQA 模型筛选高质量的常识知识作为推理基础. 现有研究尚未涉及知识获取质量的评估.

(5) 中文常识问答研究不足: 针对不同语种的常识问答研究, 离不开对应语种的数据集与知识库的构建. 在中文领域, Li 等人[90]以 Atomic 为基础, 提出并构建了第一个大规模的中文常识对话知识图谱 C^3KG. 值得指出的是, 在中文领域, 学术界仍然缺乏高质量的常识问答数据集. 由于中文与英文在表述模式 (语法结构)、常识知识范畴 (地域、习俗) 上, 存在语言层面的固有差异. 现有研究尚未在中文常识问答上展开深入的探讨.

基于上述不足, 本文建议从如下 5 个方面进一步推动 CQA 相关研究, 包括数据构建的可靠性、知识获取的主动性、推理过程的可解释性、知识蕴涵的深度挖掘和中文常识问答的深入探讨.

(1) 构建高质量的常识问答数据集: 向数据构建过程引入多类型的常识知识, 拓展知识类型的覆盖面, 从而辅

助 CQA 模型从更为宽泛的知识面中实施知识获取、线索挖掘和常识理解.

(2) 知识源更新和场景化挖掘: 研究动态跟踪知识发展的方法, 是未来 CQA 技术走向实用的重要环节. 特别地, 建立知识库和知识图谱的自动补全、自动更新及去冗余技术, 有着极为重要的应用价值. 此外, 实时挖掘相关于特定问题的知识, 是 CQA 技术中不可或缺的步骤. 预测问题所属领域, 特别是判断问题指向的真实场景, 并以领域和场景为约束, 精准获取相关知识、线索和结构, 对于建立精准化 CQA 有着极为重要的作用.

(3) 提高 CQA 推理过程的可解释性: 从自由文本和结构化知识图谱中挖掘线索, 并依据问题与线索之间的逻辑关系进行推理, 能够构成可解释性的推理链条. 但是, 线索的孤立性或局部的关联性, 对于语言的理解无法提供充分的上下文, 特别是线索的知识蕴涵或外延关系, 往往并未纳入线索挖掘的过程, 使得线索的理解存在天然的知识支撑. 因此, 跳出现有基于实体关系的单一逻辑关系, 拓展知识关联性的门类与结构, 从而发展线索的知识类画像, 以及多源关系的结构画像, 是形成可解释性答案推理的重要研究.

(4) 挖掘预训练语言模型中蕴涵的常识知识: 语言模型已在预训练阶段, 形成了感知语言特征和通用知识的能力, 其对于 CQA 模型的问题求解起到了重要的基础性作用 (即语言理解和知识感知作用). 但是, 这一感知过程是一种“黑盒”过程. 从而, 在处理各类特定领域的专有数据时, 预训练语言模型的感知内核无法得到评估和有效调整. 因此, 建立专门的任务形式和评估方法, 探测 CQA 感知内核的运算逻辑、知识认知程度和错误传递形式, 是提高预训练语言模型的迁移学习和微调效果的重要条件. 特别地, 这一技术能够为从根本上修正 CQA 内核的不足, 提供佐证与参考.

(5) 完善中文常识问答的研究: 常识问答研究已经形成了相对规范的研究模式, 主要包含基准数据集与常识知识库构建、常识知识获取和常识知识推理 3 部分. 设计符合中文常识范畴的常识问答数据集、构建大规模的常识知识库, 是展开中文常识问答研究的基础. 需要指出的是, 探究常识问答在中英文领域中的区别与联系, 有助于推动中文常识问答的研究.

References:

[1] Pujara J, Miao H, Getoor L, Cohen W. Knowledge graph identification. In: Proc. of the 12th Int'l Semantic Web Conf. Sydney: Springer, 2013. 542–557. [doi: 10.1007/978-3-642-41335-3_34]

[2] Wang X, Zou L, Wang CK, Peng P, Feng ZY. Research on knowledge graph data management: A survey. Ruan Jian Xue Bao/Journal of Software, 2019, 30(7): 2139–2174 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5841.htm [doi: 10.13328/j.cnki.jos.005841]

[3] Holland PW. Statistics and causal inference. Journal of the American statistical Association, 1986, 81(396): 945–960. [doi: 10.1080/01621459.1986.10478354]

[4] Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks. In: Proc. of the 2008 Conf. on Empirical Methods in Natural Language Processing. Honolulu: Association for Computational Linguistics, 2008. 1070–1079.

[5] Liu H, Singh P. ConceptNet—A practical commonsense reasoning tool-kit. BT Technology Journal, 2004, 22(4): 211–226. [doi: 10.1023/B:BTTJ.0000047600.45421.6d]

[6] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. arXiv:1409.2329, 2015.

[7] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

[8] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015. 1–15.

[9] Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[10] Liu Z, Lin W, Shi Y, Zhao J. A robustly optimized BERT pre-training approach with post-training. In: Proc. of the 20th Chinese National Conf. on Computational Linguistics. Huhhot: Chinese Information Processing Society of China, 2021. 1218–1227.

[11] Lewis M, Liu YH, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 7871–7880. [doi: 10.18653/v1/2020.acl-main.703]

[12] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss

A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1877–1901.

[13] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou YQ, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 2020, 21(1): 5485–5551.

[14] Yang Y, Kang S. Common sense-based reasoning using external knowledge for question answering. IEEE Access, 2020, 8: 227185–227192. [doi: 10.1109/ACCESS.2020.3045762]

[15] Lim J, Oh D, Jang Y, Yang K, Lim H. I know what you asked: Graph path learning using AMR for commonsense reasoning. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020. 2459–2471. [doi: 10.18653/v1/2020.coling-main.222]

[16] Ma KX, Francis J, Lu QY, Nyberg E, Oltramari A. Towards generalizable neuro-symbolic systems for commonsense question answering. In: Proc. of the 1st Workshop on Commonsense Inference in Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 22–32. [doi: 10.18653/v1/D19-6003]

[17] Li ZF, Zou BW, Li YQ, Jin ZL, Hong Y. Cascading commonsense question answering method base on multi-source knowledge fusion. Journal of Shanxi University (Natural Science Edition), 2022, 45(2): 264–273 (in Chinese with English abstract). [doi: 10.13451/j.sxu.ns.2021099]

[18] Wang RZ, Tang DY, Duan N, Wei ZY, Huang XJ, Ji JS, Cao GH, Jiang DX, Zhou M. K-adapter: Infusing knowledge into pre-trained models with adapters. In: Proc. of the 2021 Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021. 1405–1418. [doi: 10.18653/v1/2021.findings-acl.121]

[19] Wang PF, Peng NY, Ilievski F, Szekely P, Ren X. Connecting the dots: A knowledgeable path generator for commonsense question answering. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, 2020. 4129–4140. [doi: 10.18653/v1/2020.findings-emnlp.369]

[20] Talmor A, Herzig J, Lourie N, Berant J. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4149–4158. [doi: 10.18653/v1/N19-1421]

[21] Mihaylov T, Clark P, Khot T, Sabharwal A. Can a suit of armor conduct electricity? A new dataset for open book question answering. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 2381–2391. [doi: 10.18653/v1/D18-1260]

[22] Clark P, Cowhey I, Etzioni O, Khot T, Sabharwal A, Schoenick C, Tafjord O. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv:180305457, 2018.

[23] Sap M, Rashkin H, Chen D, Le Bras R, Choi Y. Social IQa: Commonsense reasoning about social interactions. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 4463–4473. [doi: 10.18653/v1/D19-1454]

[24] Huang LF, Le Bras R, Bhagavatula C, Choi Y. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 2391–2401. [doi: 10.18653/v1/D19-1243]

[25] Ostermann S, Modi A, Roth M, Thater S, Pinkal M. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In: Proc. of the 11th Int 'l Conf. on Language Resources and Evaluation. Miyazaki: European Language Resources Association (ELRA), 2018. 3567–3574.

[26] Ostermann S, Roth M, Pinkal M. MCScript2.0: A machine comprehension corpus focused on script events and participants. In: Proc. of the 8th Joint Conf. on Lexical and Computational Semantics. Minneapolis: Association for Computational Linguistics, 2019. 103–117. [doi: 10.18653/v1/S19-1012]

[27] Zhang S, Liu XD, Liu JJ, Gao JF, Duh K, van Durme B. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. arXiv:181012885, 2018.

[28] Boratko M, Li X, O'Gorman T, Das R, Le D, McCallum A. ProtoQA: A question answering dataset for prototypical common-sense reasoning. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 1122–1136. [doi: 10.18653/v1/2020.emnlp-main.85]

[29] Rahman A, Ng V. Resolving complex cases of definite pronouns: The Winograd schema challenge. In: Proc. of the 2012 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island: Association for Computational Linguistics, 2012. 777–789.

[30] Emami A, De La Cruz N, Trischler A, Suleman K, Cheung JCK. A knowledge hunting framework for common sense reasoning. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 1949–1958. [doi: 10.18653/v1/D18-1220]

[31] Cai Z, Tu LF, Gimpel K. Pay attention to the ending: Strong neural baselines for the ROC story cloze task. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers). Vancouver: Association for Computational Linguistics, 2017. 616–622. [doi: 10.18653/v1/P17-2097]

[32] Shwartz V, West P, Le Bras R, Bhagavatula C, Choi Y. Unsupervised commonsense question answering with self-talk. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 4615–4629. [doi: 10.18653/v1/2020.emnlp-main.373]

[33] Sun Y, Wang SH, Feng SK, Ding SY, Pang C, Shang JY, Liu JX, Chen XY, Zhao YB, Lu YX, Liu WX, Wu ZH, Gong WB, Liang JZ, Shang ZZ, Sun P, Liu W, Ouyang X, Yu DH, Tian H, Wu H, Wang HF. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv:2107.02137, 2021.

[34] Ma KX, Ilievski F, Francis J, Bisk Y, Nyberg E, Oltramari A. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. Proc. of the AAAI Conf. on Artificial Intelligence, 2021, 35(15): 13507–13515. [doi: 10.1609/aaai.v35i15.17593]

[35] Lin BY, Chen XY, Chen JM, Ren X. KagNet: Knowledge-aware graph networks for commonsense reasoning. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 2829–2839. [doi: 10.18653/v1/D19-1282]

[36] Lv SW, Guo DY, Xu JJ, Tang DY, Duan N, Gong M, Shou LJ, Jiang DX, Cao GH, Hu SL. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(5): 8449–8456. [doi: 10.1609/aaai.v34i05.6364]

[37] Li YQ, Zou BW, Li ZF, Aw AT, Hong Y, Zhu QM. Winnowing knowledge for multi-choice question answering. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. Punta Cana: Association for Computational Linguistics, 2021. 1157–1165. [doi: 10.18653/v1/2021.findings-emnlp.100]

[38] Roemmele M, Bejan CA, Gordon AS. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In: Proc. of the 2011 AAAI Spring Symp. Stanford: AAAI, 2011. 90–95.

[39] Levesque HJ, Davis E, Morgenstern L. The winograd schema challenge. In: Proc. of the 13th Int'l Conf. on the Principles of Knowledge Representation and Reasoning. Rome: Institute of Electrical and Electronics Engineers, 2012. 552–561.

[40] Sharma A, Vo NH, Aditya S. Baral C. Towards addressing the winograd schema challenge: Building and using a semantic parser and a knowledge hunting module. In: Proc. of the 24th Int'l Joint Conf. on Artificial Intelligence. Buenos Aires: AAAI, 2015. 1319–1325.

[41] Mostafazadeh N, Chambers N, He XD, Parikh D, Batra D, Vanderwende L, Kohli P, Allen J. A corpus and cloze evaluation for deeper understanding of commonsense stories. In: Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016. 839–849. [doi: 10.18653/v1/N16-1098]

[42] Zellers R, Bisk Y, Schwartz R, Choi Y. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 93–104. [doi: 10.18653/v1/D18-1009]

[43] Han K, Xiao A, Wu EH, Guo JY, Xu CJ, Wang YH. Transformer in Transformer. In: Proc. of the 35th Conf. on Neural Information Processing Systems. Sydney, 2021. 15908–15919.

[44] Improving language understanding by generative pre-training. 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[45] Yang ZL, Dai ZH, Yang YM, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized autoregressive pretraining for language understanding. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver, 2019. 5753–5763.

[46] Sap M, Le Bras R, Allaway, E, Bhagavatula C, Lourie N, Rashkin H, Roof B, Smith NA, Choi Y. Atomic: An atlas of machine commonsense for if-then reasoning. Proc. of the AAAI Conf. on Artificial Intelligence, 2019, 33(1): 3027–3035. [doi: 10.1609/aaai.v33i01.33013027]

[47] Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge: MIT Press, 1998. 21–120.

[48] Krishna R, Zhu YK, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA, Bernstein MS, Fei-Fei L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int'l Journal of Computer Vision, 2017, 123(1): 32–73. [doi: 10.1007/s11263-016-0981-7]

[49] Vrandečić D, Krötzsch M. Wikidata: A free collaborative knowledge base. Communications of the ACM, 2014, 57(10): 78–85. [doi: 10.1145/2629489]

[50] Singh P, Lin T, Mueller ET, Lim G, Perkins T, Zhu WL. Open mind common sense: Knowledge acquisition from the general public. In: Proc. of the 2002 OTM Confederated Int'l Conf. Berlin: Springer, 2002. 1223–1237. [doi: 10.1007/3-540-36124-3_77]

[51] Burton K, Java A, Soboroff I. The ICWSM 2009 Spinn3r dataset. In: Proc. of the 3rd Annual Conf. on Weblogs and Social Media. San Jose: AAAI, 2009.

[52] Modi A, Anikina T, Ostermann S, Pinkal M. InScript: Narrative texts annotated with script information. In: Proc. of the 10th Int'l Conf. on Language Resources and Evaluation. 2016. 3485–3493.

[53] Cambria E, Song YQ, Wang HX, Hussain A. Isanette: A common and common sense knowledge base for opinion mining. In: Proc. of the 11th IEEE Int'l Conf. on Data Mining Workshops. Vancouver: IEEE, 2011. 315–322. [doi: 10.1109/ICDMW.2011.106]

[54] Ilievski F, Oltramari A, Ma KX, Zhang B, McGuinness DL, Szekely P. Dimensions of commonsense knowledge. Knowledge-based Systems, 2021, 229: 107347. [doi: 10.1016/j.knosys.2021.107347]

[55] Lenat DB, Guha RV, Pittman K, Pratt D, Shepherd M. Cyc: Toward programs with common sense. Communications of the ACM, 1990, 33(8): 30–49. [doi: 10.1145/79173.79176]

[56] Mostafazadeh N, Kalyanpur A, Moon L, Buchanan D, Berkowitz L, Biran O, Chu-Carroll J. GLUCOSE: Generalized and contextualized story explanations. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 4569–4586. [doi: 10.18653/v1/2020.emnlp-main.370]

[57] Tandon N, De Melo G, Weikum G. WebChild 2.0: Fine-grained commonsense knowledge distillation. In: Proc. of the 2017 ACL, System Demonstrations. Vancouver: Association for Computational Linguistics, 2017. 115–120.

[58] Romero J, Razniewski S, Pal K, Pan JZ, Sakhadeo A, Weikum G. Commonsense properties from query logs and question answering forums. In: Proc. of the 28th ACM Int'l Conf. on Information and Knowledge Management. Beijing: ACM, 2019. 1411–1420. [doi: 10.1145/3357384.3357955]

[59] Cambria E, Li Y, Xing FZ, Poria S, Kwok K. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In: Proc. of the 29th ACM Int'l Conf. on Information & Knowledge Management. Ireland: ACM, 2020. 105–114. [doi: 10.1145/3340531.3412003]

[60] Bhakthavatsalam S, Richardson K, Tandon N, Clark P. Do dogs have whiskers? A new knowledge base of haspart relations. arXiv: 2006.07510, 2020.

[61] Wu WT, Li HS, Wang HX, Zhu KQ. Probase: A probabilistic taxonomy for text understanding. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. Scottsdale: ACM, 2012. 481–492. [doi: 10.1145/2213836.2213891]

[62] Cambria E, Song YQ, Wang HX, Howard N. Semantic multidimensional scaling for open-domain sentiment analysis. IEEE Intelligent Systems, 2014, 29(2): 44–51. [doi: 10.1109/MIS.2012.118]

[63] Tanon TP, Weikum G, Suchanek F. YAGO 4: A reason-able knowledge base. In: Proc. of the 17th Int'l Conf. on the Semantic Web. Heraklion: Springer, 2020. 583–596. [doi: 10.1007/978-3-030-49461-2_34]

[64] Guha RV, Brickley D, Macbeth S. Schema.org: Evolution of structured data on the Web. Communications of the ACM, 2016, 59(2): 44–51. [doi: 10.1145/2844544]

[65] Gangemi A, Guarino N, Masolo C, Oltramari A, Schneider L. Sweetening ontologies with DOLCE. In: Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. Berlin: Springer, 2002. 166–181. [doi: 10.1007/3-540-45810-7_18]

[66] Niles I, Pease A. Towards a standard upper ontology. In: Proc. of the 2001 Int'l Conf. on Formal Ontology in Information Systems. Ogunquit: ACM, 2001. 2–9. [doi: 10.1145/505168.505170]

[67] Bennett JS. ROGET: A knowledge-based system for acquiring the conceptual structure of a diagnostic expert system. Journal of Automated Reasoning, 1985, 1(1): 49–74. [doi: 10.1007/BF00244289]

[68] Baker CF, Fillmore CJ, Lowe JB. The berkeley framenet project. In: Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th Int'l Conf. on Computational Linguistics. Montreal: ACM, 1998. 86–90. [doi: 10.3115/980845.980860]

[69] Dodge EK, Hong J, Stickles E. MetaNet: Deep semantic automatic metaphor analysis. In: Proc. of the 3rd Workshop on Metaphor in NLP. Denver: Association for Computational Linguistics, 2015. 40–49. [doi: 10.3115/v1/W15-1405]

[70] Schuler KK. VerbNet: A Broad-coverage, Comprehensive Verb Lexicon. Philadelphia: University of Pennsylvania, 2005.

[71] Bhakthavatsalam S, Anastasiades C, Clark P. GenericsKB: A knowledge base of generic statements. arXiv:200500660, 2020.

[72] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1(8): 9.

[73] Bosselut A, Rashkin H, Sap M, Malaviya C, Celikyilmaz A, Choi Y. COMET: Commonsense transformers for automatic knowledge graph construction. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 4762–4779. [doi: 10.18653/v1/P19-1470]

[74] Xu YH, Zhu CG, Xu RC, Liu Y, Zeng M, Huang XD. Fusing context into knowledge graph for commonsense question answering. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. Association for Computational Linguistics, 2021. 1201–1207. [doi: 10.18653/v1/2021.findings-acl.102]

[75] Feng YL, Chen XY, Lin BY, Wang PF, Yan J, Ren X. Scalable multi-hop relational reasoning for knowledge-aware question answering. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 1295–1309. [doi: 10.18653/v1/2020.emnlp-main.99]

[76] Yasunaga M, Ren HY, Bosselut A, Liang P, Leskovec J. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021. 535–546. [doi: 10.18653/v1/2021.naacl-main.45]

[77] Sun YQ, Shi Q, Qi L, Zhang Y. JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering. In: Proc. of the 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: Association for Computational Linguistics, 2021. 5049–5060. [doi: 10.18653/v1/2022.naacl-main.372]

[78] Wang JX, Li XY, Tan Z, Zhao X, Xiao WD. Relation-aware bidirectional path reasoning for commonsense question answering. In: Proc. of the 25th Conf. on Computational Natural Language Learning. Association for Computational Linguistics, 2021. 445–453. [doi: 10.18653/v1/2021.conll-1.35]

[79] Yan J, Raman M, Chan A, Zhang TY, Rossi R, Zhao HD, Kim S, Lipka N, Ren X. Learning contextualized knowledge structures for commonsense reasoning. In: Proc. of the 2021 Findings of the Association for Computational Linguistics (ACL-IJCNLP 2021). Association for Computational Linguistics, 2021. 4038–4051. [doi: 10.18653/v1/2021.findings-acl.354]

[80] Ye ZX, Chen Q, Wang W, Ling ZH. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. arXiv:1908.06725, 2020.

[81] Niu YL, Huang F, Liang JM, Chen WK, Zhu XY, Huang ML. A semantic-based method for unsupervised commonsense question answering. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing (Vol. 1: Long Papers). Association for Computational Linguistics, 2021. 3037–3049. [doi: 10.18653/v1/2021.acl-long.237]

[82] Lourie N, Le Bras R, Bhagavatula C, Choi Y. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. Proc. of the AAAI Conf. on Artificial Intelligence, 2021, 35(15): 13480–13488. [doi: 10.1609/aaai.v35i15.17590]

[83] Khashabi D, Min S, Khot T, Sabharwal A, Tafjord O, Clark P, Hajishirzi H. UNIFIEDQA: Crossing format boundaries with a single QA system. In: Proc. of the 2020 Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, 2020. 1896–1907. [doi: 10.18653/v1/2020.findings-emnlp.171]

[84] Clark K, Luong MT, Le QV, Manning CD. ELECTRA: Pre-training text encoders as discriminators rather than generators. arXiv:2003.10555, 2020.

[85] Ilievski F, Szekely P, Zhang B. CSKG: The commonsense knowledge graph. In: Proc. of the 18th Int'l Conf. on the Semantic Web. Springer, 2021. 680–696. [doi: 10.1007/978-3-030-77385-4_41]

[86] Kavumba P, Inoue N, Heinzerling B, Heinzerling B, Singh K, Reisert P, Inui K. When choosing plausible alternatives, clever hans can be clever. In: Proc. of the 1st Workshop on Commonsense Inference in Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 33–42. [doi: 10.18653/v1/D19-6004]

[87] Da J, Kasai J. Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. In: Proc. of the 1st Workshop on Commonsense Inference in Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 1–12. [doi: 10.18653/v1/D19-6001]

[88] Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu YX, Miller A. Language models as knowledge bases? In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 2463–2473. [doi: 10.18653/v1/D19-1250]

[89] Poerner N, Waltinger U, Schütze H. E-BERT: Efficient-yet-effective entity embeddings for BERT. In: Proc. of the 2020 Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, 2020. 803–818. [doi: 10.18653/v1/2020.findings-emnlp.71]

[90] Li DW, Li YR, Zhang JY, Li K, Wei C, Cui JW, Wang B. C^{3}KG: A Chinese commonsense conversation knowledge graph. In: Proc. of the 2022 Findings of the Association for Computational Linguistics: ACL 2022. Dublin: Association for Computational Linguistics, 2022.

1369–1383. [doi: 10.18653/v1/2022.findings-acl.107]

附中文参考文献:

[2] 王鑫, 邹磊, 王朝坤, 彭鹏, 冯志勇. 知识图谱数据管理研究综述. 软件学报, 2019, 30(7): 2139–2174. http://www.jos.org.cn/1000-9825/5841.htm [doi: 10.13328/j.cnki.jos.005841]

[17] 李志峰, 邹博伟, 李烨秋, 金志凌, 洪宇. 基于多知识源融合的级联式常识问答方法. 山西大学学报(自然科学版), 2022, 45(2): 264–273. [doi: 10.13451/j.sxu.ns.2021099]

范怡帆(1996－), 女, 博士生, CCF 学生会员, 主要研究领域为常识问答.

李志峰(1998－), 男, 硕士生, CCF 学生会员, 主要研究领域为深度学习, 自然语言处理, 常识问答.

邹博伟(1984－), 男, 博士, 研究员, 主要研究领域为自然语言处理, 文本生成, 自动问答.

洪宇(1978－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为信息抽取, 篇章关系理解, 多模态机器翻译, 智能问答.

徐庆婷(1994－), 女, 博士生, CCF 学生会员, 主要研究领域为信息抽取.