

文章编号: 1003-0077(2023)08-0001-17

基于深度学习的多文档机器阅读理解综述

高峰^{1,2}, 倪建成³, 高鹏¹, 周子力¹, 李艳艳¹

(1. 曲阜师范大学 网络空间安全学院, 山东 曲阜 273165;

2. 华东师范大学 计算机科学与技术学院, 上海 200062;

3. 曲阜师范大学 网络信息中心, 山东 曲阜 273165)

摘要: 多文档机器阅读理解是利用计算机同时将多个文档的语义信息进行阅读理解、筛选组合后进行问题作答的过程。与传统的单文档机器阅读理解相比, 其更关注问题约束下的文档间语义逻辑关系理解, 是自然语言处理领域中新兴的研究方向之一。该文首先选取了多文档机器阅读理解的重点数据集, 涵盖多源信息型数据集、线性推理型数据集和细粒度半结构化型数据集等以分析任务发展趋势; 其次调研了深度学习领域中主流的评价指标和研究方法, 划分了具有代表性的四层阅读架构并分别展开了详细的讨论, 特别在阅读理解层中选取了文档选择、Transformer 语言模型、图卷积神经网络和外部知识融合等阅读方法以分析任务的研究现状; 最后对多文档阅读理解的研究进行了总结与展望, 提出了后续研究面临的问题与挑战。

关键词: 综述; 多文档机器阅读理解; 深度学习

中图分类号: TP391

文献标识码: A

Survey of Multi-document Machine Reading Comprehension Based on Deep Learning

GAO Feng^{1,2}, NI Jiancheng³, GAO Peng¹, ZHOU Zili¹, LI Yanyan¹

(1. School of Cyber Science and Engineering, Qufu Normal University, Qufu, Shandong 273165, China;

2. School of Computer Science and Technology, East China Normal University, Shanghai 200062, China;

3. Network Information Center, Qufu Normal University, Qufu, Shandong 273165, China)

Abstract: Multi-document machine reading comprehension requires the machine to read and comprehend multi documents simultaneously, so as to answer subsequent questions. Compared with the traditional single-document machine reading comprehension, it is more concerned with the semantic logic among the documents within the constraints of questions. The paper firstly explores the main datasets for multi-document machine reading comprehension to analyze its development trends, including multi-source datasets, linear inference datasets, and fine-grained semi-structured datasets. Secondly, it investigates the mainstream evaluation metrics, then a prototype of four-tier reading structure is discussed in detail. This paper also reviews the methods based on document selection, Transformers, Graph Convolution Networks, and external knowledge fusion. Finally, it gives an outlook on the research of multi-document reading comprehension, as well as the problems and challenges in this study.

Keywords: survey; multi-document machine reading comprehension; deep learning

1 多文档机器阅读理解的发展与任务定义

机器阅读理解(Machine Reading Comprehension, MRC)一直以来是人工智能领域中一项具有挑战

性的研究任务, 是检验机器是否具有人工智慧的一项重要参考^[1]。作为具有较长研究历史的任务, 按研究方法区分, 机器阅读理解经历了三个较为清晰的阶段: 早期系统时代(1970—2012)、传统机器学习时代(2013—2015)、深度学习时代

收稿日期: 2021-02-24 定稿日期: 2021-05-07

基金项目: 曲阜师范大学高层次人才科研基金(602801)

(2015—)[2]。得益于深度学习理论和框架的成熟,机器阅读理解任务目前正处于繁荣发展的阶段,从传统的单文档阅读理解任务逐渐转向对阅读理解能力、推理能力和概括能力要求更高的多文档机器阅读理解任务,形成了以数据驱动方法创新的局面。

1.1 单文档机器阅读理解

单文档机器阅读理解(Single-Document Machine Reading Comprehension)是阅读理解的最初任务形

式,在深度学习时代其可被形式化为一个有监督的学习问题:对于给定的问题 q 、文本段落 d ,要求模型 f 输出预测答案 a ,如式(1)所示。

$$f:q,d \rightarrow a \quad (1)$$

其中,问题可表示为 $q=\{w_1^q,w_2^q,\dots,w_m^q\}$,段落可表示为 $d=\{w_1^d,w_2^d,\dots,w_n^d\}$, w 为文本分词, m 和 n 分别表示为问题和段落的分词后长度。根据数据集任务的不同, a 的形式不尽相同,一般可以将其划分为三种形式,即分类式答案、抽取式答案和生成式答案,详细描述如表1所示。

表1 单文档机器阅读理解任务形式描述

任务形式	描述
分类式答案	完形填空:在文段的指定位置设置预测槽,要求模型根据上下文从给定的候选答案中选择最合适的单词或实体,填入预测槽后使语句通顺并符合语境语义
	多项选择:给定文段信息,要求模型能够在问题和针对此问题给出的多个潜在答案的约束下,选择出最符合文意的答案,使得文段、问题与预测的选项三者存在语义或事实关联,而根据数据集不同,候选答案数量可能保持全数据集一致,也可能在不同样本间存在显著差异
抽取式答案	也称作“语段预测”(Span Prediction),要求模型根据给定的文段和问题,从原文中提取一个或多个文本语段[3](Text Span)作为答案输出
生成式答案	当数据集的答案为人工生成而非来自原文时,要求模型从原文提取与问题相关的语义信息,在模型内部进行筛选处理并通过字典输出生成后的答案文本

1.2 多文档机器阅读理解

多文档机器阅读理解(Multi-Document Machine Reading Comprehension)是在单文档的基础上,将给定的文档数量从单个扩增为多个,并进一步关注问题与多个文档间的内部逻辑关系。因此,多文档机器阅读理解任务对模型的理解、推理和概括能力要求更高,引领了利用机器理解人类复杂语言逻辑关系的研究浪潮。在式(1)的基础上,多文档机器阅读理解任务的形式化如下:对于给定问题 q 、多个文档集合 D ,要求预测器 f 能够输出预测答案 a ,如式(2)所示。

$$f:q,D \rightarrow a \quad (2)$$

其中,文档集合 $D=\{d_1,d_2,\dots,d_r\}$, r 表示文档数量;第 i 个文档 $d_i=\{w_1^{d_i},w_2^{d_i},\dots,w_s^{d_i}\}$, s 表示文档序列分词后的长度;答案 a 可与单文档阅读理解共享任务形式。

1.3 多文档与单文档机器阅读理解的联系与区别

随着大规模预训练语言模型的出现,在多个单文档机器阅读理解数据集任务上,人类的作答结果已逐渐被机器模型超越[4]。如何以贴近人类思维的

方式完成更复杂的机器阅读理解任务,成为了阅读领域的研究热点,其中推理能力成为了研究主流方向之一,如SQuAD 2.0[5]数据集在SQuAD 1.1[6]的基础上,要求模型不仅能正确地作答问题,还要求模型在无法根据给定文档作出问题的正确答案时,主动停止预测,以判断模型是否对问题和文档有足够的理解能力,而不是强行作答。

多文档机器阅读理解可以视作单文档机器阅读理解的升级任务,其内部存在紧密的联系。首先,两者作为自然语言处理领域中的文本理解类任务,前序处理流程相似,如利用文本嵌入将自然语言映射至向量空间中,或通过编码层将线性文本信息转化为序列编码等;其次,两者可共享阅读理解任务的形式与评价指标,即使多文档阅读更注重文档间的关系理解和逻辑推理,然而在具体的任务形式上,虽然两者仍可保持一致,如中文单文档阅读数据集C[7]与英文多文档阅读数据集WikiHop[8],两者给定的阅读文档数量不同,但问题和答案的形式均是多项选择问答。

多文档阅读理解与单文档阅读理解的区别主要表现在以下三个方面:

信息源多元化 与单文档阅读理解仅有单个文

档不同,多文档阅读理解需要协调处理来自多个文档的信息。由于不同数据集的创建方式和任务定义不同,同一样本下的多个文档可能存在不同的语义关系,如文档间存在观点矛盾^[9-10],或需在多个文档中进行线性推理^[8],甚至文档集合内存在噪声文档干扰正常的阅读推理^[11]等,这些现象都增加了模型阅读理解的难度,需要进行适应化处理。

对推理能力与可解释性要求的提升 由于多文档的阅读过程更贴近人类阅读时复杂的信息选择过程,研究者们大多通过多种启发式方法为模型赋予推理能力,以提升任务表现与鲁棒性。如在阅读过程中引入事理、知识图谱等外部知识,将信息转换为模型内部的判断机理,可以进一步提升推理过程的感知和分析能力^[12]。而阅读过程的可解释性,则需要分析模型的预测过程是否符合人类思维认知的模式,进而探讨在不同数据集或应用场景下模型的泛化可能。

对时空伸缩性的要求 无文档数量上限的多文档阅读理解任务,有时会导致答案搜索空间的爆炸,这就要求算法模型在作答时具有协调时空消耗的能力:即使文档规模增大,但在回答问题时仍需要保

持相对稳定的响应时间,同时也要有策略地选择文档集合并将其调入模型,以控制阅读时机器的算力和存储开销,保证问答系统运行的稳定。

总之,多文档机器阅读理解脱胎于单文档机器阅读理解,但提出了更复杂的逻辑推理要求,其形式也更接近人类阅读的真实过程^[13],这决定了其在具有较高研究难度的同时,拥有广阔的商业应用前景。

2 多文档机器阅读理解数据集与评价指标

2.1 多文档机器阅读理解数据集

在深度学习领域,高质量的海量数据集将直接助益模型达到更好的效果^[13],进一步推动研究的发展。我们选取了8个有代表性的多文档阅读理解数据集,从数据集创建的驱动原因、语料来源、生成方式、任务特点和应用意义等角度对它们进行了总结,并在表2中详细展示各项指标参数,以期从数据集的角度呈现多文档机器阅读理解的发展趋势。

表2 多文档阅读理解数据库(2016—2020年)横向比较

数据集名称	MS MARCO	TriviaQA	DuReader	WikiHop	MedHop	HotpotQA	AmazonQA	R ⁴ C
发布时间	2016	2017	2018	2018	2018	2018	2019	2020
数据集语言	英文	英文	中文	英文	英文	英文	英文	英文
发布机构	Microsoft AI & Research	University of Washington, Allen Institute for Artificial Intelligence	Baidu Inc.,	University College London, Bloomsbury AI	University College London, Bloomsbury AI	Carnegie Mellon University, Stanford University, Université de Montréal, CIFAR, Google AI	Carnegie Mellon University	Tohoku University, RIKEN, University College London
发表地点	NIPS 2016	ACL 2017	QA@ACL 2018	TACL	TACL	EMNLP 2018	IJCAI 2019	ACL 2020
数据领域	开放域	开放域	开放域	开放域	封闭域	开放域	封闭域	开放域
任务形式	生成式	抽取式	生成式	分类式	分类式	抽取式	生成式	抽取式
数据来源	人工问题、网络文档、人工答案	人工问题、网络文档、人工答案	人工问题、网络文档、人工答案	事实问题、维基百科、事实答案	事实问题、Medline、事实答案	事实问题、维基百科、事实答案	人工问题、商品评论、人工答案	人工问题、网络文档、人工答案
数据集规模 Q: 问题 A: 答案 D: 文档 单位: 个	Q&A: 1 010 916 D: 8 841 823	Q&A: 95 956 D: 662 659	Q: 201K D: 946K A: 417K	Q: 51 318 D: 3-63/Q A: 2-79/Q	Q: 2 508 D: 5-64/Q A: 2-9/Q	Q&A: 112 779 D: 10/Q in Distractor Setting	Q: 923K D: 14M A: 3.6M	Q&A: 4 588 D: 10/Q

续表

数据集名称	MS MARCO	TriviaQA	DuReader	WikiHop	MedHop	HotpotQA	AmazonQA	R ⁴ C
数据集官网	https://microsoft.github.io/msmarco/	http://nlp.cs.washington.edu/triviaqa/	https://ai.baidu.com/broad/introduction?dataset=dureader	http://qan-garoo.cs.ucl.ac.uk	http://qan-garoo.cs.ucl.ac.uk	https://hotpotqa.github.io	https://github.com/amazonqa/amazonqa	https://naoya-i.github.io/r4c/
数据集排行榜	https://microsoft.github.io/msmarco/	https://competitions.codalab.org/competitions/17208	https://ai.baidu.com/broad/leaderboard?dataset=dureader	http://qan-garoo.cs.ucl.ac.uk/leaderboard.html	http://qan-garoo.cs.ucl.ac.uk/leaderboard.html	https://hotpotqa.github.io	N.A	https://naoya-i.github.io/r4c/
测评方式	盲测	盲测	盲测	盲测	盲测	盲测	自测	盲测
评价指标	BLEU-1 ROUGE-L	EM F_1	BLEU-4 ROUGE-L	准确率	准确率	EM 联合 F_1	BLEU ROUGE	联合 F_1
DBLP	https://dblp.org/rec/conf/nips/NguyenRSGTMD16.html	https://dblp.org/rec/conf/acl/JoshiCWZ17.html	https://dblp.org/rec/conf/acl/HeLLLZXLWWSLWW18	https://dblp.org/rec/journals/tacl/WelblSR18	https://dblp.org/rec/journals/tacl/WelblSR18	https://dblp.org/rec/conf/emnlp/Yang0ZBCSM18	https://dblp.org/rec/conf/ijcai/GuptaKCRL19	https://dblp.org/rec/conf/acl/InoueSI20

2.1.1 MS MARCO

2016 年, Nguyen 等人^[10] 认识到当时的阅读理解数据集严重依赖于文档中显式的信息, 缺少对内容的深入理解, 因此基于微软必应搜索引擎的用户搜索记录和网页文档, 构建了能反映真实世界中人类问题需求的多文档机器阅读理解数据集——MS MARCO。每样本包含 6 个字段: ①用户在搜索时预期可通过搜索引擎直接获得答案的问题; ②10 个来自必应搜索引擎并经过人工确认与问题相关的独立文档; ③由人工从文档中直接抽取的答案; ④由人工根据文档重新编写的答案, 并要求此答案比③更符合自然语言的语法与逻辑, 且保证与原文档的内容具有一定区分度; ⑤文档的原文链接、原文标题、原主体文本等未经加工的源信息; ⑥对问题的分类标签, 如“数字型”“实体型”“地点型”“人物型”或“描述型”等。MS MARCO 数据集是一项里程碑式的工作, 它较早地认识到利用搜索引擎日志, 从用户的主观意图和海量网络文档入手, 产出可反馈人类社会真实需求的多文档阅读理解数据集。

2.1.2 TriviaQA

2017 年, Joshi 等人^[14] 提出, 如果直接将智力问答竞赛中涉及多个知识领域的赛题作为阅读理解任务的问题, 不仅可以避免需要人工针对文档重新编

造问题的繁琐过程, 还能减少因问题文本语言风格或问题与文档信息相关性过强而导致的数据偏置问题。他们基于约 9.5 万个智力问答“问题-答案”对, 将维基百科与网络文档作为语料库, 创建了面向智力问答的多文档阅读理解数据集——TriviaQA。平均每个样本有 6 个相关文档用于答案的预测。

与其他数据集相比, TriviaQA 有 3 个特点: (1) 不仅涵盖了多个领域的智力挑战问题, 而且真实地迁移了人类社会在“问答”方面的兴趣爱好到机器世界中; (2) 通过对随机采样样本进行定性分析, 发现分别在维基百科抽取到和从网络文档筛选的两种文档集合中, 能够涵盖正确回答问题的比例分别为 79.7% 和 75.4%, 同时, 与 SQuAD 相比, 超过三倍以上的问题需要通过对多个句子的推理才能得出答案, 证明了其有效性和全面性; (3) TriviaQA 不仅可用于信息检索式阅读理解, 而且为构建基于结构化知识的问答系统提供了可用素材, 为文本结合外部知识的阅读问答方法研究奠定了语料基础。

2.1.3 DuReader

2018 年, 为了增添阅读理解数据集语言的多样性, He 等人^[9] 利用百度搜索引擎日志和大型互助问答社区“百度知道”, 建立了中文的多文档阅读理解问答数据集——DuReader, 创新性地引入了针对

答案是否为“主观观点/客观事实”和“实体性/描述性/是否命题”的两种分类标准,利用人工汇总多个网络文档和百度知道社区回答,生成针对问题的人工准确答案。DuReader 数据集包含了约 20 万个问题,42 万个答案和 100 万个文档,是发布时最大的中文阅读理解数据集。

2.1.4 WikiHop

2018 年, Welbl 等人^[8]认为当时的多文档阅读理解数据集中缺少对文档间逻辑推理的要求,因此他们基于 Wikidata^①的关系三元组和维基百科,利用有向二部图,构建了文档间存在线性逻辑关系的开放域多文档阅读理解数据集 QAngaroo WikiHop。首先将先验知识三元组 (Subject Entity, Relation, Object Entity), 转化为问答的形式: Question = (Subject Entity, Relation, ?), Answer = (Object Entity); 然后构建如图 1 所示的有向二部图, 左侧为实体节点, 右侧为文档节点, 当文档中包含某个实体时则从文档连接至实体; 最后利用广度优先搜索 (Breadth-First Search) 算法构建从 Subject Entity 到 Object Entity 的搜索路径, 收集路径中命中的文档与相关的实体并保存为一个样本, 其中与答案实体类型一致的其他实体将被保存为问题的干扰候选项。特别地, 若在收集的单个样本中, 存在多个同时成立的先验知识, 即在答案实体和干扰项中同时满足先验知识 (s, r, o) 与先验知识 (s, r, o') , 则舍弃候选项 o' , 以保证在同一问题下存在唯一的答案 o 。图 1 中有底色的项目表示在搜索时命中的文档和实体, 对号表示为正确答案, 叉号表示为候选项中的干扰答案。为了验证数据集的有效性, Welbl 等人对开发集与测试集中的每条数据邀请了 3 名志愿者, 进行问题与文档间逻辑关系的验证, 结果显示有约 74% 的数据遵循或可能遵循“回答问题需要在多个文档间进行逻辑推理”的假设。

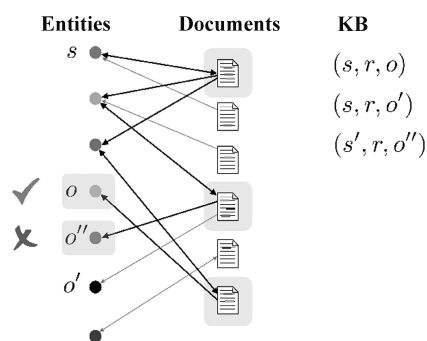


图 1 WikiHop 数据集创建示意图

引用自原论文

2.1.5 MedHop

近年来, 分子生物学 (Molecular Biology) 的出版物数量呈指数级增长^[15], 而且其相关的数据分析经历了严格生物化学实验验证, 具有较强的科学性和客观性。在此背景下 Welbl 等人^[8]以 Drug-Bank^[16-17]为结构化知识、Medline 药物文献摘要为语料库, 复用 WikiHop 的创建方式, 提出了面向封闭域的 QAngaroo MedHop^[8]数据集, 目的在于利用记录了药物生化性质的文献摘要, 通过机器阅读理解的形式进行药物间反应 (Drug-Drug Interactions, DDIs) 预测。

药物反应预测涉及药物-蛋白质反应、蛋白质-蛋白质反应等多种生物化学过程, 因此 MedHop 数据集具有一定的挑战性, 其将多文档机器阅读任务的研究领域扩展到了分子生物学, 使用完全基于自然语言理解的方式对 DDIs 进行预测。利用科学、严谨和定量分析的生物化学研究成果能促进自然语言领域的相关研究, 而自然语言领域预测的结果又可以有针对性地指导生物化学的相关实验过程, 提高实验效率, 减少实验损耗, 实现两研究领域的共同发展。

2.1.6 HotpotQA

2018 年, Yang 等人^[4]指出多文档阅读理解任务中存在的一个关键缺陷, 即当时的数据集不能在阅读理解过程中明确地提供文档间前进或后退的逻辑关系证据, 进而导致在学习过程只关注如何预测答案, 却忽视了分析预测过程的逻辑性。为此, Yang 等人通过人工标记出在人类思维阅读、推理时使用的相关句子, 以 Wikipedia 为语料库创建了带有摘选标记的多文档数据集——HotpotQA。HotpotQA 有 4 个主要特征: (1) 问题需要在提供的多个文档中进行选择和推理才能回答; (2) 问题形式多样, 且问题的生成未依赖于任何先验知识; (3) 提供了语句级别的支持性逻辑链条, 帮助模型学习人类的阅读推理过程, 并提出了对模型的回答过程进行合理性验证的要求; (4) 新增了两实体间属性比较的问题种类, 用于测评模型提取信息和比较信息的能力。

HotpotQA 数据集具有 112 779 个样本, 其中训练集、开发集和测试集包含的样本数量分别为 90 564 个、7 405 个和 7 405 个。由于此数据集的设计理念先进, 更贴合人类阅读理解习惯, 在发布后迅

① <https://www.wikidata.org>

速得到了大量研究者的关注和评测提交。

2.1.7 AmazonQA

2019 年, Gupta 等人^[18]观察到每天有几千名的顾客在亚马逊购物网站的商品详情页进行提问, 而这些问题能在广泛地浏览来自其他用户的评论后得出答案。因此, Gupta 等人从亚马逊网站 15.6 万件商品中的 1 400 万位用户的评论文档中收集了 92.3 万个问题和 360 万个答案, 开发了基于用户社区商品问答的多文档机器阅读理解数据集——AmazonQA。与其他数据集相比, 该数据集着眼于电子商务领域的信息, 完整地反映真实用户的问题需求, 是发布时最大的描述性问答数据集。为了训练与检测模型对数据的理解能力, 与 SQuAD 2.0 类似, AmazonQA 提供了是否可以通过文档回答问题的指示标记。

2.1.8 R⁴C

在实际应用中, 类似于 HotpotQA 提供的支持性事实(Supporting Facts)是以句子形式出现的, 而语句中仅有少部分内容是有效的驱动信息, 大部分为冗余信息。为了使推理过程更准确、清晰, Inoue 等人^[19]于 2020 年提出了半结构化、具有更细粒度的派生事实(Derivation)驱动的多文档推理数据集——R⁴C。从文档中抽取的派生事实 d 的形式化如式(3)所示。

$$d \equiv \langle d^h, d^r, d^t \rangle \quad (3)$$

其中, d^h 与 d^t 是名词性短语实体, d^r 是动词性短语, 用于表示 d^h 和 d^t 的逻辑关系。式(3)半结构化逻辑表示结构, 可以为模型提供更清晰明确的逻辑跳转事实, 细化、充实了推理链路的完整性、可靠性和合理性。

由于派生事实粒度较细且提取困难, R⁴C 数据集的规模小于 HotpotQA, 其包含约 4 600 个样本, 每个样本均包含 3 条人工提取的优质派生事实。尽管数据集规模较小, 但这种小粒度的推理链条, 模拟了人类阅读中细粒度可解释理解的过程, 是多文档阅读理解任务中研究的新趋势, 有助于模型在细节推理方向的创新。

2.2 多文档机器阅读理解评价指标

2.2.1 准确率

对于分类式阅读理解任务, 一般采用准确率(Accuracy)评价模型的预测结果^[6, 20-22], 准确率是预测结果正确的样本数占样本总数的比值。

2.2.2 联合 F_1 和 EM

对于逻辑链路预测要求的多文档阅读理解任

务, 一般需要采用联合 F_1 (Joint F_1) 指标以评价模型在预测过程中生成链路的质量^[11, 19]。计算时, 以词语为粒度分别计算正确答案、正确推理链路 with 预测答案、预测推理链路之间的答案查准率 P_{ans} 、推理链路查准率 P_{sup} 、答案查全率 R_{ans} 、推理链路查全率 R_{sup} , 并分别将答案与推理链路的查准率和查全率作积得到联合查准率与联合查全率。联合查准率 P_{joint} 、联合查全率 R_{joint} 和联合 F_1 的定义如式(4)~式(6)所示。

$$P_{joint} = P_{ans} P_{sup} \quad (4)$$

$$R_{joint} = R_{ans} R_{sup} \quad (5)$$

$$\text{联合 } F_1 = \frac{2P_{joint}R_{joint}}{P_{joint} + R_{joint}} \quad (6)$$

EM 精确匹配(Exact Match) 是一项较为严格的评价指标, 对于模型给出的候选答案词组、句子, 只有在候选答案与参考答案完全相同时才可视为匹配成功。若数据集中共有 N 个问题, 模型给出 M 个匹配成功的候选答案, 其 EM 得分计算如式(7)所示。

$$EM = \frac{M}{N} \quad (7)$$

2.2.3 BLEU

对于生成式阅读理解任务, 可使用 BLEU (Bilingual Evaluation Understudy)^[23]进行答案质量评价。该指标最初被应用在机器翻译任务中, 评价生成译文相对于标准译文的质量, 如式(8)~式(10)所示。

$$p_n = \frac{\sum_{C \in \text{Candidates}} \sum_{n\text{-gram} \in C} \text{Count}_{\text{clip}}(N\text{-gram})}{\sum_{C' \in \{\text{Candidates}\}} \sum_{n\text{-gram}' \in C'} \text{Count}(N\text{-gram}')} \quad (8)$$

$$BP = \begin{cases} 1, & c > r \\ e^{1-\frac{r}{c}}, & c \leq r \end{cases} \quad (9)$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N w_n \log P_n\right) \quad (10)$$

其中, N -gram 表示采用 N 元精度(N -gram precision), 一般选用四元精度(4-gram Precision)进行句子级别的评价; BP(Brevity Penalty)为简洁惩罚因子, 以惩罚候选语句在低召回率下语句过短的情况。

2.2.4 ROUGE

对于生成式阅读理解任务, 还可使用 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)^[24]进行答案质量评价。该指标起初主要用于评价自动摘要生成任务的生成语句^[25], 与 BLEU 不同的是其

可从查全率的角度对候选语句与参考语句进行评价。在生成式机器阅读理解任务的实际应用中,最常用的指标是 ROUGE-L,假定对于某模型有长度为 m 的参考序列文本答案 X 和长度为 n 的候选序列文本答案 Y , ROUGE-L 得分 F_{lcs} 计算过程如式(11)~式(13)所示。

$$R_{\text{lcs}} = \frac{\text{LCS}(X, Y)}{m} \quad (11)$$

$$P_{\text{lcs}} = \frac{\text{LCS}(X, Y)}{n} \quad (12)$$

$$F_{\text{lcs}} = \frac{(1 + \beta^2) R_{\text{lcs}} P_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 P_{\text{lcs}}} \quad (13)$$

其中, $\text{LCS}(X, Y)$ 为 X 与 Y 的最长公共子序列长度; R_{lcs} 为查全率; P_{lcs} 为查准率; β 为平衡因子, 在机

器阅读理解任务中一般可取值为 1.2; 而当 β 足够大时, 得分 $F_{\text{lcs}} = R_{\text{lcs}}$, 即指标关注查全率。

3 多文档机器阅读理解模型

基于深度学习的多文档机器阅读理解模型, 可按照模型的线性处理顺序, 抽象为如图 2 所示的自底向上的模型层次图, 包含映射自然语言至向量空间的嵌入层、捕捉线性文本序列特征的编码层、理解文本与实体间语义关系的阅读理解层和适应任务特性的答案输出层。通过研究模型在各层, 尤其是阅读理解层的技术方法, 实现研究者的设计思路以适应数据集任务的特点, 是提升模型在多文档机器阅读理解任务中表现的重要手段。

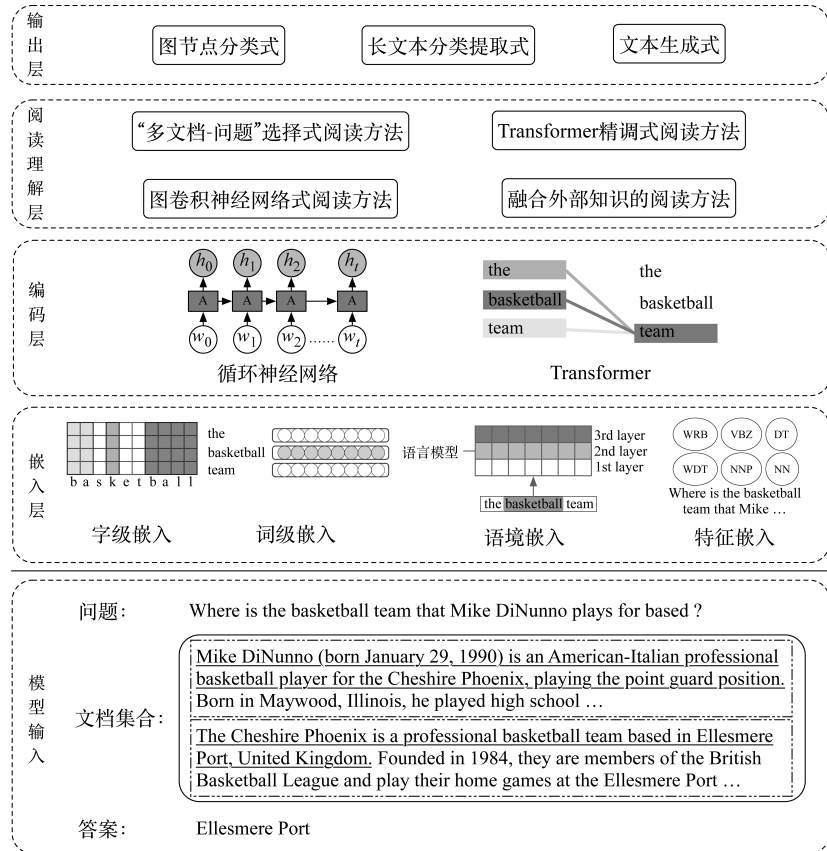


图 2 多文档机器阅读理解模型层次图

3.1 嵌入层

嵌入层 (Embedding Layer) 负责将问题、文档和答案等文字信息投射到高维稠密的向量空间中, 实现文本信息的数字化与向量化表示。依据嵌入过程中的粒度级别、语境依赖和词性等特征, 嵌入方法可以分为字级嵌入、词级嵌入、语境嵌入和特征嵌入等方法。在应用中可组合多种嵌入方法以提高模型

效果。

3.1.1 字级嵌入方法

如图 3 所示, 字级嵌入 (Character Embedding) 通常是以字母为单位并结合诸如卷积神经网络 (Convolutional Neural Networks)^[26] 等深度学习框架的文本嵌入方法。此方法可用来解决噪声字符 (Noisy Character)、未登录词 (Out-of-Vocabulary) 和生僻词 (Rare Word) 等问题^[27-33], 在 Tu 等人^[34]

提出的多文档阅读模型 HDE 中,将 N -gram 字级嵌入^[29]作为文本特征的一部分,以缓解因粗粒度嵌入方法导致的信息表示不足的问题。

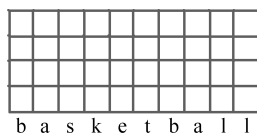


图3 字级嵌入示意图

3.1.2 词级嵌入方法

如图4所示,词级嵌入(Word Embedding)是将文本以单词为单位,基于单词分布式假设进行的实值向量化表示。2013年,Mikolov等人^[35-36]提出了基于连续词袋模型和连续跳跃元语法模型,用于单词相似性检验的词向量嵌入模型 Word2Vec,实现了对词语关系的建模。2014年,为了更好地利用大规模语料库的语言共现(Co-occurrence)统计特性,Pennington等人^[37]基于全局对数双线性回归方法提出了 GloVe 模型,结合了局部上下文窗口和全局矩阵分解两种方法的优点,并在多个多文档阅读理解模型的应用中提高了答案预测的准确率^[34,38-39]。



图4 词级嵌入示意图

3.1.3 语境嵌入方法

虽然在词级嵌入方法中固定的实值嵌入方法取得了很大的成功,但是研究者发现如图5所示结合了上下文的语境嵌入方法展现出更强大的语言表示能力。此类方法一方面可以区分一词多义,另一方面还可以通过利用预训练深度网络模型中的高低阶层来表示语境中蕴含的单词含义和句法结构信息^[40]。2018年,Peters等人^[40]基于双向 LSTM 提出了一种利用长短时记忆网络(Long Short-term Memory, LSTM)基于特征(Feature-Based)的语言模型 ELMo。该模型利用单词的上下文对文字进行特定语境的双向嵌入表示,有效提升了其应用到自然语言处理的其他下游任务的表现。2019年,Devlin等人^[4]在双向 Transformer 的基础上提出了利用预训练+精调(Fine-Tuned)策略的双向特征表示语言理解模型 BERT,在单文档阅读理解任务 SquAD 1.1 上超越了人类的作答结果。2020年,针对“预训练+精调”的两段式语言嵌入框架需要大量领域标注数据,有悖于人类在少量样本上就可学习

语言规律的问题,Brown等人^[41]提出了基于小样本(Few-Shot)的自回归语言模型(Autoregressive Language Model) GPT-3。得益于该模型先进的设计理念,在多文档阅读理解数据集 TriviaQA^[14]上,超越了当时效果最好的“预训练+精调”式模型 RAG^[42]的预测结果。

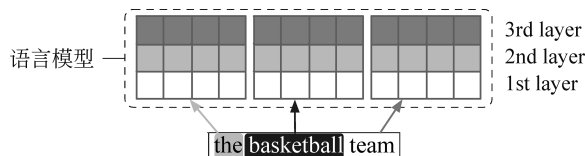


图5 语境嵌入示意图

3.1.4 特征嵌入方法

如图6所示,特征嵌入(Feature Embedding)可以将诸如命名实体识别(Named-Entity Recognition, NER)、词性标注(Part-of-Speech, POS)、位置特征(Position)等文本低维度特征与其他特征嵌入进行拼接,进而提升模型的阅读理解能力。Cao等人^[39]在 BAG 模型中,将文本的 NER、POS 特征和语义嵌入拼接形成联合嵌入后,提升了多文档阅读理解任务的答案预测准确率。

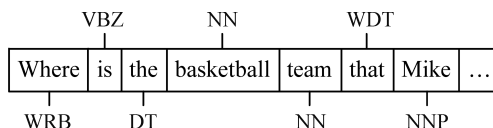


图6 特征嵌入示意图

3.2 编码层

在多文档阅读理解模型中,编码层(Encoder Layer)主要对线性文本序列的上下文语境进行编码,是模型中通用的基础部分。多文档机器阅读理解模型中常用的编码层结构有循环神经网络(Recurrent Neural Networks, RNN)及其变种和 Transformer 及其变种。

3.2.1 循环神经网络

如图7所示,RNN^[43]及其变种 LSTM^[44]、GRU^[45]可以通过时间步的前进方向顺序地处理文本。由于此类网络模型包含学习单元和参数共享机制,使其在应用中具有记忆功能。在实际应用中,除了有从前向后的时间步顺序,还可以结合从后向前的反向时间步顺序,拼接生成双向的(Bidirectional)特征表示^[31,46-50],从而更好地学习文本特征。

3.2.2 Transformer

虽然 RNN 模型展现出了强大的时序信息编码

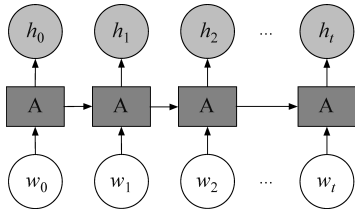


图 7 循环神经网络时间序列展开图

能力,但是其所依赖的时序学习模式不仅计算效率较低,还会导致长序列文本的前文信息出现丢失。2017 年,Vaswani 等人^[51]提出了不依赖时间步的完全基于自注意力机制的 Transformer 模型。如图 8 所示,该模型可通过并行计算模式提高计算效率,有效地解决了 RNN 模型中信息丢失的问题。

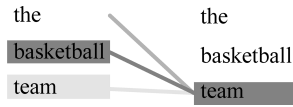


图 8 基于自注意力机制的 Transformer 编码示意图

Transformer 模型在编码和解码过程中,把分别由源输入序列和目标输出序列充当的线性文本序列划分为 3 个角色:查询 Q (Query)、键 K (Key)、值 V (Value),并通过计算序列间的缩放点积注意力 (Scaled Dot-product Attention) 进行特征学习,如式(14)所示。

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (14)$$

其中, $\frac{1}{\sqrt{d_k}}$ 为用于保证训练过程中梯度稳定的归一化缩放因子; $\text{Softmax}(\cdot)$ 为非线性激活函数,用于获取值的权重。得益于 Transformer 模型强大的语言特征捕捉能力,许多基于 Transformer 的多文档阅读理解模型效果达到了发布时的最高水平^[52-57],成为搭建编码层的有力备选。

3.3 阅读理解层

阅读理解层是整个模型中最灵活的自定义处理层之一,也是研究者们集思广益后提出的针对不同多文档阅读理解任务特点而进行特定处理的模块。根据模型的侧重点,阅读理解层包含 4 种主流研究方法:“多文档-问题”选择式、Transformer 精调式、图卷积神经网络式和外部知识融合式。

3.3.1 “多文档-问题”选择式阅读方法

鉴于多文档阅读理解任务需要对多个文档进行阅读和答案预测,研究者们提出了根据多个文档与

问题的关联性和重要性,有侧重地进行文档选择的方法。2018 年,Clark 等人^[32]认为在多个文档中预测答案,会被不相干的文档干扰,造成正确率下降,因此提出了利用 TF-IDF 算法先选择与问题相关的文档子集,后将子集内各文档预测的答案语段进行全局共享的归一化操作,避免了不同文档间各答案的特征差异而导致的输出偏置,迫使模型可以产出在不同文档间具有可比性的候选答案集合。

2018 年,Wang 等人^[46]提出的跨文档验证式阅读理解模型将多个文档拼接为一个超长序列,根据负采样对数概率在各文档的子序列部分各抽取一个候选答案语段 A_n (n 为文档及答案的序号),并在每个序列分词上设置是否与候选答案重叠的标记(与答案重合的分词标记为 1,其他位置标记为 0),以保证抽取到的答案语段相对于该文档的预测质量。最后利用 Softmax 函数对来自 n 个文档的候选答案语段联合表示 r^{A_n} 进行答案与多个段落的相关性验证,并输出最终预测的答案。

然而,将多个文档进行简单的拼接会因文档数量的增加而导致阅读效率下降。为解决这个问题,Yan 等人^[47]提出了面向多文档阅读的深度级联模型,通过“文档—段落—答案”不断细化的阅读过程,平衡了答案预测的正确率和阅读效率。多个文档在阅读过程中经历了文档排序、段落排序、文档提取、段落提取和答案提取 5 个处理过程,用于过滤与问题不相关的文档与段落,逐渐缩小答案搜索空间,保证了推理过程的合理性,在不断细化的级联过程中实现对答案语段的预测。

不同于上述两种模型从文档与答案的角度进行预测,在 Mao 等人^[55]从丰富问题语义度的角度提出的 GAR 模型,扩充了阅读理解中“问题”的语义信息,以提高答案检索时相关文档的数量与质量。Mao 提出了三种扩展问题语句的方式:(1) 将原始问题语句直接送入预训练语言模型中(如 GPT-3^[41]或 BART-large^[58])以生成“伪答案”语句,这些生成语句可能直接包含正确答案;(2) 给定文档中包含“伪答案”的原文语句;(3) 从维基百科中提取的,与原始问题语句相关且包含“伪答案”的语句。通过从文档中直接抽取答案和由模型接生成答案的两项实验,证明了以上三种扩增语句对提升答案预测准确率的有效性。

综上,“多文档-问题”选择式阅读方法是在单文档阅读理解任务的研究基础上,为满足多文档阅读理解任务要求而进行的必要扩展,为后续开展在多

文档信息逻辑推理奠定了坚实基础。

3.3.2 Transformer 精调式阅读方法

尽管 Transformer 模型具有较强的信息捕捉能力,然而考虑到模型复杂度,在应用中需要限制文本序列的长度。例如朴素 Transformer 模型支持的文本序列长度上限为 512,极大制约了 Transformer 模型在多文档或长序列文本上的应用。为了降低 Transformer 模型的复杂度,Beltagy 等人^[57]提出了稀疏自注意力模型 Longformer。该模型利用滑动窗口注意力、扩张滑动窗口注意力和全局滑动窗口注意力三种机制,将 Transofrmer 模型的复杂度从指数级降为线性级。同时,全局滑动窗口注意力机制的引入,能够获取问题文本的相对于整个文本序列的全局特征注意力,确保问题特征能被后续文本捕捉到。通过将多文档信息拼接为单个超长文本序列,Longformer 模型在发表时取得了 WikiHop 和 TriviaQA 数据集上的最佳表现。

受图的稀疏化启发,Zaheer 等人^[54]提出了 Big Bird 模型,引入了随机注意力、窗口注意力和全局注意力三种机制,复杂度仅为线性级。在 HotpotQA、TriviaQA 和 WikiHop 数据集上的实验结果表明该模型具有较强的长文本特征捕捉能力。

总之,基于 Transformer 的精调式阅读方法是将多个文档拼接成单个文档,利用或改进单文档阅读理解的方法完成答案预测,性能主要受文本序列长度的影响,而基于图神经网络的多文档阅读理解

方法则从特征提取的拓扑结构上实现创新。

3.3.3 图卷积神经网络式阅读方法

图卷积神经网络(Graph Convolutional Networks, GCNs)是一种有效表示真实世界中复杂关系的网络拓扑结构^[59],可对应在多文档阅读理解中的实体逻辑关系。

2019 年,Cao 等人^[60]认为当时的机器阅读理解方法主要应用于单个文档的信息检索,为了整合和推理来自多个文档的消息,提出了基于 GCN 的跨文档阅读理解模型——Entity-GCN。在阅读过程中将从文档中提取出的候选答案实体固化为图中的提及节点(Mention Nodes),利用“同段”(连接来自同一文档段落的节点)、“同指”(连接同一候选答案的不同节点)和“共指”(利用指代消解连接同一实体的不同节点)的加边方式建立为图 9 的拓扑结构。图中虚线矩形框代表单个文档;节点是文档中与候选答案匹配的文本语段(Text Span);来自同一个文档的所有节点用实线连接,而不同文档的相同文字节点用虚线连接。使用多关系型 GCN^[61]更新节点特征,如式(15)所示。

$$u_i^{(l+1)} = f_s(u_i^{(l)}) + \frac{1}{|N_i|} \sum_{j \in N_i} \sum_{r \in R_{ij}} f_r(u_j^{(l)}) \quad (15)$$

其中, u 为节点特征, l 为 GCN 的迭代层数, N_i 为节点 i 的邻居节点集合, R 为图中存在的关系集合, f_s 和 f_r 分别为学习自身节点特征和在某特定关系下邻居节点特征的线性学习矩阵。

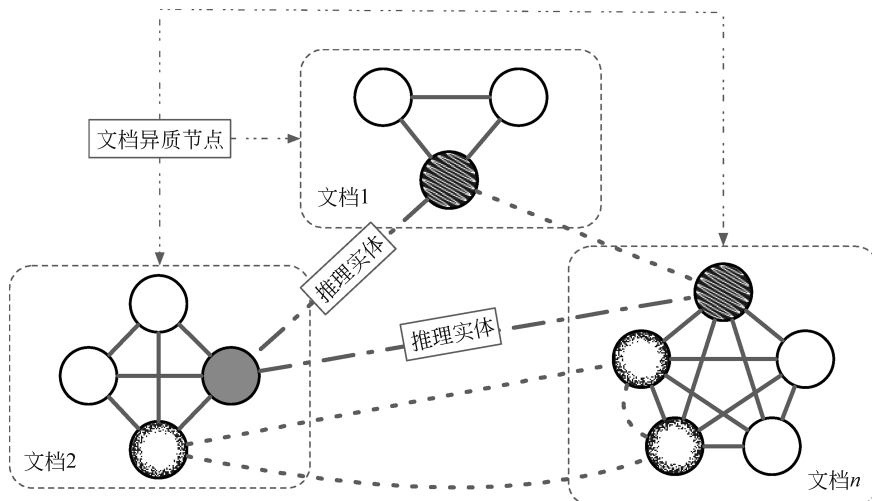


图 9 基于图卷积神经网络的多文档阅读理解方法建模图

2019 年,为了更好地融合问题与节点特征,Cao 等人^[39]将单文档阅读理解方法中的双向注意力机制^[28]引入到 Entity-GCN 的节点特征中,提出了

BAG 模型,实现了将双向注意力机制由序列文本到图结构的应用。问题与节点特征的相似度矩阵 S 的计算过程如式(16)所示。

$$\mathbf{S} = \text{avg}_{-1} f_a (\text{Concat}(\mathbf{h}_n, \mathbf{f}_q, \mathbf{h}_n \circ \mathbf{f}_q)) \quad (16)$$

其中, $\mathbf{h}_n$ 为 GCN 模块输出的节点特征, $\mathbf{f}_q$ 为编码的问题特征, $\circ$ 为矩阵元素点积 (Dot Product), Concat 表示矩阵合并, $\mathbf{f}_a$ 为线性学习矩阵, avg_{-1} 表示维度平均。利用 Softmax 函数, 可以分别得到节点对问题的感知注意力 $\tilde{a}_{n2q}$ 和问题对节点的感知注意力 $\tilde{a}_{q2n}$ 。最后生成对问题感知的节点特征表示 $\tilde{a}$, 如式(17)所示。

$$\tilde{a} = \text{Concat}(f_n, \tilde{a}_{n2q}, f_n \circ \tilde{a}_{n2q}, f_n \circ \tilde{a}_{q2n}) \quad (17)$$

其中, f_n 表示在进入 GCN 模块前的原始节点特征表示, $\tilde{a}$ 最终被送入输出层用于答案预测。

2020 年, Tang 等人^[38]从补全多跳阅读理解实体推理链条的角度提出了 Path-based GCN 模型, 其主要思想是除提取与候选答案一致的文本语段作为节点之外, 还抽取了与节点有关的命名实体和名词性短语作为推理节点补充到阅读图。图 9 中方角矩形的推理实体连接了两答案节点, 构建潜在的推理链路。此外, Tang 等人还提出了问题感知的门控机制 (Question-aware Gate), 不同于 BAG 模型中问题节点特征的引入方式, 问题感知门控作用于 GCN 的推理过程, 可控制节点的特征更新。

与节点均为实体的同质图不同, Tu 等人^[34]为了表示多语义角度, 将涉及的多个文档提取为与其他节点异质的文档节点挂载到图网络, 最终提出了包含不同信息粒度的异质图神经网络 HDE^[34]。如图 9 的圆角矩形所示, 将文档节点连接到其自身包含的实体节点, 得到异质图网络。

受认知科学启发, Ding 等人^[62]提出的 CogQA 模型将多文档阅读分为两个过程: 隐式信息提取和显式逻辑推理, 分别用 BERT 语言模型和图卷积神经网络 GCN 实现。在阅读时, 迭代地使用 BERT 从文档中提取实体作为潜在的跳跃实体和答案实体, 再利用 GCN 进行推理, 直到节点穷尽或 GCN 达到一定规模。CogQA 的实体提取过程如式(18)~(21)所示。

$$P_{\text{ans}}^{\text{start}}[i] = \frac{e^{\mathbf{S}_{\text{ans}} \cdot \mathbf{T}_i}}{\sum_j e^{\mathbf{S}_{\text{ans}} \cdot \mathbf{T}_j}} \quad (18)$$

$$P_{\text{ans}}^{\text{end}}[i] = \frac{e^{\mathbf{E}_{\text{ans}} \cdot \mathbf{T}_i}}{\sum_j e^{\mathbf{E}_{\text{ans}} \cdot \mathbf{T}_j}} \quad (19)$$

$$P_{\text{hop}}^{\text{start}}[i] = \frac{e^{\mathbf{S}_{\text{hop}} \cdot \mathbf{T}_i}}{\sum_j e^{\mathbf{S}_{\text{hop}} \cdot \mathbf{T}_j}} \quad (20)$$

$$P_{\text{hop}}^{\text{end}}[i] = \frac{e^{\mathbf{E}_{\text{hop}} \cdot \mathbf{T}_i}}{\sum_j e^{\mathbf{E}_{\text{hop}} \cdot \mathbf{T}_j}} \quad (21)$$

其中, $\mathbf{S}_{\text{ans}}$ 、 $\mathbf{E}_{\text{ans}}$ 、 $\mathbf{S}_{\text{hop}}$ 和 $\mathbf{E}_{\text{hop}}$ 为指针向量, 分别用于计算文档中第 i 个分词成为答案节点和跳跃节点的概率, $\mathbf{T}$ 表示分词的向量特征, $\mathbf{T}_j$ 表示与第 i 个分词位于同一语句的其他分词的特征。提取到的节点在基于 GCN 的逻辑推理模块中依照拓扑结构进行特征更新。

3.3.4 融合外部知识的阅读方法

2019 年, 为了提升阅读过程的推理能力, Ye 等人^[63]提出了基于知识增强的图阅读模型 KGNN。该模型从多个文档中提取节点, 利用知识图谱 (Knowledge Graph) 中现存的实体关系连接已知的且有知识关系的节点, 进而建立关系矩阵 $\mathbf{E}_r$ 。同时, KGNN 模型将共指的节点进行连接, 利用如式(22)所示特定关系的信息传播方式完成推理过程。

$$v_i^u = \sum_r \frac{\alpha_r}{|N_r(v_i)|} \sum_{v_j \in N_r(v_i)} \varphi_r(v_j) \quad (22)$$

其中, $N_r(v_i)$ 表示在 r 关系下节点 v_i 的邻居节点, α_r 是将问题和关系 r 匹配的注意力权重矩阵, 利用 $\varphi_r(v_j) = \text{FFN}(\mathbf{v}_j + \mathbf{E}_r)$ 的方式将邻居节点和特定连接关系进行融合, 这里 $\text{FFN}(\cdot)$ 为全联接的前向传播层。KGNN 模型从建模实体间关系的角度出发, 利用了知识图谱的先验知识指导节点的连接, 对提升模型推理链合理性和表现起到助力作用。

除利用知识外部知识指导关系建模外, Yang 等人^[64]提出了一种从知识图谱提取知识嵌入到阅读图的模型 KT-NET。对于文档和问题分词的全集 W , 该模型从知识库 C 中匹配相关的知识概念 $c \in C(w)$, 并通过相似度矩阵 α 匹配 w 与 c 的语义关系, 将与 w 相关的知识向量嵌入融合到 BERT 模型给出的语义向量中进行后续预测。KT-NET 模型的优点在于不仅可以整合知识图谱与文档语义的知识信息, 还可以从多个知识图谱全局的角度考虑补充文档中未显示给出的知识嵌入信息, 协调了语言模型和先验知识的应用关系。

为了在图神经网络中利用知识三元组信息, Sun 等人^[65]提出了包含文档、实体和知识三元组的异质图网络模型 PullNet。该模型通过迭代地从知识图谱中“拉取”三元组的方式, 在文档和知识图谱中找到与问题相关的实体并补充到阅读图中, 最终使图包含来自知识图谱中的实体节点集合 V_e 、表示单句话的文档节点集合 V_d 和包含知识三元组信息的事实节点集合 V_f 。令 E 表示图网络的边集合, 图中存在两种边关系: (1) 若 V_e 中的节点 v_s, v_o 包含于知识图谱三元组中, 则分别连接 v_s, v_o 和对应

的 v_f ; (2) 若实体节点 v_e 在文本序列 v_d 中被提及, 则连接 v_d 与 v_e 。在建立阅读图后, 基于节点分类的方式完成答案预测。PullNet 是整合多文档文本信息和知识图谱知识的模型, 其创新在于通过建立包含先验知识信息的异质推理图, 进行融合先验知识的阅读理解过程。

为了充分理解问题语义, 结合问题类型实现有针对性的答案预测, 谭红叶等人^[66]提出了一种基于外部知识和层级篇章表示的多文档阅读理解方法。该方法从利用外部知识加强对问题的理解程度出发, 先定义了识别问题语句中重要词的多种规则, 再引入了重要词在《现代汉语词典》(第六版) 的字典释义和 HowNet 义原, 使用特征嵌入拼接的方式, 得到融合字典释义的问题特征表示。实验表明, 《现代汉语词典》对重要词的覆盖率为 88.1%, HowNet 对重要词的覆盖率为 92.7%, 因此能较好地适应问题中复杂的语义关系, 实现丰富问题语义知识表示的目的, 进一步提高了模型通过问题特征捕捉原文信息进行阅读理解的能力。

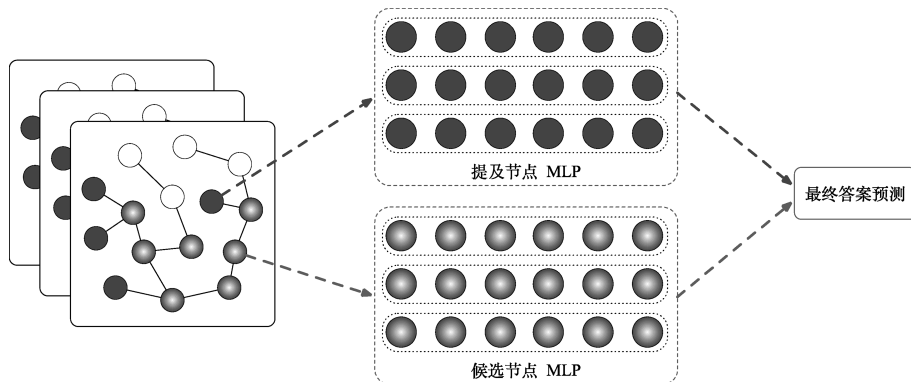


图 10 图节点分类式输出层示意图

其中, H^C 、 H^E 分别表示阅读图中候选节点和对应的实体提及节点特征, 在两个独立的双层感知机 f_C 、 f_E 的作用下, 得到每个节点的预测得分, $ACC_{\max}(\cdot)$ 表示从一组数据中提取的最大值。将两者按照类别相加, 得到各类的分布得分。

3.4.2 长文本分类提取式

在序列式文本中, 模型会预测每个分词作为答案文本语段的起始概率 p_{start} 和结束概率 p_{end} , 并将两者的最大概率对应的文本语段作为输出, 形式与 CogQA 模型提取跳跃实体文本语段或候选答案语段的方法类似, 不再赘述。

对于给定多个候选答案的分类式阅读理解任务, 一些长文本阅读模型采取拼接多个候选项和

3.4 输出层

输出层的主要目的是适应阅读理解层的输出与数据集的任务形式, 生成模型的最终答案预测。根据阅读理解层的特点与多文档阅读理解的任务特性, 常用的输出层有三种类型: 图节点分类式、长文本分类提取式和文本生成式。

3.4.1 图节点分类式

利用 GCN 对实体和文档进行关系建模的方法^[34,38-39,60]在阅读理解层后连接多层感知机 (Multilayer Perceptron, MLP) 作为输出层, 在正确答案与候选答案间使用交叉熵函数计算损失, 进而优化模型参数。如图 10 所示, HDE 模型^[34]从 GCN 提取提及节点 (Mention Nodes) 和候选节点 (Candidate Nodes) 的特征, 并将它们输入到两个独立的 MLP 计算每个类别的分布得分 a , 如式 (23) 所示。

$$a = f_C(H^C) + ACC_{\max}(f_E(H^E)) \quad (23)$$

增加特殊标记的方式, 计算分类预测结果的得分。如 Longformer 模型^[57]将针对问题给出的多个候选答案 (candidates)、问题 (question) 与多个文档 (context) 拼接为单个长序列文本: $[q]$ question $[/q]$ $[ent]$ candidate1 $[/ent]$ $\dots$ $[ent]$ candidateN $[/ent]$ $</s>$ context1 $</s>$ $\dots$ $</s>$ contextM $</s>$, 这里 $[q]$ 、 $[/q]$ 、 $[ent]$ 、 $[/ent]$ 和 $</s>$ 均为用于微调模型的标记。在微调的过程中, 标记被赋予特殊含义, 预测时将每个实体前的 $[ent]$ 标签输入到输出层, 得到该实体作为答案预测的概率。

3.4.3 文本生成式

由于在多文档机器阅读理解任务中作答时需整

合多篇文档的语义信息,直接从文档中抽取文本语段作为答案可能会导致答案片面狭隘,无法统筹所有文档的语义信息。为解决该问题,部分模型使用基于序列到序列(Sequence-to-Sequence)的生成式输出层对文档与答案进行编解码,获取符合多个文档语义信息的答案。Masque 模型^[52]遵循 Transformer 的编解码模式,在输出答案最终分布概率前,考虑了两种不同来源的答案分布概率,即扩展词汇分布(Extended Vocabulary Distribution)和拷贝分布(Copy Distributions),前者负责计算从词典 V_{ext} 中输出词汇 y_t 的概率 $P^v(y_t)$,后者负责协调从多个文档和问题原文中拷贝词汇 y_t 作为答案的概率 $P^p(y_t)$ 、 $P^q(y_t)$ 。将词汇 y_t 的三个概率加权求和,得到如式(24)所示的 y_t 输出概率分布,如式(24)所示。

$$P(y_t) = \lambda^v P^v(y_t) + \lambda^q P^q(y_t) + \lambda^p P^p(y_t)$$

(24)

其中, λ^v 、 λ^q 、 λ^p 为学习得到用于控制三者分布的平衡参数,词典 V_{ext} 中每个词汇 y_t 的输出概率由标准答案、问题文本和多文档文本综合调控,充分地利用了多文档机器阅读理解具有多个源文档的任务特性,优势较为显著。

表 3 展示了本文涉及模型按照数据集分组后,输出层预测结果的横向比较。其中,加粗的字体表示同组模型中更优的表现,人工作答表现用斜体标记。此外,为公平比较模型的预测结果,在表 3 中均选取单个模型的预测结果进行比较。在 HotpotQA 数据集评价排名中,“D”“F”分别表示 Distractor Setting 和 Fullwiki Setting 两种不同的评价环境。

表 3 数据集中代表模型的横向比较结果

MS MARCO	指标	
模型	BLEU-1	ROUGE-L
Wang 等人 ^[46]	44.47	46.15
Nishida 等人 ^[52]	42.37	50.93
TriviaQA	指标	
模型	EM	F_1
BiDAF ^[14]	40.74	47.05
Clark 等人 ^[32]	61.10	66.04
Yan 等人 ^[47]	68.65	73.07
Zaheer 等人 ^[54]	—	79.50
Mao 等人 ^[55]	62.70	—
Beltagy 等人 ^[57]	—	75.20

续表		
DuReader	指标	
模型	BLEU-4	ROUGE-L
BiDAF ^[9]	31.80	39.00
人工表现 ^[9]	56.10	57.40
Wang 等人 ^[46]	40.97	44.18
Yan 等人 ^[47]	49.39	50.71
Tan 等人 ^[66]	39.70	45.3
HotpotQA	指标	
模型	EM	联合 F_1
Ding 等人 ^[62]	F: 12.40	F: 34.90
Beltagy 等人 ^[57]	—	D: 64.40
Zaheer 等人 ^[54]	—	D: 67.80
Ye 等人 ^[63]	D: 22.40 F: 7.03	D: 52.82 F: 24.66
WikiHop	指标	
模型	准确率	
BiDAF ^[8]	42.90	
Tu 等人 ^[34]	70.90	
Tang 等人 ^[38]	72.50	
Cao 等人 ^[39]	69.00	
Zaheer 等人 ^[54]	75.90	
Beltagy 等人 ^[57]	75.00	
Cao 等人 ^[60]	67.60	

4 总结与展望

本文从机器阅读理解发展的角度,梳理了多文档阅读理解作为机器阅读新兴研究方向的发展历程;辨析了多文档阅读理解与单文档阅读理解的区别与联系,确立该任务的研究意义;介绍了 8 个大型多文档阅读理解数据集和 5 个通用的评价标准;调研了当前多文档阅读理解的主流研究方法,按照处理顺序把模型划分为 4 个层次;在层次一致的前提下,介绍并比较了各方法的独到之处。

总体上,多文档阅读理解是在单文档阅读理解的研究基础上发展而来的新任务,它加深了机器从文本中显式地提取答案,到隐式地学习推理多文本间潜在语义逻辑关系的研究深度,是推广机器阅读理解技术在人类世界应用的必经研究阶段。然而,多文档机器阅读理解任务还存在以下三方面值得进

一步研究。

多文档阅读过程的推理能力要求 目前多文档阅读理解数据集对问题的回答提出了两个要求：①多源信息的筛选拟合；②多文档的语义逻辑推理。尽管诸如将多文档拼接为单文档的阅读理解方法能够对信息进行一定的筛选，但是此类改进的单文档阅读理解方法在多文档间进行逻辑推理和可解释的能力较弱。同时，针对长文本序列的阅读方法虽然在一定程度上缓解了因文本规模增大而导致的信息丢失和存储开销增加等问题，但是面对日益增长的信息量，长文本序列的建模方法显得捉襟见肘，伸缩能力的缺乏仍有可能导致其预测效率降低。因此，从提升模型推理能力的角度展开研究，不仅可以解释阅读的推理决策过程，还可以避免从全文盲目寻找答案的过程。基于级联推理和 GCN 的方法，借助文本序列的空间拓扑结构，为模型赋予伸缩能力与逻辑推理能力。

多文档推理过程的细粒度化 与人类阅读过程相似，多文档阅读理解模型需要在推理过程中筛选有意义的信息，逐渐缩小答案搜索空间，进而输出准确的预测。HotpotQA 数据集提供了语句级的支持性语句，帮助模型学习语句的选择过程，而人类阅读时不总是在整段文本上进行推理，常常以更细粒度的形式选择有用信息，比如利用命名实体在多个相关文档间进行无监督的逻辑推理跳转^[7]，构建推理链路。借助 R⁴C 数据集给出的比语句更细化、表述更清晰的三元组推理依据，模型可进行有监督的实体逻辑推理学习过程。此外，借助模型在细粒度数据上的学习结果，可以尝试分析其在其他数据集上的泛化能力，进一步解释模型在预测过程中的选择决策依据，提升模型的综合鲁棒性。

与外部知识结合的阅读理解 与阅读理解任务相比，基于知识库的知识问答任务也具有相当长的研究历史，且取得了一批有价值的研究成果。由于文档受篇幅或叙述角度的局限，模型阅读时可用语料信息不足或将导致推理过程不合理或失效。当前已有模型借助外部知识库建模文本的实体关系，或将知识三元组直接引入到阅读过程中，取得了一定的成功。因此，利用结构化的外部知识提升无结构文本阅读理解性能的方法值得进一步研究。

参考文献

[1] ADAMS M, BRUCE B. Background knowledge and

reading comprehension [J]. Reader Meets Author: Bridging the Gap, 1982, 13: 2-25.

[2] CHEN D. Neural reading comprehension and beyond [D]. PhD Diss., Stanford University, 2018.

[3] 苏立新, 郭嘉丰, 范意兴, 等. 面向多片段答案的抽取式阅读理解模型[J]. 计算机学报, 2020, 43(05): 856-867.

[4] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.

[5] RAJPURKAR P, JIA R, LIANG P. Know what you don't know: Unanswerable questions for SQuAD [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 784-789.

[6] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100,000+ Questions for machine comprehension of text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016: 2383-2392.

[7] SUN K, YU D, YU D, et al. Investigating prior knowledge for challenging Chinese machine reading comprehension[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 141-155.

[8] WELBL J, STENETORP P, RIEDEL S. Constructing datasets for multi-hop reading comprehension across documents[J]. Transactions of the Association for Computational Linguistics, 2018, 6: 287-302.

[9] HE W, LIU K, LIU J, et al. DuReader: A Chinese machine reading comprehension dataset from real-world applications[C]//Proceedings of the Workshop on Machine Reading for Question Answering @ ACL, 2018: 37-46.

[10] NGUYEN T, ROSENBERG M, SONG X, et al. MS MARCO: A human generated machine Reading comprehension dataset[C]//Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches Co-located with the 30th Annual Conference on Neural Information Processing Systems, 2016.

[11] YANG Z, QI P, ZHANG S, et al. HotpotQA: A dataset for diverse, explainable multi-hop question answering [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 2369-2380.

- [12] 王军平, 张文生, 王勇飞, 等. 面向大数据领域的事理认知图谱构建与推断分析[J]. 中国科学(信息科学), 2020, 50(07): 988-1002.
- [13] BARADARAN R, GHIASI R, AMIRKHANI H. A survey on machine reading comprehension systems [J]. *Natural Language Engineering*, 2022, 28(6): 683-732.
- [14] JOSHI M, CHOI E, WELD D, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1601-1611.
- [15] COHEN K B, HUNTER L. Natural language processing and systems biology[M]. *Artificial Intelligence Methods and Tools for Systems Biology*. Dordrecht: Springer Netherlands, 2004: 147-173.
- [16] LAW V, KNOX C, DJOUMBOU Y, et al. DrugBank 4.0: Shedding new light on drug metabolism [J]. *Nucleic Acids Research*, 2014, 42: 1091-1097.
- [17] WISHART D S, FEUNANG Y D, GUO A C, et al. DrugBank 5.0: A major update to the DrugBank database for 2018[J]. *Nucleic Acids Res*, 2018, 46(D1): D1074-D1082.
- [18] GUPTA M, KULKARNI N, CHANDA R, et al. AmazonQA: A review-based question answering task [C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019: 4996-5002.
- [19] INOUE N, STENETORP P, INUI K. R4C: A benchmark for evaluating rc systems to get the right answer for the right reason[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 6740-6750.
- [20] ONISHI T, WANG H, BANSAL M, et al. Who did what: A large-scale person-centered cloze dataset [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016: 2230-2235.
- [21] BERZAK Y, MALMAUD J, LEVY R. STARC: structured annotations for reading comprehension [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 5726-5735.
- [22] KHASHABI D, CHATURVEDI S, ROTH M, et al. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 252-262.
- [23] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.
- [24] LIN C Y. ROUGE: A Package for Automatic Evaluation of Summaries[G]//Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics, 2004: 74-81.
- [25] 李金鹏, 张闯, 陈小军, 等. 自动文本摘要研究综述[J]. 计算机研究与发展, 2021, 58(1): 1-21.
- [26] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014: 1746-1751.
- [27] DHINGRA B, ZHOU Z, FITZPATRICK D, et al. Tweet2Vec: Character-based distributed representations for social media[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 269-274.
- [28] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension [C]//Proceedings of the 5th International Conference on Learning Representations, 2017.
- [29] HASHIMOTO K, XIONG C, TSURUOKA Y, et al. A joint many-task model: Growing a neural network for multiple NLP tasks[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 1923-1933.
- [30] ARORA M, KANSAL V. Character level embedding with deep convolutional neural network for text normalization of unstructured data for twitter sentiment analysis[J]. *Social Network Analysis and Mining*, 2019, 9(1): 1211-1214.
- [31] NISHIDA K, NISHIDA K, NAGATA M, et al. Answering while summarizing: Multi-task learning for multi-hop QA with evidence extraction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2335-2345.
- [32] CLARK C, GARDNER M. Simple and effective multi-paragraph reading comprehension [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 845-855.
- [33] 顾迎捷, 桂小林, 李德福, 等. 基于神经网络的机器阅读理解综述[J]. 软件学报, 2020, 31(07): 2095-2126.
- [34] TU M, WANG G, HUANG J, et al. Multi-hop

- reading comprehension across multiple documents by reasoning over heterogeneous graphs[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2704-2713.
- [35] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]//Proceedings of the 1st International Conference on Learning Representations, 2013.
- [36] LE Q V, MIKOLOV T. Distributed representations of sentences and documents[C]//Proceedings of the 31th International Conference on Machine Learning, 2014: 1188-1196.
- [37] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2014: 1532-1543.
- [38] TANG Z, SHEN Y, MA X, et al. Multi-hop reading comprehension across documents with path-based graph convolutional network[C]//Proceedings of the 29th International Joint Conference on Artificial Intelligence, 2020: 3905-3911.
- [39] CAO Y, FANG M, TAO D. BAG: Bi-directional attention entity graph convolutional network for multi-hop reasoning question answering[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 357-362.
- [40] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 2227-2237.
- [41] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the Annual Conference on Neural Information Processing Systems, 2020.
- [42] LEWIS P S H, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Proceedings of the Annual Conference on Neural Information Processing Systems, 2020.
- [43] WILLIAMS R J, ZIPSER D. A learning algorithm for continually running fully recurrent neural networks[J]. Neural Computation, 1989, 1(2): 270-280.
- [44] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [45] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J/OL]. arXiv preprint arXiv: 1412.3555, 2014.
- [46] WANG Y, LIU K, LIU J, et al. Multi-passage machine reading comprehension with cross-passage answer verification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1918-1927.
- [47] YAN M, XIA J, WU C, et al. A deep cascade model for multi-document reading comprehension[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019: 7354-7361.
- [48] KUNDU S, NG H T. A question-focused multi-factor attention network for question answering[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018: 5828-5835.
- [49] WEISSENBORN D, KOČISKÝ T, Dyer C. Dynamic integration of background knowledge in neural NLU systems[J/OL]. arXiv preprint arXiv: 1706.02596, 2017.
- [50] JIANG Y, JOSHI N, CHEN Y C, et al. Explore, propose, and assemble: An interpretable model for multi-hop reading comprehension[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2714-2725.
- [51] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the Annual Conference on Neural Information Processing Systems, 2017: 5998-6008.
- [52] NISHIDA K, SAITO I, NISHIDA K, et al. Multi-style generative reading comprehension[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2273-2284.
- [53] AINSLIE J, ONTAÑÓN S, ALBERTI C, et al. ETC: Encoding long and structured inputs in transformers[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 268-284.
- [54] ZAHEER M, GURUGANESH G, DUBEY K A, et al. Big Bird: Transformers for longer sequences[C]//Proceedings of the Annual Conference on Neural Information Processing Systems, 2020.
- [55] MAO Y, HE P, LIU X, et al. Generation-Augmented Retrieval for Open-Domain Question Answering[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the

- 11th International Joint Conference on Natural Language Processing, 2021: 4089-4100.
- [56] IZACARD G, GRAVE E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021: 874-880.
- [57] BELTAGY I, PETERS M E, COHAN A. Longformer: the long-document transformer[J/OL]. arXiv preprint arXiv:2004.05150, 2020.
- [58] LEWIS M, LIU Y, GOYAL N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 7871-7880.
- [59] LIU Z, ZHOU J. Introduction to graph neural networks[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2020, 14(2): 1-127.
- [60] CAO N D, AZIZ W, TITOV I. Question answering by reasoning across documents with graph convolutional networks[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 2306-2317.
- [61] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]//Proceedings of the European Semantic Web Conference, 2018: 593-607.
- [62] DING M, ZHOU C, CHEN Q, et al. Cognitive graph for multi-hop reading comprehension at scale[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2694-2703.
- [63] YE D, LIN Y, LIU Z, et al. Multi-paragraph reasoning with knowledge-enhanced graph neural network[J/OL]. arXiv preprint arXiv:1911.02170, 2019.
- [64] YANG A, WANG Q, LIU J, et al. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2346-2357.
- [65] SUN H, BEDRAX WEISS T, COHEN W W. PullNet: Open domain question answering with iterative retrieval on knowledge bases and text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 2380-2390.
- [66] 谭红叶, 李宣影, 刘蓓. 基于外部知识和层级篇章表示的阅读理解方法[J]. 中文信息学报, 2020, 34(04): 85-91.



高峰(1995—), 博士研究生, 主要研究领域为机器阅读理解和知识库问答。

E-mail: isfengg@outlook.com



高鹏(1991—), 博士, 教授, 主要研究领域为模式识别、计算机视觉和机器学习。

E-mail: pgao@qfnu.edu.cn



倪建成(1971—), 通信作者, 博士, 教授, 主要研究领域为计算机视觉、机器学习和分布式计算。

E-mail: nijch@163.com