神经机器阅读模型综述

骆 丹[1,2], 张 鹏[1,2], 马 路[1,2], 王 斌[3], 王丽宏[4]

[1] 中国科学院 信息工程研究所 第二研究室 北京 中国 100093
[2] 中国科学院大学 网络空间安全学院 北京 中国 100049
[3] 小米 AI 实验室 北京 中国 100085
[4] 国家计算机网络应急技术处理协调中心 北京 中国 100029

摘要 近年来, 随着互联网的高速发展, 网络内容安全问题日益突出, 是网络治理的核心任务之一。文本内容是网络内容安全最为关键的研究对象, 然而自然语言本身固有的模糊性和灵活性给网络舆情监控和网络内容治理带来了很大的困难。因此, 如何准确地理解文本内容, 是网络内容治理的关键问题。目前, 文本内容理解的核心支撑技术是基于自然语言处理的方法。机器阅读理解作为自然语言处理领域中的一项综合性任务, 可以深层次地分析、全面地理解网络内容, 在网络舆论监测和网络内容治理上发挥着重要作用。近年来, 深度学习技术已在图像识别、文本分类、自然语言处理等多个领域中取得显著成果, 基于深度学习的机器阅读理解方法也被广泛研究。特别是近年来各种大规模数据集的公开, 加快了神经机器阅读理解的发展, 各种结合不同神经网络的机器阅读模型被相继提出。本文旨在对神经机器阅读模型进行综述。首先介绍机器阅读理解的发展历史和研究现状; 然后阐述机器阅读理解的任务定义, 并列举出有代表性的数据集以及神经机器阅读模型; 再介绍四种新趋势目前的研究进展; 最后提出神经机器阅读模型当前存在的问题, 并且分析机器阅读理解如何应用于网络内容治理问题以及对未来的发展趋势进行展望。

关键词 网络内容安全; 网络舆情监测; 机器阅读理解; 自然语言处理; 深度学习; 神经网络
中图法分类号 TP181 DOI 号 10.19363/J.cnki.cn10-1380/tn.2024.03.10

A Survey on Neural Machine Reading Comprehension Model

LUO Dan[1, 2], ZHANG Peng[1, 2], MA Lu[1, 2], WANG Bin[3], WANG Lihong[4]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Xiaomi AI Lab, Beijing 100085, China
[4]National Computer Network Emergency Response Technical Team Coordination Center of China, Beijing 100029, China

Abstract After witnessing the soaring development of the Internet in the past decades, the problem of Cyber Content Security, which is considered as one of the core tasks of network governance, has become increasingly prominent. Text content is the most pivotal research object of cyber content security. However, the inherent ambiguous and flexibility of natural language bring great difficulties to public opinion monitoring and cyber content governance on the Internet. Therefore, how to accurately understand the text content is the key issue of cyber content governance. At present, the core supporting technology of text content understanding is based on Natural Language Processing. As a comprehensive task in the field of Natural Language Processing, Machine Reading Comprehension can analyze the network content in depth and achieve a comprehensive understanding, which plays an important role in the monitoring of network public opinion and the governance of cyber content. In recent years, Deep Learning technology has made remarkable achievements in many fields, such as Pattern Recognition, text classification and Natural Language Processing. Likewise, Machine Reading Comprehension methods based on Deep Learning have been widely studied. Especially in recent years, the publication of various large-scale datasets has accelerated the development of neural Machine Reading Comprehension, and various machine reading models combining different neural networks have been proposed successively. The purpose of this paper is to review various neural machine reading models. Firstly, the development history and research status of Machine Reading Comprehension are introduced. Then, the task definition of Machine Reading Comprehension is expounded, and representative datasets and neural machine reading models are presented. The latest research progress of four new trends is introduced. Finally, the existing problems of the neural machine reading model are put forward, how Machine Reading Com-

通信作者: 张鹏, 博士, 副研究员, Email: pengzhang@iie.ac.cn。
本课题得到国家重点研究发展计划 (No. 2016QY03D0503, No. 2016YFB081304)、中国科学院战略性先导项目 (No. XDC02040400)、中国科学院青年创新促进会项目(No. 2020163)资助。
收稿日期: 2020-05-21; 修改日期: 2020-09-08; 定稿日期: 2022-12-07

prehension methods are applied to solve the problem of Cyber content governance is analyzed, and the future development trend is forecasted.

Key words cyber content security; public opinion monitoring; machine reading comprehension; natural language processing; deep learning; neural network

1 引言

网络内容安全是国家安全的重要领域, 关系着国家政权安全和社会大局稳定。随着互联网应用技术的推广普及, 社交网络和自媒体等内容平台强化了虚拟世界和现实社会的耦合。在网络繁荣发展的背后也滋生着大量的违法违规内容, 如何从海量的网络数据中高效地过滤色情、广告、涉政、暴恐等违法信息以及敏感词, 是解决网络内容安全的关键。以前的人工审核方式成本高且效率低, 如今随着人工智能(Artificial Intelligence, AI)的不断发展及其在多个领域的广泛应用, 网络内容治理和网络舆论监控也正在迎接智能时代的到来。

传统的基于人工建模和特征提取的网络内容治理方法仅仅是通过关键词匹配来识别出违法违规和不良有害信息。语言表达具有模糊性和灵活性, 传统方法忽略了上下文语义, 从而导致了识别准确率低且误报率高。比如, 中文分词不明确出现误判现象, 谐音表达刻意回避敏感字眼导致漏报, 多样的表达增加了人工审核的工作量和工作难度。基于深度学习的机器阅读理解(Machine Reading Comprehension, MRC)可以结合上下文理解文本深层的语义, 提高识别安全威胁的准确度及反应速度, 从而提高网络舆情监控和网络安全治理的质量和效率。因此, 机器阅读理解技术的发展将会给予网络内容安全有力的保障。

机器阅读理解是自然语言处理(Natural Language Processing, NLP)领域中的核心任务之一, 让计算机能够处理文本, 并且理解文本的内在含义, 这是实现机器智能化、完成人机交互的关键基础。机器阅读理解是让智能体理解全文语境并且回答与文本相关的问题, 以此衡量机器对自然语言的理解能力。MRC 注重于更大范围、更深层次的上下文语义信息, 需要综合运用文本表示、语义理解、知识推理等多种复杂技术, 是一项充满魅力和挑战的任务。

本文的主旨是对神经机器阅读模型进行综述。早期的MRC系统使用人工生成且规模较小的数据集, 并且使用基于词袋表示、人工规则、信息抽取等简单的启发式方法, 因此模型性能低下, 不能在实际应用中使用。随着深度学习技术的飞速发展, 基于深度学习的 MRC(又称为神经机器阅读理解, neural machine reading comprehension)在挖掘上下文语义信息方面显示出其优越性, 比传统的基于规则的方法效果提升显著。另一方面, 随着各种大规模阅读理解数据集的公开发布(例如 CNN&Daily mail[1]、SQuAD[2-3]、MS MARCO[4]、DuReader[5]等), 神经 MRC 技术也取得了突破性地进展。

神经机器阅读理解正处于快速发展的阶段, 并且已经成为了学术界(Stanford, Carnegie Mellon 等)和工业界(Google, Facebook, Microsoft 等)的研究重点。如表 1 所示, 经过对国内外各界的研究现状进行大量调研之后, 本文将目前机器阅读理解的研究方向分为四种任务类型和四个新研究趋势。四大类任务中前三种研究比较成熟, 自由作答由于存在数据集难以构建、评价指标难以统一等局限性, 发展比较缓慢; 四个新趋势得到了研究人员的重视, 相关的研究成果也相继发布。

表 1 机器阅读理解研究方向介绍

Table 1 Introduction of machine reading comprehension research direction

四大类任务	完形填空	答案为篇章中的某个实体或者单词
	多项选择	从多个候选答案中选择唯一正确的选项
	跨度抽取	答案是来自篇章中的文本片段
	自由作答	答案是任何自由文本
四种新趋势	不可回答问题	判断出问题是否可作答, 若能, 就给出答案; 不能, 就拒绝回答
	多篇章 MRC	从多篇文档中推理预测出问题的答案
	基于知识的 MRC	结合外部知识进行阅读理解作答
	对话式 MRC	基于对话历史进行阅读理解作答

目前, 机器阅读理解相关的研究工作很多, 也有一些综述性论文发表。例如, 文献[6]侧重于对深度机器阅读理解模型基本结构的组件进行综述; 文献[7]是综述基于预训练模型的机器阅读理解研究进展。本文旨在对各种神经机器阅读模型进行研究综述。文章的结构如下: 第 2 节介绍 MRC 的发展历史和研究现状; 第 3 节介绍四类 MRC 任务定义和对应的数据集; 第 4 节介绍四种新研究趋势; 第 5 节从注意力机制和推理策略创新方面介绍神经机器阅读模型; 第 6 节分析了 MRC 目前存在的问题并且对未来进行展望; 第 7 节是总结。

2 机器阅读理解的研究现状

机器阅读理解大致经历了三个发展阶段: 系统基于规则的早期开发阶段、尝试建立机器学习模型来解决 MRC 任务的中期过渡阶段、当前基于深度学习的飞速发展阶段。

第一阶段, 基于规则的早期开发阶段。1977 年, Lehnert[8]首次提出了机器阅读理解的概念, 并且设计了 QUALM 系统, 重点关注于语用问题和故事上下文。Hirschman 等[9]在 1999 年发布首个机器阅读理解数据集, 该数据集一共包括了 120 个 3～6 年级的材料故事, 其中开发集和测试集各占一半。Hirschman 在该文献中还提出了 DEEP READ 系统[9], 该系统使用基于规则的词袋表示方法, 并且利用了词干提取、语义类识别和代词解析等技术。Riloff 等[10]基于此数据集设计出 QUARC 系统, 该系统依旧使用手动生成的规则。总的来说, 这一阶段的数据集规模很小, 没有强大的文本表示方法, 使用人工设计的启发式规则进行模式匹配, 模型拟合数据的能力低下, 因此无法在实际应用中使用。

第二阶段, 基于机器学习的中期过渡阶段。2013 年, 微软公司的 Richardson 等人[11]发布了一个高质量的单选题数据集 MCTest, 该数据集包含 500 个虚构的故事和 2000 个问题, 是以三元组(篇章、问题、答案)的形式收集的人类标记数据。虽然 MCTest 的出现促进了 MRC 的研究[12-14], 但是它的主要缺点是尺寸太小, 因为标记的样本数据通常需要相当专业的知识和整洁的设计, 这使得注释过程的代价非常昂贵, 所以 MCTest 不适用于一些需要大量标记数据的技术, 无法训练出性能良好的模型。这一阶段的机器学习模型相比早期基于规则的启发式方法, 使用了手工设计的特征, 如句法依赖关系、共指消解、篇章关系、单词嵌入等, 模型性能得到了一定程度的提升。然而非神经 MRC 模型具有一定的缺陷, 比如简单的文本表示、难以设计高质量的特征、数据量太小等, 模型仍然没能达到适合实际场景的性能要求。

第三阶段, 基于深度学习的飞速发展阶段。MRC 发展的转折点在于 2015 年谷歌 DeepMind 的 Hermann[1]提出了一个创新性的想法, 利用美国有线电视新闻网(CNN)[15]和每日邮报(Daily Mail)[16]的大量新闻报道的文章, 使用启发性的方法构造了大规模的有监督训练数据集 CNN&Daily Mail。通过将注意力机制整合到递归神经网络(recurrent neural network, RNN)结构中, Hermann[1]首次提出了两种神经机器阅读模型 Attentive Reader 和 Impatient Reader, 在识别词汇表示和语义匹配上性能表现远超传统基于特征的分类器。随后出现的 Stanford Attentive Reader[17]、AS Reader[18]都是基于 Attentive reader 进行简单改进的模型。2016 年, 哈工大讯飞联合实验室采用类似 Hermann[1]的方法构建了中文填空式 MRC 数据集: HFL-RC 数据集[19], 并且提出了 CAS Reader[19]模型。在此工作基础上, 又于 2017 年提出了 AoA Reader[20]模型, 该模型提高了特征提取的能力。

针对 CNN&Daily Mail 数据集, Chen 等人[17]对其进行了手工采样分析, 表明当时的神经网络已经可以达到该数据集的性能上限。斯坦福大学的 Rajpurkar 等人[2]在 2016 年发布了 SQuAD 数据集, 不同于填空式 MRC 数据集, 该数据集通过众包的方式收集了大量真实的查询和答案, 要求答案是原文的一个片段而不是单一的实体对象。斯坦福大学发起的 SQuAD 挑战赛, 吸引了学术界和工业界的广泛关注, 就好比是武林大会的召开, 各种英雄豪杰在此平台进行切磋较量, 使得 MRC 的发展得到了质的提高。Wang 等人[21]在 2016 年基于自然语言推理任务设计了一种特殊的 LSTM 模型 Match-LSTM, 后来将这种模型延用到阅读理解任务上, 并且利用指针网络(Pointer Network, Ptr-Net)[22]提出了两种答案预测方法, 这两种 Match-LSTM 模型[23]都大大优于 Rajpurkar 等人[2]使用逻辑回归和手工构建特征的基线方法。Xiong 等人[24]引入了协同注意机制, 并且设计了一个动态解码器迭代地预测答案开始位置和结束位置。Seo 等人[25]提出类似 DCN[24]的双向注意力流模型, 相比较而言 BiDAF 模型[25]更简单, 但是效果更好。2018 年初, 阿里巴巴的 SLQA+模型[26]和微软亚洲研究院的 R-Net 模型[27]在 EM 值上分别以 82.440 和 82.650 的分数先后超过人类水平。

微软发布的 MS MARCO 数据集[4]可以被看作是 MRC 继 SQuAD[2]之后的又一个里程碑。MS MARCO

是建立在经过匿名处理的真实世界问题和文档的数据基础之上, 更贴近现实世界。英文领域的 MRC 数据集和竞赛发展得如火如荼, 中文 MRC 的研究由于缺乏大规模高质量的数据集而进展缓慢。但百度在 2018 年发布了与微软 MS MACRO 结构类似的中文 MRC 数据集 DuReader[5], 打破了这一僵局。

自 CNN&Daily Mail[1]数据集之后, 各种大规模的数据集被相继公开发布, 除了上面介绍的以外, 代表性的工作还有 CBT[28]、RACE[29]、NewsQA[30]、TriviaQA[31]、SearchQA[32]、NarrativeQA[33]等。谷歌与卡内基梅隆大学联合发文[34]审视了数据对训练深度学习模型的重要性, 每一个新的数据集都提出了新的问题, 新问题促使研究人员不断探索新模式, 从而推动机器阅读理解领域的发展。

3 机器阅读理解的任务定义

MRC 任务的定义可以概括总结为: 给定三元组 (P,Q,a) , 其中 P 是与问题 Q 相关的篇章, a 是问题对应的正确答案。训练模型使得 a 作为问题 Q 的答案概率最大, 即条件概率 $\mathrm{p}(a\,|\,P,Q)$ 最大化。

Chen 在其博士毕业论文[35]中根据答案的类型将 MRC 问题分为四类: 完形填空、多项选择、跨度抽取、自由作答。本文延用这四种分类, 介绍各类 MRC 任务的区别和相应的评价指标, 并且例举一些有代表性的数据集, 以及解决每一类任务的神经机器阅读模型。

3.1 完形填空

受语言能力考试启发, 完型填空是最早出现的 MRC 任务。在完形填空类型的数据集中, 问题是通过删除文章中的一些单词或实体来产生的。完形填空最突出的特点是, 答案是上下文中的单词或实体, 这个任务可以看作是单词或实体的预测。这类任务使用简单的准确度作为评价指标, 即系统给出正确答案的百分比。

CNN&Daily Mail[1]是最典型的完形填空类 MRC 数据集, 包含 93000 篇来自美国有线电视新闻网[15]的报道文章和 22000 篇来自每日邮报[16]的新闻数据。这些新闻报道都会有一句抽象概括而非简单复制原文句子的总结, Hermann[1]通过使用一个占位符取代总结句中的某个实体, 来构造完形类型的问题。为了避免实体共现统计和外部知识对 MRC 的影响, 将所有出现的命名实体进行匿名化, 让模型更关注于挖掘上下文的语义关系, 以提高模型的泛化能力。哈工大讯飞联合实验室使用类似的方法, 基于人民日报和“儿童读物”构建了填空型中文 MRC 数据集 HFL-RC[19], 填补了中文阅读理解的空白。完型填空类的数据集还有 CBT[28]、CLOTH[36]、CliCR[37]等。

Attentive Reader 和 Impatient Reader 是针对 CNN&Daily Mail 数据集首次使用神经网络来解决 MRC 任务的模型。Attentive Reader[1]先用 BiLSTM 编码篇章和问题的上下文语义表示, 将问题表示为头部单词反向隐藏状态和尾部单词正向隐藏状态的连接, 然后用 tanh 函数来计算问题与每个篇章词之间的注意力权重, 再经过非线性变换, 最终输出篇章的注意力加权表示和问题表示的联合表示来预测答案。与 Attentive Reader 不同, Impatient Reader[1]是对于每个问题词递归地计算相应的篇章注意力加权表示, 直到遍历完所有问题词得到最终的篇章表示。Attentive Reader 的设计思路是带着整个问题出发阅读文章找答案; 而 Impatient Reader 则是每读一个问题词就回顾一下整个文章, 多次反复, 直到读完整个问题再作答。Chen[17]针对 Attentive Reader 进行了优化, 设计出一个双线性匹配函数取代 tanh 函数来计算注意力权重, 在得到注意力加权的篇章表示之后, 没有再次通过非线性层将其和问题的表示进行结合, 而是直接将其归一化后做了最终预测。AS Reader[18]也是 Attentive Reader 的变体, 使用点积简化了注意力函数的计算, 并且直接累加相同篇章词的注意力权重作为最后的答案概率, 使得模型更简单。Cui 等人[19]受 Impatient Reader[1]和 AS Reader[18]的启发, 为每个问题词分别计算篇章词的注意力分布, 再合并所有独立的结果得到最终篇章的注意力分布, 最后答案预测层与 AS Reader 一样。基于 CAS Reader[19]的工作, Cui 在 2017 年又提出了 AoA Reader[20], 模型计算了对于整个篇章而言平均的查询词注意力分布, 再基于此分布来将所有独立的篇章注意力分布进行加权求和, 而不是像 CAS Reader 直接进行累加, 从而提高了模型提取特征的能力。

表 2 模型在 CNN 数据集上的性能比较

Table 2 Model comparision on the CNN dataset

模型名称	CNN News	
	Valid	Test
Attentive Reader[1]	61.6	63.0
Impatient Reader[1]	61.8	63.8
Standford Attentive Reader[17]	72.4	72.4
AS Reader[18]	68.6	69.6
CAS Reader[19]	68.2	70.0
AoA Reader[20]	73.1	74.4

3.2 多项选择

机器阅读理解的多项选择任务, 与人类语言能力考试相似, 要求机器根据给定的文章从多个候选答案中选择唯一正确的选项。相比于完形填空, 多项选择题的答案形式更加灵活, 不再局限于上下文中的单词或实体, 但这个任务必须提供候选答案。多项选择类的 MRC 任务和完形填空一样, 都以系统回答的准确率评价其性能。

微软在 2013 年发布的 MCTest[11]是第一个多项选择类型的高质量数据集。它是由 500 个虚构的故事组成, 每个故事有 4 个相关的问题, 每个问题有 4 个选项。因为故事是虚构的, 所以答案只能在故事里找到, 从而避免了外部知识的影响。MCTest 最大的缺点就在于数据量太小, 无法用来训练复杂的神经网络模型。RACE[29]数据集是来源自中国初高中英语考试试题, 文章类型多样, 克服了其他数据集中普遍存在的主题偏差问题。这些文章和问题都是专家专门设计来评估学生阅读理解能力的, 简单的上下文匹配技术可能表现不佳, 因此推理能力是正确回答这些问题必不可少的。与 MCTest 数据集相比, RACE 包含了 27933 篇文章和 97687 个问题, 拥有足够支持深度学习模型训练的数据量。2019 年, 腾讯 AI 实验室发布了首个中文多项选择 MRC 数据集 C3[38], 类似于 RACE[29], C3 的阅读材料来自汉语水平考试(HSK)和民族汉语考试(MHK)的试题和练习题, 一共包含 13369 篇文章和 19577 个问题。

Zhu 等人[39]提出了分层注意流, 从词到句逐层计算注意力, 充分建模文章、问题和候选项之间的相互作用。利用候选项来提高从篇章中收集证据的效果, 并且基于注意力机制建模候选项之间的相关性, 得到更好的候选项表示。DFN[40]不仅在推理步骤上是动态的, 而且在执行注意力的方式上也是动态的。通过动态选择最优的注意力策略, DFN 对篇章、问题和候选项的表示进行融合, 以满足不同类型问题对不同理解能力的需求。Wang 等人[41]提出一种 co-matching 方法来将篇章并发地与问题和候选项进行匹配, 并且使用层次的文本建模方法, 使得信息从单词级聚合到句子级, 再从句子级聚合到文档级。DCMN+[42]是对 HCM[41]的改进, 将人类通常使用的两种阅读策略整合到模型中: 1)篇章句选择, 从篇章中找出最明显的支持性句子来回答问题; 2)答案候选项交互, 引入答案候选项之间的比较信息, 使得每个选项不再是孤立的。此外, 对于(问题、篇章、候选答案)三元组, 使用对偶的双向 co-matching 网络来建模两两之间的关系, 并且利用门控机制融合来自两个方向的表示。哈工大讯飞联合实验室[43]提出了一种基于卷积空间注意力的模型, 能够充分提取出篇章、问题和候选项之间的交互信息, 形成丰富的表征。通过整合文章和问题信息, 形成文章、问题和候选项之间的三维注意力空间, 重点对候选项的不同语义方面进行建模。此外, 提出了一种卷积空间注意力(CSA)机制, 从空间注意力中动态提取具有代表性的特征。

表 3 模型在 RACE 数据集上的性能比较

Table 3 Model comparision on the RACE dataset

模型名称	RACE-M	RACE-H	RACE
HAF[39]	45.0	46.4	46.0
DFN[40]	51.5	45.7	47.4
HCM[41]	55.8	48.2	50.4
DCMN+[42]	73.2	64.2	67.0
CSA[43]	52.2	50.3	50.9

3.3 跨度抽取

在实际应用中的阅读理解数据, 有时候单一的实体或单词不足以回答问题, 有时候没有提供候选答案。跨度抽取式 MRC 是一种更接近真实场景的任务, 要求机器从文章中抽取一段文本作为答案。这类问题也被称为抽取式问答, 其评价指标通常使用 F1 值和精确匹配(Exact Match, EM)值来进行综合评价。

2016 年, 斯坦福大学自然语言计算组发起的 SQuAD[2]机器阅读理解挑战赛, 被誉为“机器阅读理解界的 ImageNet[44]”。该项目组基于 536 篇维基百科文章通过众包的方式提出了 107785 个问题, 并让标注人员从文章中选择任意长度的文本片段来回答问题。与以前的数据集相比, SQuAD 数据集规模大且质量高, 促进了各种神经机器阅读模型架构的不断改进。在中文机器阅读理解方面, 第二届“讯飞杯”中文机器阅读理解测评开放了首个人工标注的中文篇章片段抽取型阅读理解数据集 CMRC2018[45], 由人类专家在 Wikipedia 段落上注释的近 20000 个真实问题组成。类似的跨度抽取式 MRC 数据集还有 NewsQA[30]、TriviaQA[31]等。

SQuAD 挑战赛受到了学术界和产业界的高度重视, 各种模型也应运而生。2016 年, Wang 等人提出了 Match-LSTM[21]来解决文本蕴含问题。由于 SQuAD 的数据特性, Wang 等人[23]继续采用 Match-LSTM 来对篇章和问题之间的匹配进行建模, 并且基于指针网络(Ptr-Net)[22]设计了序列预测和边界预测两种答案预测方法。序列模型依次预测每一个答案词的位置, 最终预测出的答案不一定是篇章

中连续的片段; 而边界模型简化了整个过程, 只预测答案跨度的开始和结束位置。BiDAF[25]分别从字符级别、单词级别和上下文嵌入级别的不同粒度来表示篇章和问题, 并且在问题篇章交互层中计算了两者的对齐矩阵, 基于该矩阵计算双向注意力得到问题感知的篇章表示。特别地, 为了避免早期总结造成的信息损失, 将前面的嵌入表示和注意力表示一起流入建模层, 进一步捕获基于查询的篇章词之间的交互, 最终使用两个全连接层分别预测答案跨度的开始和结束位置。类似于双向注意力, Xiong 等人[24]引入了协同注意力机制同时关注篇章和问题, 并融合两者的注意力表示来捕捉篇章和问题之间的交互。由于篇章中可能存在几个直观的答案跨度, 每个都对应着局部最大值, 所以一个动态指针解码器被设计来交替预测答案跨度的开始位置和结束位置, 使得模型能够从错误答案对应的局部最大值中恢复。后来, Xiong 等人又提出了 DCN+[46]模型, 在 DCN 协同编码器的基础上进行了两点改进: 1)通过堆叠协同注意力层来实现带有自注意力的协同注意力编码器, 从而构建更丰富的表示; 2)通过残差连接将每一层协同注意力的输出进行合并, 减少了信号路径的长度。

2018 年初, 阿里巴巴的 SLQA+模型[26]和微软的 R-Net 模型[27]先后在 EM 值上历史性地超越人类性能, 双方由于不同维度的评测结果各有千秋, 当时获得了 SQuAD 挑战赛并列第一的成绩。这两个模型都是模仿人类做阅读理解问题的方式: 1)编码层: 首先通读全文初步了解文章主题, 读题以了解提问内容; 2)交互匹配层: 将篇章和问题联系起来, 带着问题去理解篇章; 3)自匹配层: 综合全文去收集证据找到一个大概的答案范围; 4)答案预测层: 回到问题, 根据前面找到的答案范围选择最佳答案。R-Net 模型[27]受 Match-LSTM[21]的启发, 在基于注意力的 RNN 中加入了门控机制, 门控值大小表明了篇章中的词语在回答问题时具有不同的重要性, 忽略掉不相关的篇章部分而强调重要的部分。此外, 在得到问题感知的篇章表示之后, R-Net 模型[27]引入了自匹配机制来更进一步细化篇章的表示, 对于篇章中每个当前词从其他篇章词处收集证据。SLQA+[26]模型构建的主要思想是借助分层策略, 逐步集中注意力捕捉篇章中与问题关联的特定区域, 使答案边界清晰; 最大的创新点在于使用融合函数来结合不同粒度的语义内容, 从而更好地理解篇章, 避免对细节过度关注。

3.4 自由作答

从完形填空到多项选择再到跨度抽取, 虽然 MRC 研究在一步步逼近现实数据, 但现有的 MRC 任务限制了答案是来自篇章的连续文本片段, 这距离实现真正机器智能化还相差甚远。自由作答 MRC 任务不再限制答案的形式, 允许答案是任何自由文本, 更接近实际应用的阅读理解。由于标准答案是人为生成的, 所以在评价指标上还未达成共识, 目前广泛采用自然语言生成任务中的评价指标 BLUE[47]值和 ROUGE[48]值。

表 4 模型在 SQuAD 数据集上的性能比较

Table 4 Model comparision on the SQuAD dataset

模型名称	EM	F1
Match-LSTM[23]	64.7	73.7
BiDAF[25]	68.0	77.3
DCN[24]	66.2	75.9
DCN+[46]	75.1	83.1
SLQA+[26]	82.4	88.6
R-Net[27]	82.6	88.5

基于必应搜索引擎从匿名的真实用户查询中构建的 MS MARCO[4]是最具代表性的自由作答式 MRC 数据集。该数据集的构建过程为: 1)从搜索日志中过滤掉非问题的查询; 2)使用搜索引擎检索每个问题的相关篇章; 3)从这些篇章中自动提取相关段落; 4)对包含有用信息的段落进行人工标注, 并概括这些信息形成对应问题的答案。MS MARCO 包含了 1010916 个问题, 以及从 3563535 个网页中抽取的 8841823 个篇章段落。这些问题是来自真实用户的查询, 并且其中有一部分没有答案, 更能够代表信息需求的“自然”分布。2018 年, 谷歌 DeeppMind 推出了第一个基于整本书或者整个剧本的大规模问答数据集 NarrativeQA[33], 需要机器理解更复杂的文本, 回答更难的问题。NarrativeQA 数据集包含了 1572 个故事和 46765 个问题, 问题和答案都是人为设计的, 大多数是复杂的形式, 因此需要更高层次的抽象和推理。DuReader[5]是第一个中文自由作答类型的 MRC 数据集, 与 MS MARCO[4]一样面向真实应用, 问题和篇章来自于百度知道和百度搜索引擎, 答案是标注者手动生成的。DuReader 共包含了 30 万问题、66 万的答案和 140 万的篇章, 并提供了包括是非类和观点类的更加丰富的问题类型。DuReader 是目前最大的中文 MRC 数据集, 在中文应用中具有开创意义。

微软提出的 gQA[49]是生成式问答模型, 不仅可以有效地对问题和篇章之间的关系进行建模, 而且根据数据的性质决定答案是抽象生成还是从篇章中抽取。编码器模拟篇章和问题之间的关系, 并将它们

编码成一个向量, 从而方便解码器直接从中形成答案。解码器中引入了复制机制和覆盖机制, 复制机制学习何时直接从篇章中复制一个重要的实体, 而不是从词汇表中生成任何内容; 覆盖机制通过跟踪哪些篇章词被注意了太多次来避免重复生成。Tan 等人[50]提出一个先抽取再合成的框架 S-Net。证据抽取模块的目的是预测篇章中最重要的子跨度, 通过设计一个多任务联合学习的框架, 利用篇章排序来提高文本跨度预测的准确率。答案合成模块使用序列到序列的模型, 将抽取出的证据作为额外的特征来生成初步的答案, 并且基于规则对其进行处理得到最终的答案。Masque[51]是一种多风格的抽象摘要模型, 突破了抽取式答案的限制, 采用摘要的方式生成指定风格的答案。通过指定答案开始位置的人造词条, 在单个系统中实现多种风格答案的生成。模型架构由阅读器和解码器组成, 阅读器为了提高阅读理解的能力, 将阅读理解、篇章排序和问题可回答性进行多任务联合学习。解码器中的多源指针-生成器机制, 使得模型既可以从词汇表中生成单词也可以从问题或者篇章中复制单词。

3.5 小结

如表 5 所示, 在 MRC 研究的进程中, 为了让机器更智能化, 能够处理更接近真实世界的数据, 各种带着新问题的数据集相继被发布, MRC 任务的研究方向也越来越多样化。了解现有模型的性能有助于进一步识别现有数据集的局限性, 促使研究者寻求更好的方法来构建更具挑战性的数据集, 以实现机器理解文本的最终目标。

完型填空形式简单易于评估, 但答案被限制为篇章中的某个单词或者实体, 与实际情况不符, 不能很好地测试机器是否理解了文本内容。多项选择为问题提供了候选答案, 虽然取消了对答案形式的限制, 但仍与实际应用有很大的出入。候选项为问题提供了一些额外的信息, 减少了机器对篇章内容的理解, 比如可以通过排除法进行作答, 因此也不能很好地评估机器阅读理解的能力。自由作答任务的答案形式具有灵活性, 需要机器具有真正的理解能力和强大的推理能力, 最能够反映机器的阅读理解

表 5 机器阅读理解数据集对比

Table 5 Comparison of different machine reading comprehension datasets

数据集名称	语言	答案类型	新趋势	问题数	篇章数	语料来源
CNN&Daily Mail[1]	英文	完形填空	—	387K+997K	93K+22K	新闻报道
HFL-RC[19]	中文	完型填空	—	877K	60K	人民日报和儿童童话
CBT[28]	英文	完形填空	—	687K	108	儿童图书
CliCR[37]	英文	完形填空	—	105K	12K	医学的临床病例报告
CLOTH[36]	英文	完形填空/多项选择	—	99K	7K	初高中英语考试试题
MCTest[11]	英文	多项选择	—	2640	660	小说故事
RACE[29]	英文	多项选择	—	97K	28K	初高中英语考试试题
MCScript[52-53]	英文	多项选择	基于知识的问答	20K	3500	日常生活场景故事
DREAM[54]	英文	多项选择	对话式问答	10197	6444	英语考试试题
C3[38]	中文	多项选择	—	20K	13K	汉语考试试题
WikiHop[55]	英文	多项选择	多篇章	49K	669K	Wikipedia 作为文档语料库, Wikidata 作为知识库
SQuAD[2-3]	英文	跨度抽取	不可回答问题	88K/130K	536/505	维基百科文章
NewsQA[30]	英文	跨度抽取	不可回答问题	120K	13K	新闻报道
TriviaQA[31]	英文	跨度抽取	多篇章	95K	650K	Trivia web 爱好者撰写问答对, 维基百科和必应搜索
CMRC2018[45]	中文	跨度抽取	—	19K	5K	中文维基百科
HotpotQA[56]	英文	跨度抽取	多篇章	113K	536	维基百科
MS MARCO[4]	英文	自由作答	多篇章/不可回答问题	1M	3.6M 网页数 (8.8M 段落数)	必应搜索
DuReader[5]	中文	自由作答	多篇章	300K	1.4M	百度搜索和百度知道
SearchQA[32]	英文	自由作答	多篇章	140K	6.9M	智力问答和谷歌搜索
CoQA[57]	英文	自由作答	对话式阅读理解	127K	8K	7 个不同的领域
NarrativeQA[33]	英文	自由作答	—	47K	1572	书籍、电影脚本

(注: “+”表示两个数据集的规模; “/”表示不同版本的规模)

能力, 也是最接近于实际应用的。然而, 自由形式答案的数据集难以构建, 文本生成技术不成熟, 以及评价指标没有统一的标准, 使得自由作答 MRC 研究发展缓慢。跨度抽取任务要求机器从文章中抽取一段文本作为答案, 相对完形和多选任务更贴近真实的应用, 与自由作答相比数据集更容易构建并且有标准的评价指标, 是最好的中间选择。跨度抽取任务得到了学术界和工业界的广泛关注, 目前研究技术比较成熟, 机器阅读理解的性能已经超过了人类水平, 很多自由作答的研究工作也是使用跨度抽取的技术进行最终答案预测的。

4 机器阅读理解的发展趋势

为了让机器更智能, 更接近真实的应用场景, 机器阅读理解任务的难度越来越大, 除了答案形式上从完形填空到自由作答, 此外还出现了很多新的研究趋势, 本节将详细介绍。

4.1 不可回答问题

Jia 等人[58]基于一组简单的规则在段落中自动生成一个与问题相关但非答案句的干扰句, 以此来混淆模型。结果分析表明, 在这种对抗性的实验中, 在 SQuAD 上成功的模型性能下降了一半, 原因在于 SQuAD 数据集的前提是问题一定能在文章中找到答案。因此, 斯坦福大学在 2018 年发布了升级版本的 SQuAD2.0[3], 在 SQuAD1.1[2]版本的数据基础上添加了 5 万多个无法回答的问题。这些新增的问题是由标注者根据最初的答案人为反向设计出来的, 与原文高度相关, 并且都有貌似合理的答案, 这就要求模型具有能够判断出问题是否可作答的能力。

从 1.0、1.1 再到 2.0, SQuAD 数据集和比赛难度持续升级。SQuAD2.0[3]不仅要求阅读理解模型在问题可回答时给出答案, 还要能判断出不可回答问题并拒绝回答。Clark 等人[59]在模型的最后额外添加了一个层来计算“无答案分数”, 表示问题不可回答的概率。在无答案分数和答案跨度分数上应用一个共享的 softmax 函数来标准化, 从而产生一个联合的目标函数。U-Net[60]将具有无法回答问题的 MRC 分解为三个子任务: 答案预测、无答案检测和答案验证, 并且利用多任务联合学习来训练模型。答案预测器预测问题的候选答案跨度; 无答案检测器在问题没有答案时拒绝选择任何文本跨度; 答案验证器基于候选答案跨度来决定问题不可回答的概率。此外, 还引入了一个具有抽象语义的通用节点, 作为一种特殊的信息载体, 篇章和问题的注意力信息都集中于这个节点, 使它们之间的联系比传统的双向注意力交互更加紧密。Hu 等人[61]提出了先阅读再验证的架构, 通过神经机器阅读模型来预测候选答案并给出初步的无答案分数, 再经过答案验证器来判断抽取出的候选答案是否合理, 即问题是否无答案。对于机器阅读模型, 通过引入两个辅助损失来缓解答案抽取和无答案检测任务之间的冲突。在答案验证方面研究了三种不同的网络结构: 第一种是简单的序列模型, 将包含候选答案的句子、问题和抽取出的答案依次连接作为一个整体序列来处理; 第二种是交互式体系结构, 通过计算答案句和问题之间的匹配关系来预测答案不可回答性; 第三种是将前两种方法集成在一起, 连接两个模型的输出得到一个联合表示来计算无答案概率。

4.2 多篇章阅读理解

多篇章阅读理解是单篇章的扩展, 任务要求 MRC 模型从多篇文档中推理预测出问题的答案。在多篇章 MRC 的设置下, 能否快速、准确地从大量的文档语料库中检索到最相关的文档, 直接关系到问答系统的最终性能。对于一些复杂的问题, 证据片段可能出现在一篇文章的不同段落甚至不同的文章中, 问答模型需要收集这些散落的证据线索, 并且从中推理出正确的答案。因此, 多篇章阅读理解任务更具挑战性。MS-MARCO[4]是典型的多篇章数据集, 其中每个问题平均有 10 篇相关的文档, 这些文档是由必应搜索引擎检索到的相关网络文章。类似地, DuReader[5]数据集中每个问题有 5 篇来自百度搜索和百度知道的相关文章。

针对 MS-MARCO[4]和 DuReader[5]这样的多篇章数据集, Wang 等人[62]提出了多任务联合学习训练的 V-Net 模型, 主要包含三个不同的模块: 答案边界预测、答案内容建模和答案验证。答案边界预测和答案内容建模是从不同角度对候选答案进行建模, 因为不同的候选答案通常有着相同的答案边界、不同的答案内容。不同的篇章会产生多个候选答案, 边界预测和内容建模都是聚焦于单篇文章的信息处理, 而没有考虑不同篇章之间的交互, 所以篇章之间的答案验证能够将不同篇章中的信息聚集起来为每一个候选答案计算验证分数。Liu 等人[63]针对多答案问题提出了一个多答案多任务的框架。不仅提出了三种损失函数来整合同一个问题的多个不同的参考答案; 而且通过添加辅助损失来预测正确的篇章并从中抽取答案; 此外, 还使用最小风险训练来解决同一个答案多次出现的问题。阿里巴巴团队[64]提出一个在线问答系统, 设计了一个深度级联学习的框架来平衡效率和效果, 可有效降低召回阶段延时, 并

最大化答案准确率。框架主要包括三个级联的模块: 文档检索模块、段落检索模块和答案抽取模块。使用简单的特征和模型过滤掉不相关的文档内容, 逐步从文档级别到段落级别演化, 最后在有限的备选段落中进行更精确地答案抽取。这个系统结合了流水线和联合学习的优点, 从粗到细粒度地定位答案。Hu 等人[65]提出一个端到端的统一问答模型 RE^3QA, 将上下文检索、阅读理解和答案重排结合在一起来预测最终的答案。不同的组件之间共享上下文文本表示, 使用高质量的上游输出直接监督下游模块。通过对文本的不断剪枝, 一步步地缩小搜索范围, 并且引入了非最大抑制算法来删除冗余的候选答案。

表 6　模型在 MS MARCO 数据集上的性能比较
Table 6　Model comparision on the MS MARCO dataset

模型名称	ROUGE-L	BLEU-1
S-Net[50]	45.23	43.78
V-Net[62]	46.15	44.47
Masque[51]	59.87	54.11

更多的文档意味着更多的信息, 有助于得到更完整的答案, 这种扩展可用于处理基于大量非结构化文本的开放域问答任务。开放域问答旨在从大量的知识源中为广泛的问题寻找答案, 例如结构化的知识库以及来自搜索引擎的非结构化文档。SearchQA[32]是最具代表性的开放域问答数据集, 构建过程如下: 首先从美国益智问答游戏节目《危险边缘》收集所有的问题-答案对; 再使用谷歌搜索问题得到一组实际的相关/不相关的文档; 最后过滤掉那些无法在检索到的文档中找到答案的问题, 以及那些谷歌返回的相关网页数少于 40 的问题。最终数据集包含了超过 140000 个问答对, 平均每对有 49.6 个相关文本片段。

针对开放域问答, 单独的 MRC 系统不能完成整体的任务。Chen 等人[66]将维基百科作为唯一的开放域知识源, 先利用高效的文档检索系统来缩小搜索空间, 再基于 Stanford Reader[17]从单个文档或小的文档集合中提取答案。Wang 等人[67]提出了一种新的开放域问答系统, 首先使用信息检索模型得到与问题最相关的前几篇文档, 然后经过 R^3 模型选择文档并从中预测答案。R^3 模型主要由排序器和阅读器组成, 排序器输出一个篇章采样策略或者篇章的概率分布, 选择最可能包含答案的篇章作为阅读器的输入, 阅读器从该篇章中抽取问题对应的答案跨度。排序器和阅读器是联合训练的, 排序器基于强化学习进行训练, 奖励取决于阅读器如何从采样的篇章中抽取答案; 阅读器使用标准的随机梯度下降法训练, 以最大化预测正确答案的概率。Wang 等人[68]提出了两种答案重排策略, 以进一步明确收集来自不同篇章的证据。首先, 与错误的答案相比, 正确的答案往往被更多的文章反复提及, 因此提出了基于证据强度的重排策略, 对候选答案的出现进行计数或者概率计算。此外, 有时答案需要满足问题所有方面的补充证据, 这些证据可能分布在不同的篇章中, 因此提出了基于证据覆盖的重排器, 根据不同篇章中的联合证据对问题的覆盖程度来重排候选答案。DS-QA 模型[69]是一个从粗到细粒度降噪的远距离监督的开放域问答系统。首先通过信息检索从一个大型语料库中根据问题检索段落; 然后采用段落选择器来快速地浏览所有检索到的段落并过滤掉噪声段落; 再通过一个精确的段落阅读器, 对每个选定的段落进行仔细阅读并抽取答案; 最后将所有选定段落的抽取结果聚合起来得到最终的答案。Pang 等人[70]提出了一个层次答案跨度模型, 为开放域问答任务设计了一个新的概率公式, 基于问题、篇章和答案跨度三个级别的层次结构决定最终的答案概率。根据总概率定律, 给定问题的答案概率可以定义为篇章概率和条件答案概率的乘积。篇章级别的概率用来衡量篇章的质量, 使得模型对于无答案问题具有鲁棒性。条件答案概率是计算给定篇章中答案跨度的概率, 使用聚合器来结合多个相同文本跨度的信息。跨度级别的概率表示文本跨度作为问题答案的概率, 使用条件指针网络定义答案跨度的概率。

4.3　基于知识的阅读理解

由于自然语言本身存在模糊性和含蓄性, 所以外部知识在人类阅读理解和语言理解中起着至关重要的作用。之前的很多研究都是针对阅读理解其他方面的问题, 有时甚至明确地排除外部知识的影响来构建数据集(CNN&Daily Mail[1]、CBT[28]等), 这与实际的人类阅读理解过程不符。因此, 外部知识的运用是人类和机器阅读理解最大的区别。MCScript[52-53]数据集是关于人类日常活动的故事, 比如去看电影或者在餐厅里吃饭等, 很多问题超出了文本的范围, 需要使用常识进行推理。该数据集通过众包工人生成文本、问题和答案, 大约包括了 3500 个文本和 20000 个问题, 为结合常识进行推理提供了一个评估框架。

外部知识的重要性逐渐引起了研究者的关注, 也取得了一些研究成果。Long 等人[71]提出了一种针对罕见实体进行完形填空的预测任务, 利用 Freebase 数据库中抽取的实体描述作为外部知识, 辅助预测

缺失的命名实体。Mihaylov 等人[72]提出了一种将外部知识整合到 AS Reader[18]模型中的方法, 将知识信息编码到与普通词条表示相同的向量空间中, 再分别与篇章和问题表示进行结合来丰富文本的表示, 从而提高预测正确答案的能力。KT-NET[73]从知识库中自适应地选择所需的知识, 然后采用一种注意力机制将外部知识与 BERT[74]的隐藏状态进行融合, 从而得到知识感知的表示用作答案预测。

4.4 对话式阅读理解

对话式阅读理解将对话合并到 MRC 中, 结合了对话和阅读理解的难度, 因此更具有挑战性。对话式 MRC 有两种形式, 一种是基于某段文本展开多轮对话问答, 每个问题都需要根据文本内容并且结合会话历史来理解作答; 另一种是给定一段对话内容, 基于此提出问题并推理作答。CoQA[57]是第一种形式的数据集, 话题涉及七个不同领域, 包括了 8000 篇短文和 127000 个会话问答对。DREAM[54]数据集是第二种形式, 是基于对话文本的选择题阅读理解数据集, 数据来自英语外语考试, 包含了 6444 段对话和 10197 个问题。

Reddy 等人[57]在发布 CoQA 数据集的同时提出了一个基线模型, 该基线是 DrQA[66]模型和 PGNet[75]模型的组合。首先将 DrQA 预测出的上下文文本跨度作为答案依据, 再利用 PGNet 生成答案, 此外, 通过将历史问答序列附加到篇章词序列后面, 从而实现在理解文本的基础上结合会话历史来推理答案。GRAPHFLOW[76]是一种基于图神经网络(GNN)的模型, 可以捕捉对话中的会话流。模型动态地构建了一个问题感知的上下文图, 图中的顶点是由每个篇章词组成。在此基础上, 模型还提出了一种新的流机制来模拟上下文图序列中的时间依赖关系。Sun 等人[54]在发布 DREAM 数据集的同时提出采用端到端的预训练语言模型作为基线, 利用 FTLM 包含丰富的外部语言知识, 并且提出 speaker 嵌入来捕捉对话结构。针对 DREAM 的数据形式, Zhu 等人[77]提出了一种简单但非常有效的对偶多头协同注意力机制, 可以充分表达文章和问题之间的关系, 同时有效地与大型预训练的语言模型配合。

5 神经机器阅读模型

深度学习技术已经在计算机视觉、机器翻译等多个领域中取得了显著成果。各种大规模 MRC 数据集的发布, 使得机器阅读理解任务能够更好地运用深度学习技术。本节主要从注意力机制和推理策略的创新角度介绍各种神经机器阅读模型。

5.1 注意力机制的创新

注意力机制受人类视觉感知注意力的启发, 最早在计算机视觉和图像处理领域被提出。注意力机制被初次应用于自然语言处理领域得益于 Bahdanau 等人[78]的工作, 将翻译和对齐进行联合学习, 使得机器翻译性能得到了显著提升。注意力机制可以缓解目标序列和源序列之间的远距离依赖问题, 因此对于 MRC 任务中的长篇章推理问题, 注意力机制能够有效地建模信息交互。近几年的 MRC 研究中出现了许多基本注意力形式的变体, 包括多层次多粒度的注意力关注点和多角度的注意力建模的改进, 下面介绍各种基于注意力的方法。

Hermann 等人[1]首次在机器阅读理解任务中引入了注意力机制, 使用词条级别的注意力机制, 即计算篇章中每个词条的权重, 提出了 Attentive Reader 和 Impatient Reader 模型。Stanford Reader[17]和 AS Reader[18]都是针对 Attentive Reader 中的注意力计算函数进行的改进, 前者使用双线性函数而后者使用点积来计算注意力权重。BiDAF[25]在问题篇章交互层中引入了双向注意力机制, 基于篇章和问题的对齐矩阵分别计算 Context-to-query 和 Query-to-context 的注意力。类似地, Xiong 等人[24]在 DCN 模型中引入了协同注意力机制同时关注篇章和问题, 并融合两者的注意力表示来捕捉篇章和问题之间的交互。微软亚研院提出的 R-Net [27]在基于注意力的 RNN 上加入了额外的门控来决定篇章中信息对于问题的重要性; 此外, R-Net 首次引入了自匹配注意力机制, 从整个篇章收集与当前篇章词相关的证据, 更进一步细化了篇章的表示。SLQA+[26]模型采用层次注意力, 使用多粒度的融合函数来结合不同层次对齐的上下文表示。FusionNet[79]从三个角度对基本形式的注意力机制进行了改进: 1)提出“history-of-word”概念, 使用从词向量到语义表示的完整信息来构建注意力; 2)采用一个对称形式的注意力评分函数, 以有效地利用单词历史; 3)引入全感知多层次的注意力机制, 独立地捕捉不同层次的信息, 并且将捕捉到的所有信息结合起来得到全感知的篇章表示。Hu 等人[80]在 R.M-Reader 中提出了重注意力机制, 在多轮对齐结构中不断优化对齐矩阵, 从而避免了注意力冗余和注意力不足的问题。Zhuang 等人[81]提出一个动态自注意力架构, 使用门控机制动态决定哪些词条用来构建篇章内部或者篇章之间词条级别的语义表示。通过门控来获得输入序列中每个词条的重要性, 抽取出最重要的部分词条用于计算自注意力, 而不是基于整个文本词条序列来计算。

表 7 注意力函数 $S(x, y)$

Table 7 Attention function $S(x, y)$

函数名称	公式
加性注意力	$s^{\mathrm{T}}\tanh\left(W_1x+W_2y\right)$
乘法注意力	$x^{\mathrm{T}}U^{\mathrm{T}}Vy$
缩放的乘法注意力	$\frac{1}{\sqrt{k}}x^{\mathrm{T}}U^{\mathrm{T}}Vy$
缩放的非线性乘法注意力	$\frac{1}{\sqrt{k}}f\left(Ux\right)^{\mathrm{T}}f\left(Vy\right)$
对称形式	$x^{\mathrm{T}}U^{\mathrm{T}}DUy$, D 是对角矩阵
非线性的对称形式	$f\left(Ux\right)^{\mathrm{T}}Df\left(Uy\right)$

(注: f 函数表示非线性函数)

表 8 注意力创新

Table 8 Attention innovation

模型名称	创新点
Stanford Reader[17]、AS Reader[18]	注意力函数的不同
DCN[24]、BiDAF[25]	双向注意力流
R-Net[27]	自注意力机制
SLQA+[26]	多粒度融合
FusionNet[79]	全感知注意力
R.M-Reader[80]	重注意力
Dyn-SAN[81]	动态自注意力

5.2 推理策略的创新

自然语言在表述过程中存在很大的模糊性甚至出现歧义, 因此需要结合语境分析语义。对于计算机而言, 这种方式很难通过符号表达进行逻辑推理。提高机器阅读理解的性能, 关键是赋予机器推理和归纳能力。本节主要介绍两个经典的多跳推理 MRC 数据集以及一些有代表性的神经机器阅读推理模型。

Johannes 等人[55]在 QAngaroo 项目中派生了两个多跳推理数据集: WIKIHOP 和 MEDHOP。这两个数据集发布的目的就是评估系统跨多个文档的多跳推理能力。数据集由三元组(主语实体 s, 关系 r, 宾语实体 o)组成, 与实体相关的文章被加入到证据文档集 D 中。通过去除掉三元组中的宾语, 得到查询 $q=\left(s,r,?\right)$, 以及答案 $a^*=o$。为了达到多跳推理的目的, 在语料库构建的基础上构造了二分图, 一边的顶点对应着知识库中的实体, 另一边是 D 中的文档, 边表示实体出现在对应的文档中。对于给定的 $\left(q,a^*\right)$ 对, 利用广度优先搜索遍历二分图识别得到候选答案 C_q, 访问到的文档将成为支持文档 S_q。Wikihop 数据集是将 Wikipedia 作为文档语料库, 将 Wikidata 当作结构化的知识三元组。Medhop 是为分子生物学领域构建的数据集, 使用 DrugBank 作为结构化的知识库, 来自 MEDLINE 的研究论文摘要作为文档语料库。Yang 等人[56]提出了面向自然和多跳问答的数据集 HotpotQA, 包含了 113000 组问答数据。Hotpot 数据集有三个主要特点: 1)通过众包的方式收集问题和答案, 保证每个问题必须进行多步推理才能得到答案; 2)问题类型多样, 不受限于任何预设的知识库; 3)提供回答所需的句子层面的推理线索, 提高答案推理的可解释性。

MemNNs[28]是最早提出的推理模型, 将基本的记忆网络进行堆叠扩展成多跳, 不断地更新问题表示和篇章表示来完成推理过程。Trischler 等人[82]提出了一个由抽取器和推理器组成的 EpiReader 模型。抽取器就是使用 AS Reader 模型比较问题与篇章的浅层关系, 过滤筛选出少量的候选答案。推理器将候选答案插入问题中形成假设, 然后基于文本蕴含的概念来计算每个假设与篇章句的一致性, 得到每个假设成立的概率。最后将推理器的证据概率和抽取器的理论概率结合, 得到最终的候选答案排序。IA Reader[83]设计了一个迭代推理的过程。对于每个时间步, 交替更新问题和篇章的注意力分布, 通过门机制得到新的问题和篇章表示, 再基于这些表示来更新推理状态。最后, 使用最终的篇章注意力权重分布来进行答案预测。与之前方法中使用固定的推理跳数不同, ReasoNet[84]在推断过程中引入了一个终止状态来释放对推理深度的约束。这种状态可以决定是否在消化中间信息后继续推理到下一轮, 或者在推断出现有信息足以产生答案时终止整个推理。推理轮次由篇章和查询动态建模, 并根据问题的难易程度自动学习。DCN 模型[24]提出了一种动态指针解码器, 交替预测答案跨度的开始和结束位置, 允许从初始不正确答案的局部最大值中恢复。Rajarshi 等人[85]针对开放域问答任务提出了一个检索器和阅读器交互迭代的框架。根据给定的查询, 段落检索器利用快速最近邻搜索算法执行最大内积搜索得到最相关的几个段落; 阅读器可以使用各种先进的 MRC 模型来替换, 目的是预测出一个答案文本跨度。该框架的最大创新点是多步推理器, 根据当前步的阅读器状态和查询向量更新产生新的查询表示, 新的查询表示被送回到检索器中, 基于此对语料库中的段落进行重排, 再经过阅读器重新预测答案。因此, 多步推理器为检索器和阅读器提供了一种相互通信的方式。TraCRNet[86]是一种基于 Transformer 的端到端的体系结构, 能够有效地处理大文档集。输入编码层是基于 Transformer 的层次模型, 从单词级别到文档级别编码处理每一篇文档以及问题; 多跳推理层使用

Universal Transformer 编码器对所有文档和问题的总表示进行多步迭代更新, 编码全局信息; 输出解码层基于堆叠的 Transformer 解码器来生成答案。在多步推理器的每一步中, 自注意力基于所有文档及问题来理解每个文档, 深度自回归也在所有文档之间建立连接, 以整合跨文档的全局信息。

除了上述的推理模型工作外, 还有一类研究是基于图的推理模型。如图 1 所示, 利用命名实体识别和指代消解技术从篇章中抽取出实体提及, 将实体提及作为顶点构建成一个图, 基于实体图进行线索推理和证据整合。Dhingra 等人[87]提出了一个共指依赖循环层, 用 Coref-GRU 来代替 GA Reader[88]中常规的 GRU, 根据相同实体的多次出现来追踪问题答案的线索。Coref-GRU 相当于带有共指边的有向无环图, 但它的缺点主要是共指边通常是某个句子的局部引用, 没有考虑其他有用的全局信息。为了实现全局语境的推理, MHQA 模型[89]除了共指边还引入了另外两种类型的边, 即不同篇章中相同实体之间的边以及上下文窗口中出现的两个实体之间的边。模型首先从篇章的编码表示中抽取出实体提及的表示, 与问题表示计算注意力值来初始化图的隐藏状态。然后使用图递归网络 GRN 或者图卷积网络 GCN 对每个图顶点的相关信息进行集成, 多次更新得到一系列的图状态。对于每个步骤得到的图隐藏状态, 分别将其与问题表示相匹配, 然后对所有步骤的匹配进行加权求和得到整体的匹配结果, 最终得到每个实体的概率分布。与 MHQA 的图结构相比, Entity-GCN[90]添加了第四种类型的关系, 即在任何两个孤立点之间添加补边, 防止出现非连通图。模型使用预训练的 ELMo 来初始化实体提及的上下文表示, 而不用训练编码器, 预处理之后存储于本地, 避免了昂贵的文档处理。通过转换顶点表示来处理多步推理, 每一步根据直接邻居信息来更新图中顶点的表示。在 Relation-GCN[91]中引入了门控机制来控制传播的信息量, 防止过去的信息被完全覆盖。与之前只有单类型结点的图模型不同, Tu 等人[92]提出了包含不同类型结点的异质图。每篇文档、每个候选以及从文档中抽取出的实体作为 HDE 图中的三种顶点, 并且根据文档、候选和实体之间的关系构建七种类型的边, 表示 HDE 图中不同的结构信息。分别计算文档、候选和实体与查询之间的协同注意力, 再经过自注意力池化, 得到不同粒度级别的查询感知表示, 用来初始化相应的顶点状态。在 HDE 图上的信息传播使用聚合-组合的方式, 即从邻居顶点聚合的信息再与顶点本身的信息进行组合, 基于门控来更新顶点的表示, 经过多次信息传播后, 所有文档、候选和实体顶点都有其最终的表示。受知识图的多跳推理启发, Kundu 等人[93]提出了基于路径进行多跳推理的 PathNet 模型, 直接操作非结构化的文本信息。首先通过实体之间的共现关系从不同的篇章中抽取得到(问题实体 → 中间实体 → 候选实体)的路径, 路径的长度表示推理的跳数; 然后基于上下文表示和篇章表示分别对这些路径进行编码以及打分; 最后将同一候选的所有路径分数相加得到该候选的概率分数。之前的方法模型只是隐式地结合了所有篇章中

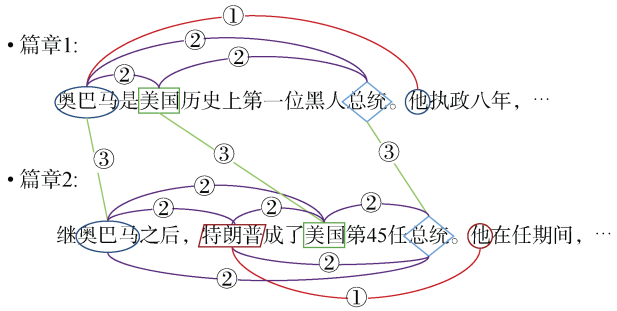

图 1 三种类型边的实体图

Figure 1 Entity graph considering three types of edges

(注: ①表示共指边; ②表示上下文窗口内不同实体之间的边; ③表示不同篇章中相同实体之间的边)

表 9 推理方法分类

Table 9 Classification of reasoning methods

基于文本蕴含的方法		EpiReader[82]: 将候选答案插入问题形成假设, 篇章句作为前提
基于答案迭代的方法		DCN[24]、DCN+[46]: 交替预测答案的开始位置和结束位置
基于表示更新的方法		MemNNs[28]、IA Reader[83] 、ReasoNet[84]、TraCRNet[86]: 迭代更新问题和篇章的表示
基于图的方法		Coref-GRU[87]: 共指依赖边
	基于同质图的方法	MHQA[89]: 共指边、相同实体边、上下文边
		Entity-GCN[90]: 共指边、相同实体边、上下文边、补边
	基于异质图的方法	HDE[92]: 不同类型的结点和边
	基于路径的方法	PathNet[93]: 基于实体共现得到推理路径

的知识, 无法提供明确的推理路径, 而 PathNet 能够通过显式的路径来解释其推理过程。

5.3 小结

神经机器阅读模型的关键在于注意力机制和推理策略的结合, 注意力机制匹配问题和篇章之间的关系, 推理发现线索为最终的答案找到证据。表 8 显示了各种模型的注意创新点, 双向注意力是从问题和篇章两个方向分别计算注意力权重; 自注意力是学习篇章内部的词依赖关系; 多粒度融合和全感知注意力是为了捕捉不同层次的信息; 重注意力是迭代更新注意力表示; 动态自注意力是基于重要的部分计算自注意力。表 9 对目前主流推理策略进行了归纳分类, 除了基于文本蕴含的方法, 其他的推理方法都是通过加深网络层数来模拟推理步数的不断增加。基于注意力机制进行有效地信息交互, 采用高效的推理策略来发现有力的证据并进行最终的答案预测, 这种注意力机制和推理策略结合的方式是现阶段解决 MRC 问题的主流技术思路。

6 存在的问题和应用展望

6.1 MRC 现阶段的不足

由于互联网信息化的飞速发展以及互联网固有的开放性和普及性, 不良敏感信息给网络空间造成的危害范围大、持续性强, 因此, 网络内容安全面临的形式和挑战尤为严峻复杂。基于机器学习的自然语言处理方法可以通过训练集自动学习知识, 克服了人工建模和特征提取的缺点, 从长远来看是对传统低效的人工审核方式的最佳替代。为了应对网络内容治理的挑战, 提高机器对自然语言的理解能力势在必行, 各种新数据集带着新的任务被相继发布, 促使研究人员不断地探索, 推动了 MRC 技术的飞速发展。尽管近年来在这一领域的研究中, 一些神经机器阅读模型在 SQuAD[2-3]等代表性数据集上的表现已经超越了人类, 但由于四种基本 MRC 任务固有的缺点, 与实际应用相差甚远, 因此出现了四种新兴的研究趋势。MRC 的四种新趋势是更接近现实应用的表现, 目前虽然有一定的成果, 但距离机器真正理解文本还有很长的路要走。对不可回答问题的准确判断、多篇章信息的联合推理、多轮对话中的指代消解、外部知识的有效融合等问题都还没得到很好的解决。下面具体分析神经机器阅读模型目前仍然存在的几点不足:

1) 系统的鲁棒性。在文献[58]设置的对抗性实验中, 人类并没有受到这些干扰句的影响, 但是神经机器阅读模型的性能显著下降, 这反映了机器并不能真正地理解自然语言。为了缓解这个问题, SQuAD2.0[3]通过总结对抗性攻击的模式, 基于人工规则经过复杂设计创建对抗性示例来扩充训练数据集。但是这种人工分析设计的方法缺乏通用性, 设计的对抗样本无法覆盖所有类型, 因此, 如何在更为通用的意义上解决 MRC 模型的鲁棒性问题仍有待研究。

2) 模型推理能力。对于多篇章 MRC, 需要整合不同篇章中散落的证据; 对于对话式 MRC, 由于人类对话的自然简洁等特点, 每个问题的理解都需要结合对话历史, 因此, 模型推理能力是找到答案线索的关键。尽管各种创新性的推理机制被提出, 但目前的神经机器阅读模型大多是基于篇章和问题之间粗浅的语义匹配来进行答案预测的, 效果仍然不佳。如何强化机器的推理能力, 发现深处隐藏的线索, 将是推动 MRC 发展的关键问题。

3) 外部知识引入。常识的累积和知识的记忆是人类智能的重要体现, 在人类阅读理解的过程中通常会利用外部知识来回答问题, 比如在对话时为了进行高效的口头交流, 人们很少明确地陈述不言而喻的场景信息, 这就需要结合常识来理解对话。为了模仿这一点, 文献[71-73]尝试利用外部知识来提高机器阅读性能, 但仍然存在不足。首先, 由于知识库中知识的稀疏性, 如何构建全面的知识库对阅读理解模型性能的提高是关键的。其次, 如何快速准确地从知识库中检索到相关的外部知识也影响着模型的效率。最后, 如何将结构化存储的知识与非结构化的篇章和问题进行有效地融合仍然是一个持续的挑战。

4) 答案生成技术。为了更接近现实场景, MS MARCO[4]和 DuReader[5]等数据集被发布, 它们的答案是通过众包抽象生成的自然语言文本, 然而目前的解决模型几乎仍然采用的是基于指针网络的边界预测方法。随着文本生成技术[94-95]的发展, 文献[49-51]初步尝试了生成新单词和复制源文本词相结合的指针-生成器混合网络, 但是方法简单, 效果一般, 得到的答案文本十分生硬, 远远没达到自然通顺流畅的需求。如何让机器生成高质量的答案依然需要更加深入的研究。

6.2 应用与展望

随着各种新技术新形态的发展, 具有对抗性的不良有害信息也不断涌现出新变种, 仅仅依赖于网民举报和人工审核的传统监测方式难以解决海量内容的审查问题。近年来基于深度学习的算法在网络文本内容识别、分析、分类等任务中取得杰出表现, “人工智能+安全”也成为了各界解决网络内容治理

问题的主流思路。例如, 腾讯云天御、网易易顿等产品都使用了深度学习技术; 学术上, 汪等人[96]在传统文本内容安全识别技术的基础上, 提出了利用深度学习的融合识别模型来提升文本内容识别的性能。

传统关键词匹配方法由于忽略上下文语义, 会导致误报漏报等问题, 而 MRC 利用人工智能技术为计算机赋予了阅读、分析和归纳文本的能力, 能够结合上下文语义来挖掘文本的深层含义, 因此, 有助于进一步提高不良有害信息的识别准确率。相比于传统方法, MRC 可以解决“敏感词误判、谐音表达、同义句检测”等问题。首先, 在中文的行文中词与词之间没有明显的分界符, 仅仅靠敏感词识别会出现误判的现象, 而 MRC 中的注意力机制可以用来解决中文分词不明确导致的误报, 类似文献[97]使用共享的全连接自注意力机制, 能根据不同的标准进行分词。例如, 微博发文“黑夜总会过去, 白天总会到来”由于出现了“夜总会”这一敏感词而被屏蔽, 但通过注意力机制可以发现“黑夜”和“白天”之间有着对应关系, 因此前半句以“夜总会”分词出现的概率比较低。其次, 人们通过加工处理(如谐音表达)有意识地回避敏感字眼以达到传播垃圾敏感信息的目的, 使得基于关键词匹配的技术无法起效。这个问题可以借鉴 MRC 中完型填空的形式, 对原始词进行预测, 从而提高识别准确率。最后, 多样的表达方式可以传达相同的语义, 增加了人工审核的工作量和工作难度, MRC 中的推理模型则可以解决相同语义的多样表达问题, 如文献[98-99]就是通过文本蕴含进行推理来实现句子复述检测的。总的来说, 机器阅读理解技术有利于提高有害信息识别的准确性, 并且在提高效率、解放人力成本方面起着重要作用。

为了让机器阅读理解技术更好的应用于网络内容治理, 未来还需要在以下方向取得突破。首先, 机器阅读理解模型可用的训练数据涉及的领域较少, 并且构建高质量的语料库必须明确具体的违规类别标准, 需要大量的人力来收集和标注样本, 未来如何高效地收集更多领域的高质量训练数据, 是机器阅读理解更好地应用于网络内容治理的关键。其次, 不良有害信息的类型和尺度因不同的平台业务而异, 未来如何实现每个平台之间的模型隔离, 以及如何批量处理 MRC 模型, 将是降低系统维护成本的重要因素。此外, 网络舆情表达了广大网民的观点和情感, 是客观信息和主观信息的结合, 但是当前机器阅读理解关注的都是对客观信息的推理分析, 让机器理解文字所表达的主观信息将会更加有效地监控网络舆情, 是非常值得关注的研究方向。

7 总结

我国的互联网用户人数近年来迅速增加, 网络内容受众非常广泛, 然而网络上的信息良莠不齐, 网络内容治理和网络舆情监管的难度也与日俱增。现阶段的文本筛选、文本分类等审核技术都是对浅层信息的处理, 深层次的文本内容理解还需要依赖于机器阅读理解。

机器阅读理解是自然语言处理领域的一项核心任务。机器阅读理解经历了三个发展阶段, 有四种基本的任务形式, 由于基本形式的数据与真实应用相差甚远, 因此出现了四种新的研究趋势, 也带来了更多新挑战。本文是关于神经机器阅读模型的综述, 从基本任务定义和新的研究趋势入手, 介绍了各种相关的神经机器阅读模型; 对各种注意力机制和推理模型的创新进行了总结和分类; 分析了目前机器阅读理解研究存在的问题同时对未来的研究方向进行了展望。

回顾近年来在机器阅读理解领域取得的主要成果, 更多高质量的数据、更优秀的架构、更强大的算力, 确实能够给机器理解文本带来更好的结果。机器阅读理解方向的发展非常快, 因此, 很难将所有新提出的研究工作都包括在本次调研中。例如近两年提出的 BERT[74]等预训练语言模型, 由于其能够很好地利用大量无标注文本, 并借助性能出色的深度学习架构, 从而能够学习到更深层的语义和更复杂的语言现象, 已经在机器阅读理解任务上取得了最先进的效果, 并且相比于传统深度学习方法提高了一大截, 一次又一次地刷新了各种机器阅读理解的榜单。在实际应用层面, 神经机器阅读理解已经逐渐在智能客服、问答机器人、智能法律文案分析、智能搜索、网络舆情监测等应用中崭露头角, 减少人力和管理成本的同时提高了服务效率。

参考文献

[1] Hermann K M, Kočiský T, Grefenstette E, et al. Teaching Machines to Read and Comprehend[J]. *Advances in Neural Information Processing Systems*, 2015, 2015-January: 1693-1701.

[2] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100, 000+ questions for machine comprehension of text[C]. *The 2016 Conference on Empirical Methods in Natural Language Processing*, 2016: 2383-2392.

[3] Rajpurkar P, Jia R, Liang P. Know what You don't know: Unanswerable questions for SQuAD[C]. *The 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Pa-*

pers), 2018: 784-789.

[4] Tri Nguyen, Mir Rosenberg, Xia Song, et al. MS MARCO: A human generated machine reading comprehension dataset. *In Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems*, 2016.

[5] He W, Liu K, Liu J, et al. DuReader: a Chinese machine reading comprehension dataset from real-world applications[C]. *The Workshop on Machine Reading for Question Answering*, 2018: 37-46.

[6] Li Z J, Wang C B. Survey on Deep-Learning-Based Machine Reading Comprehension[J]. *Computer Science*, 2019, 46(7): 7-12.
(李舟军, 王昌宝. 基于深度学习的机器阅读理解综述[J]. *计算机科学*, 2019, 46(7): 7-12.)

[7] Zhang C R, Qiu H P, Sun Y, et al. Review of Machine Reading Comprehension Based on Pre-Training Language Model[J]. *Computer Engineering and Applications*, 2020, 56(11): 17-25.
(张超然, 裘杭萍, 孙毅, 等. 基于预训练模型的机器阅读理解研究综述[J]. *计算机工程与应用*, 2020, 56(11): 17-25.)

[8] Lehnert W G. The Process of Question Answering[M]. London: Routledge, 2022.

[9] Hirschman L, Light M, Breck E, et al. Deep Read: a reading comprehension system[C]. *The 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999: 325-332.

[10] Riloff E, Thelen M. A Rule-Based Question Answering System for Reading Comprehension Tests[C]. *The 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems - Volume 6*, 2000: 13-19.

[11] Richardson M, Burges C J C, Renshaw E. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text[J]. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing*, *Proceedings of the Conference*, 2013: 193-203.

[12] Sachan M, Dubey K, Xing E, et al. Learning answer-entailing structures for machine comprehension[C]. *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*Volume 1*: *Long Papers*), 2015: 239-249.

[13] Narasimhan K, Barzilay R. Machine comprehension with discourse relations[C]. *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*Volume 1*: *Long Papers*), 2015: 1253-1262.

[14] Wang H, Bansal M, Gimpel K, et al. Machine comprehension with syntax, frames, and semantics[C]. *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*Volume 2*: *Short Papers*), 2015: 700-706.

[15] www.cnn.com

[16] www.dailymail.co.uk

[17] Chen D Q, Bolton J, Manning C D. A thorough examination of the CNN/daily mail reading comprehension task[C]. *The 54th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 2016: 2358-2367.

[18] Kadlec R, Schmid M, Bajgar O, et al. Text understanding with the attention sum reader network[C]. *The 54th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 2016: 908-918.

[19] Yiming Cui, Ting Liu, Zhipeng Chen, et al. Consensus attention-based neural networks for chinese reading comprehension[C]. *In Proceedings of the 26th International Conference on Computational Linguistics*, 2016: 1777-1786.

[20] Cui Y M, Chen Z P, Wei S, et al. Attention-over-attention neural networks for reading comprehension[C]. *The 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 2017: 593-602.

[21] Wang S H, Jiang J. Learning natural language inference with LSTM[C]. *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, 2016: 1442-1451.

[22] Vinyals O, Fortunato M, Jaitly N. Pointer Networks[J]. *Advances in Neural Information Processing Systems*, 2015, 2015-January: 2692-2700.

[23] Shuohang Wang and Jing Jiang. Machine comprehension using match-lstm and answer pointer[C]. *In the 5th International Conference on Learning Representations*, Poster, 2017.

[24] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering[C]. *In the 5th International Conference on Learning Representations*, Poster, 2017.

[25] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, et al. Bidirectional attention flow for machine comprehension[C]. *In the 5th International Conference on Learning Representations*, Poster, 2017.

[26] Wang W, Yan M, Wu C. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering[C]. *The 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 2018: 1705-1714.

[27] Wang W H, Yang N, Wei F R, et al. Gated self-matching networks for reading comprehension and question answering[C]. *The 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 2017: 189-198.

[28] Hill F, Bordes A, Chopra S, et al. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations[EB/OL]. 2015: arXiv: 1511.02301. https://arxiv.org/abs/1511.02301

[29] Lai G K, Xie Q Z, Liu H X, et al. RACE: large-scale ReAding comprehension dataset from examinations[C]. *The 2017 Conference on Empirical Methods in Natural Language Processing*, 2017: 785-794.

[30] Trischler A, Wang T, Yuan X D, et al. NewsQA: A machine comprehension dataset[C]. *The 2nd Workshop on Representation Learning for NLP*, 2017: 191-200.

[31] Mandar Joshi, Eunsol Choi, Daniel Weld, et al. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension[C]. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Volume 1: Long Papers,

2017: 1601-1611.

[32] Dunn M, Sagun L, Higgins M, et al. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine[EB/OL]. 2017: arXiv: 1704.05179. https://arxiv.org/abs/1704.05179.

[33] Kočiský T, Schwarz J, Blunsom P, et al. The NarrativeQA Reading Comprehension Challenge[J]. *Transactions of the Association for Computational Linguistics*, 2018, 6: 317-328.

[34] Sun C, Shrivastava A, Singh S, et al. Revisiting unreasonable effectiveness of data in deep learning era[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 843-852.

[35] Danqi Chen. Neural Reading Comprehension and Beyond. PhD thesis, Stanford University, 2018.

[36] Xie Q Z, Lai G K, Dai Z H, et al. Large-Scale Cloze Test Dataset Designed by Teachers[EB/OL]. 2017: arXiv: 1711.03225. https://arxiv.org/abs/1711.03225.

[37] Suster S, Daelemans W. CliCR: a dataset of clinical case reports for machine reading comprehension[C]. *The 2018 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *Volume 1* (*Long Papers*), 2018: 1551-1563.

[38] 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.

[39] Zhu H C, Wei F R, Qin B, et al. Hierarchical Attention Flow for Multiple-Choice Reading Comprehension[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, 32(1): 6077-6085.

[40] Xu Y C, Liu J J, Gao J F, et al. Dynamic Fusion Networks for Machine Reading Comprehension[EB/OL]. 2017: arXiv: 1711.04964. https://arxiv.org/abs/1711.04964.

[41] Wang S H, Yu M, Jiang J, et al. A Co-matching model for multi-choice reading comprehension[C]. *The 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers*), 2018: 746-751.

[42] Zhang S L, Zhao H, Wu Y W, et al. DCMN+: Dual Co-Matching Network for Multi-Choice Reading Comprehension[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(5): 9563-9570.

[43] Chen Z P, Cui Y M, Ma W T, et al. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 6276-6283.

[44] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009: 248-255.

[45] Cui Y M, Liu T, Che W X, et al. A span-extraction dataset for Chinese machine reading comprehension[C]. *The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019: 5882-5888.

[46] Xiong C M, Zhong V, Socher R. DCN+: Mixed Objective and Deep Residual Coattention for Question Answering[EB/OL]. 2017: arXiv: 1711.00106. https://arxiv.org/abs/1711.00106.

[47] Papineni K, Roukos S, Ward T, et al. BLEU: A Method for Automatic Evaluation of Machine Translation[C]. *The 40th Annual Meeting on Association for Computational Linguistics*, 2002: 311-318.

[48] Lin C Y. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out, 2004.

[49] Mitra R. A Generative Approach to Question Answering[EB/OL]. 2017: arXiv: 1711.06238. https://arxiv.org/abs/1711.06238.

[50] Tan C Q, Wei F R, Yang N, et al. S-Net: From Answer Extraction to Answer Generation for Machine Reading Comprehension[EB/OL]. 2017: arXiv: 1706.04815. https://arxiv.org/abs/1706.04815.

[51] Nishida K, Saito I, Nishida K, et al. Multi-style generative reading comprehension[C]. *The 57th Annual Meeting of the Association for Computational Linguistics*, 2019: 2273-2284.

[52] Ostermann S, Modi A, Roth M, et al. MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge[EB/OL]. 2018: arXiv: 1803.05223. https://arxiv.org/abs/1803.05223.

[53] Ostermann S, Roth M, Pinkal M. MCScript2.0: A machine comprehension corpus focused on script events and participants[C]. *The Eighth Joint Conference on Lexical and Computational Semantics*, 2019: 103-117.

[54] Sun K, Yu D, Chen J S, et al. DREAM: A Challenge Data Set and Models for Dialogue-Based Reading Comprehension[J]. *Transactions of the Association for Computational Linguistics*, 2019, 7: 217-231.

[55] 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.

[56] Yang Z L, Qi P, Zhang S Z, et al. HotpotQA: A dataset for diverse, explainable multi-hop question answering[C]. *The 2018 Conference on Empirical Methods in Natural Language Processing*, 2018: 2369-2380.

[57] Reddy S, Chen D Q, Manning C D. CoQA: A Conversational Question Answering Challenge[J]. *Transactions of the Association for Computational Linguistics*, 2019, 7: 249-266.

[58] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems[C]. *The 2017 Conference on Empirical Methods in Natural Language Processing*, 2017: 2021-2031.

[59] Clark C, Gardner M. Simple and effective multi-paragraph reading comprehension[C]. *The 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 2018: 845-855.

[60] Sun F, Li L Y, Qiu X P, et al. U-Net: Machine Reading Comprehension with Unanswerable Questions[EB/OL]. 2018: arXiv: 1810.06638. https://arxiv.org/abs/1810.06638.

[61] Hu M H, Wei F R, Peng Y X, et al. Read + Verify: Machine Reading Comprehension with Unanswerable Questions[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 6529-6537.

[62] Wang Y Z, Liu K, Liu J, et al. Multi-passage machine reading comprehension with cross-passage answer verification[C]. *The*

56th Annual Meeting of the Association for Computational Linguistics (*Volume 1*: *Long Papers*), 2018: 1918-1927.

[63] Liu J H, Wei W, Sun M S, et al. A multi-answer multi-task framework for real-world machine reading comprehension[C]. *The 2018 Conference on Empirical Methods in Natural Language Processing*, 2018: 2109-2118.

[64] Yan M, Xia J N, Wu C, et al. A Deep Cascade Model for Multi-Document Reading Comprehension[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 7354-7361.

[65] Hu M H, Peng Y X, Huang Z, et al. Retrieve, read, rerank: Towards end-to-end multi-document reading comprehension[C]. *The 57th Annual Meeting of the Association for Computational Linguistics*, 2019: 2285-2295.

[66] Chen D Q, Fisch A, Weston J, et al. Reading wikipedia to answer open-domain questions[C]. *The 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 2017: 1870-1879.

[67] Wang S H, Yu M, Guo X X, et al. R^3: Reinforced Ranker-Reader for Open-Domain Question Answering[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, 32(1): 5981-5988.

[68] Wang S H, Yu M, Guo X X, et al. R^3: Reinforced Ranker-Reader for Open-Domain Question Answering[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, 32(1).

[69] Lin Y K, Ji H Z, Liu Z Y, et al. Denoising distantly supervised open-domain question answering[C]. *The 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 2018: 1736-1745.

[70] Pang L, Lan Y Y, Guo J F, et al. HAS-QA: Hierarchical Answer Spans Model for Open-Domain Question Answering[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 6875-6882.

[71] Long T, Bengio E, Lowe R, et al. World Knowledge for Reading Comprehension: Rare Entity Prediction with Hierarchical LSTMs Using External Descriptions[C]. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017: 825-834.

[72] Mihaylov T, Frank A. Knowledgeable reader: enhancing cloze-style reading comprehension with external commonsense knowledge[C]. *The 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 2018: 821-832.

[73] Yang A, Wang Q, Liu J, et al. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension[C]. *The 57th Annual Meeting of the Association for Computational Linguistics*, 2019: 2346-2357.

[74] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1: Long and Short Papers, 2019: 4171-4186.

[75] See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks[C]. *The 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 2017: 1073-1083.

[76] Chen Y, Wu L F, Zaki M J. GraphFlow: Exploiting Conversation Flow with Graph Neural Networks for Conversational Machine Comprehension[EB/OL]. 2019: arXiv: 1908.00059. https://arxiv.org/abs/1908.00059.

[77] Pengfei Zhu, Hai Zhao and Xiaoguang Li. Dual Multi-head Co-attention for Multi-choice Reading Comprehension. 2020: ArXiv Preprint ArXiv:2001.09415.

[78] Bahdanau D, Cho K H, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. *3rd International Conference on Learning Representations*, *ICLR 2015 - Conference Track Proceedings*, 2015.

[79] Huang H Y, Zhu C G, Shen Y L, et al. FusionNet: Fusing via Fully-Aware Attention with Application to Machine Comprehension[EB/OL]. 2017: arXiv: 1711.07341. https://arxiv.org/abs/1711.07341.

[80] Hu M H, Peng Y X, Huang Z, et al. Reinforced Mnemonic Reader for Machine Reading Comprehension[C]. *The 27th International Joint Conference on Artificial Intelligence*, 2018: 4099-4106.

[81] Zhuang Y M, Wang H D. Token-level dynamic self-attention network for multi-passage reading comprehension[C]. *The 57th Annual Meeting of the Association for Computational Linguistics*, 2019: 2252-2262.

[82] Trischler A, Ye Z, Yuan X D, et al. Natural language comprehension with the EpiReader[C]. *The 2016 Conference on Empirical Methods in Natural Language Processing*, 2016: 128-137.

[83] Sordoni A, Bachman P, Bengio Y. Iterative Alternating Neural Attention for Machine Reading[EB/OL]. 2016: arXiv: 1606.02245. https://arxiv.org/abs/1606.02245.

[84] Shen Y L, Huang P S, Gao J F, et al. ReasoNet: Learning to Stop Reading in Machine Comprehension[C]. *The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017: 1047-1055.

[85] Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, et al. Multi-step Retriever-Reader Interaction for Scalable Open-domain Question Answering[C]. *In Proceedings of 7th International Conference on Learning Representations*, Poster, 2019.

[86] Dehghani M, Azarbonyad H, Kamps J, et al. Learning to Transform, Combine, and Reason in Open-Domain Question Answering[C]. *The Twelfth ACM International Conference on Web Search and Data Mining*, 2019: 681-689.

[87] Dhingra B, Jin Q, Yang Z L, et al. Neural models for reasoning over multiple mentions using coreference[C]. *The 2018 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *Volume 2* (*Short Papers*), 2018: 42-48.

[88] Dhingra B, Liu H X, Yang Z L, et al. Gated-attention readers for text comprehension[C]. *The 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 2017: 1832-1846.

[89] Song L F, Wang Z G, Yu M, et al. Exploring Graph-Structured Passage Representation for Multi-Hop Reading Comprehension with Graph Neural Networks[EB/OL]. 2018: arXiv: 1809.02040.

https://arxiv.org/abs/1809.02040.

[90] de Cao N, Aziz W, Titov I. Question answering by reasoning across documents with graph convolutional networks[C]. *The 2019 Conference of the North*, 2019: 2306-2317.

[91] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling Relational Data with Graph Convolutional Networks[M]. The Semantic Web. Cham: Springer International Publishing, 2018: 593-607.

[92] Tu M, Wang G T, Huang J, et al. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs[C]. *The 57th Annual Meeting of the Association for Computational Linguistics*, 2019: 2704-2713.

[93] Kundu S, Khot T, Sabharwal A, et al. Exploiting explicit paths for multi-hop reading comprehension[C]. *The 57th Annual Meeting of the Association for Computational Linguistics*, 2019: 2737-2747.

[94] Guu K, Hashimoto T B, Oren Y, et al. Generating Sentences by Editing Prototypes[J]. *Transactions of the Association for Computational Linguistics*, 2018, 6: 437-450.

[95] Gupta A, Agarwal A, Singh P, et al. A Deep Generative Framework for Paraphrase Generation[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, 32(1): 5149-5156.

[96] S.M. Wang, Z. Wang and H. Ren. Research on fusion Model based on deep learning for text content security enhancement. *Telecommunications Science*, 2020, no.5: 25-30.
(汪少敏, 王铮, 任华. 利用深度学习融合模型提升文本内容安全的研究. *电信科学*, 2020 年 05 期:25-30.)

[97] Qiu X P, Pei H Z, Yan H, et al. A Concise Model for Multi-Criteria Chinese Word Segmentation with Transformer Encoder[EB/OL]. 2019: arXiv: 1906.12035. https://arxiv.org/abs/1906.12035.

[98] I. Dagan, and O. Glickman. Probabilistic textual entailment: Generic applied modeling of language variability[C]. *In Proceedings of PASCAL Workshop on Learning Methods for Text Understanding and Mining*, 2004.

[99] Rus V, McCarthy P M, Lintean M C, et al. Paraphrase Identification with Lexico-Syntactic Graph Subsumption[J]. *Proceedings of the 21th International Florida Artificial Intelligence Research Society Conference*, *FLAIRS-21*, 2008: 201-206.

骆丹 于 2016 年在中国科学技术大学信息安全专业获得学士学位。现在中国科学院信息工程研究所攻读博士学位。研究领域为自然语言处理。研究兴趣包括: 深度学习、机器阅读理解。Email: luodan@iie.ac.cn

张鹏 于 2013 年在中国科学院计算技术研究所计算机系统结构专业获得博士学位。现在中国科学院信息工程研究所任副研究员。研究领域为数据挖掘和网络空间安全。研究兴趣包括: 流量分析和数据挖掘。Email: pengzhang@iie.ac.cn