小 型 微 型 计 算 机 系 统 2020 年 3 月 第 3 期
Journal of Chinese Computer Systems Vol. 41 No. 3 2020

机器阅读理解的技术研究综述

徐霄玲,郑建立,尹梓名

(上海理工大学 医疗器械与食品学院,上海 200093)
E-mail:172702146@st.usst.edu.cn

摘 要: 机器阅读理解(MRC,Machine Reading Comprehension)是自然语言处理领域一个重要的研究方向.相关模型研究在直接提取篇章内容作为答案方面已经有了较大进展;现阶段研究重点是:在获取关键信息的基础上,如何整合外部知识,为人们提供更准确、更符合人类语言习惯的答案.本文对近几年机器阅读理解研究进展从四个方面进行综述,首先介绍了该任务构成要素和发展情况;其次梳理了四种类型数据集在数量、内容、难度上的变化;然后对预训练模型、注意力机制、记忆网络等方法进行介绍,比较了各个模型在不同类型数据集上的表现;最后,在上述内容基础上,对现有数据集的局限性、模型间的依赖性、未来研究热点等多方面提出思考.

关 键 词: 机器阅读理解;深度学习;预训练模型;注意力机制;记忆网络
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2020)03-0464-07

Overview of Technical Studies on Machine Reading Comprehension

XU Xiao-ling,ZHENG Jian-li,YIN Zi-ming
(School of Medical Instrument and Food Engineering,University of Shanghai for Science and Technology,Shanghai 200093,China)

Abstract: Machine Reading Comprehension(MRC) is an important research direction in the field of natural language processing. Relevant model research has made great progress in extracting text content directly as an answer. How to integrate external knowledge to get more accurate and natural answers on acquired key information is the focus of current research. In this paper, the researches on MRC in recent years are summarized from four aspects. Firstly, the elements and development of the task are introduced. Secondly, the changes in quantity, content and difficulty of four types of datasets are sorted out. Then the methods of pre-training model, attention mechanism and memory network are introduced, and the performance of different datasets with the models is compared. In the end, the limitations of existing data sets, dependence between models, future research hotspots and other aspects are put forward.

Key words: machine reading comprehension; pre-trained model; attention mechanism; memory network; deep learning

1 引 言

机器阅读理解(MRC,Machine Reading Comprehension)是自然语言处理的长期目标,是人工智能向前迈进的关键一步.互联网日益普及,深度学习等人工智能技术蓬勃发展,人们在图像识别、语音识别、围棋 AI 等领域已经使计算机达到接近人类甚至超越人类的水平.于是,人们开始往更为复杂的机器阅读理解领域进行探索.机器阅读理解是为了培养计算机对自然文本理解的能力,让其能像人类一样对文本进行阅读、推理,也就是意味着计算机在接受自然语言输入后能够给出正确的反馈[1].此概念在 1972 年首先[2]被提出.经过几十年的变化发展,已经由最初依据规则和词性、依存句法、语义角色等传统特征,演变为基于大数据和深度学习进行阅读推理.本文将从其具体任务,数据集和关键技术三方面,对机器阅读理解做出进一步阐述.

2 机器阅读理解任务概述

机器阅读理解实际上是由自然语言理解所衍生的子任务,用以衡量计算机"理解"自然语言所达到的程度.首先由 Hirschmann[3]等人提出利用文本阅读并通过回答问题的形式评估机器阅读理解,此种评估方式延续至今.通常情况下,机器阅读理解任务主要由 Document(需要机器阅读的篇章)、Question(需要机器回答的问题)、Answer(机器阅读理解的答案)三个要素构成.根据任务的不同,Answer 可能是篇章中的单个实体或者是篇章中的片段,也可能是机器生成的句子.当任务是阅读选择题时,在上述基础上需要增加 Candidate(候选答案)要素,Answer 来自于候选答案.近年来,在篇章数据集上学者们做了大量工作,使阅读理解更加贴近真实应用场景:内容上,由虚构故事向真实问答靠拢;回答方法上,由单纯依靠篇章回答向依赖外部知识推理发展;数据量上,从以前的几百到现在动辄上万.数据集的具体比较详见第 3 节.

机器阅读理解虽然在认知智能领域是一个极具挑战的任务,但却有着较为悠久的历史.最初由 Terry Winograd 提出构想[2],认为语法、语义和推理是实现阅读理解的三大要素.1999 年,出现首个自动阅读理解测试系统 Deep Read[3],该系统以故事为基础衡量阅读理解任务,利用词袋模型 BOW 和

收稿日期:2019-02-21 收修改稿日期:2019-07-31 基金项目:国家自然科学基金项目(81801797)资助. 作者简介:**徐霄玲**,女,1994 年生,硕士研究生,研究方向为自然语言处理;**郑建立**,男,1965 年生,博士,副教授,研究方向为医学信息集成;**尹梓名**,男,1986 年生,博士,讲师,研究方向为医学人工智能、医疗数据挖掘.

人工编写的规则进行模式匹配,达到了 40% 的正确率. 考虑到阅读理解需要大量常识,Schubert[4] 等人在 2000 年率先提出一个基于情节逻辑的叙事理解框架,情节逻辑被用于语义表示和外部知识表示. 总的说来,机器阅读理解早期发展速度缓慢,大量依靠手工提取的语法特征以及三元组信息,具有耗时长、鲁棒性差等缺点. 直到 Hermann 等人[5] 提出使用神经网络模型,该领域近年来才开始逐步发展起来. 其提出的 Deep LSTM Reader、Attentive Reader 和 Impatient Reader 三种神经网络模型,奠定了机器阅读领域的方法基础. 在此之后,Match-LSTM[6]、BiDAF[7]、Dynamic Coattention Networks[8] 等大量优秀模型频现,权威刷榜评测任务排名不断更新,为机器阅读理解提供了统一衡量标准,极大地促进了自然语言理解的发展.

3 MRC 数据集

机器阅读理解实际上是一个数据驱动型任务,因此数据集是其技术发展的基础. 无论是基于人工规则还是基于深度学习等热门手段,数据集的质量和难度都直接关系到模型的质量和实用性,每次不同形式数据集的出现都会带来模型的创新. 随着数据集规模增大和考查形式的变化,任务难度不断上升,对模型的要求也越来越高[9]. 到目前为止,已经出现很多经典英文数据集. 这两年,国内对阅读理解任务逐步重视,积极向国际靠拢,开放了 DuReader[10] 等中文数据集.

表 1 各个数据集基本统计信息比较

Table 1 Comparisons of basic statistical Information in datasets

数据集名称	语言类型	文章数量	问题数量
MCTest	英文	500	2000
RACE	英文	27933	97687
CNN	英文	92579	387420
Daily Mail	英文	119506	997467
CBT	英文	108	687343
BookTest	英文	–	14140825
PD&CFT	中文	28000	100000
SQuAD	英文	536	107785
NewsQA	英文	12744	120000
MS MARCO	英文	8841823	1010916
DuReader	中文	1000000	200000
NarrativeQA	英文	1572	46765

3.1 选择型数据集

选择题能有效避免模棱两可的答案,因此于 2013 年微软推出 MCTest[11]. MCTest 是一个面向开放领域的数据集,文章内容是适合 7 岁孩子理解的童话故事,提问形式为四选一选择题,且问题选项基本来自于原文,这说明基于此数据集的 MRC 评估模型基本不需要推理能力. MCTest 虽然通过众包的方式反复检查校验以确保高质量,但由于其数据规模较小(仅包含了近 500 篇文章和 2000 个问题),无法满足神经网络等更加复杂的训练模型. 2017 年学界开放了 RACE 数据集[12]. RACE 同样利用选择题的方式评估 MRC 任务. 相较于 MCTest,它数据量上占绝对优势,详见表 1. RACE 数据来源于中国 12-18 岁中学生的英语考试试题,由语言专家出题,59.2% 的问题需要联系上下文进行推理,能更加真实地以人类标准衡量机器阅读理解的能力. 在 SemEval-2018 任务 11 发布了基于常识的阅读理解[13],要求模型引入外部知识,从两个候选答案中选出一个作为正确答案.

3.2 填空型数据集

填空就是要求读者补充句子中缺失的词语[14]. 以填空形式构造问题,数量上可以任意扩充. CNN/Daily Mail[5] 率先解决了 MRC 领域数据量不足的问题. Hermann 等人从美国有线电视新闻网和每日邮报网中收集了近 100 万新闻数据,利用实体检测和匿名化算法,将新闻中概括性语句转换为 < 文章(c),问题(q),答案(a) > 三元组. 文章中的实体用随机数字代替,模型利用数字回答相应问题,有利于帮助研究者注重语义关系. CBT[15] 和 BookTest[16] 等也是填空型数据集. 两者任务类似,都是从书中抽取连续 21 个句子,前 20 句子作为文章,预测第 21 句中缺失的词. 但是 BT 数据规模更大,将近是 CBT 的 60 倍,更能满足复杂深度学习模型的数据需求. 哈尔滨工业大学讯飞联合实验室于 2016 年 7 月提出首个中文填空型阅读理解数据集 PD&CFT[17],增加了该领域语言的多样性,促进了中文阅读理解的发展.

3.3 篇章片段型数据集

篇章片段数据集指的是:在该数据集中,问题的答案不再是单一实体,而是文章中的片段(span). 既可以是单一片段,也可以是多个片段的组合,答案类型更加丰富. 由于答案的特殊性,因此多采用 F1 值、EM(准确匹配)、Bleu[18] 和 Rouge[19] 等作为衡量预测值和真实值重叠程度的指标.

SQuAD[20] 和 NewsQA[21] 是篇章片段数据集的代表,数据分别来自于维基百科和 CNN 新闻. 目前,SQuAD 数据集已经成为权威刷榜评测任务,且到发文为止在 SQuAD1.1 数据集中,机器表现已经超越人类. 由于 SQuAD1.1 数据主要集中在可回答的问题,因此斯坦福在其基础上增加了 50000 个不可回答问题,提出 SQuAD2.0[22],进一步提升了数据集难度. 2018 年第二届“讯飞杯”在其评测任务中发布了首个人工标注的中文篇章片段抽取型阅读理解数据集,填补了中文在这方面的空白.

3.4 多任务型数据集

最近研究表明,现有机器阅读理解模型,虽然能在大多数数据集上表现良好,但却无法真正实现机器理解. 究其原因,其推理能力十分有限,与人类存在较大差距[23],因此构建高难度的真实世界数据集十分迫切. MS MARCO[24] 和 DuReader[10] 数据分别取自必应搜索引擎,百度知道和百度搜索. 其中,DuReader 是迄今为止最大的中文 MRC 数据集. 上述两个数据集问题都取自真实世界的问答和搜索数据,答案不再能直接从原文获取,需要根据多篇文章进行推理得到答案. 其中,问题的答案都是由人工构建而成. 除此之外,Kočiský T 等人认为现在数据集提问过于浅显,单纯依靠模式匹配就能回答大多数问题. 为了促进阅读理解的发展,其提供了一个全新的数据集——NarrativeQA[25]. 该数据集收集了 1572 个故事,提问更加深入,要求模型必须通篇阅读并且联系上下文才能得到答案,其回答问题的模式跟人类更加贴近.

4 机器理解方法分析与研究

解决机器阅读理解问题需要关注以下三个问题:

1)问题和文档表示:将自然语言文本转换为计算机能够理解的形式;

2)检索上下文:联系上下文并适当推理,检索出文档中与问题最相关的文章片段;

3)获取答案:对检索出的文章片段进行归纳总结,得到答案.

用于解决机器阅读理解问题方法有传统方法和深度学习方法.传统方法更多地是在句子粒度上回答问题.将问题和文档提取特征后表示成矩阵,或利用人工规则,对问题 Q 的每个候选答案句应用相应类型规则集中的所有规则,累计计算得分,总得分最高者为问题 Q 的答案句[26];或把阅读理解当成分类任务,根据已经得到的特征,利用 SVM 等传统机器学习算法,得到答案 A[27].传统方法核心是特征抽取,包括抽取浅层特征和深层语义特征.目前被认为有效的特征主要有依存句法、词频共现、语篇关系等.虽然传统方法能在一些数据集上取得较好结果,但是由于特征需要专家根据数据集制定,鲁棒性差;再加之,只能从现有文本中提取特征,不能对文本进行推理,因此无法真正解决机器理解问题.

如今随着数据量几何级增长,硬件计算能力不断增强,深度学习方法被广泛运用到词粒度的机器阅读理解任务中.深度学习的最大优势在于能够通过通用的端到端的过程学习数据的特征,自动获取到数据的高层次表示,而不依赖于人工设计特征[28].用于 MRC 任务的深度学习模型基本包含嵌入层、编码层、语义交互层和答案抽取层.嵌入层将文章和问题映射成包含相关文本信息的向量表示,便于计算机理解;编码层利用 RNN、LSTM 等神经网络对文章和问题编码,得到上下文语义信息;匹配层根据将上述文章和问题编码信息进行融合匹配,最终得到混合两者语义的交互向量,这是整个模型中最重要的部分;答案预测层,根据语义交互向量,或选择答案,或抽取答案边界,或生成答案[29]

4.1 预训练模型

研究工作表明,预训练模型能有效提升大多数自然语言处理任务效果,MRC 任务同样也适用.预训练模型是前人为了解决类似问题所创造出来的模型,该模型参数能直接应用于当前任务中,既能弥补在语料不足的情况下构造复杂神经网络,又能在语料充足的情况下加快收敛速度.预训练模型的输出值,一般被应用于嵌入层,用以得到通用文本特征.

自然语言处理的所有任务本质上都是对向量的进一步使用,词作为语言表示中的基本单位,如何将其转化为向量是基础工作之一.通常情况下,词向量是预先训练好的,可以将其看成单层的预训练模型.在深度学习时代未到来以前,最为简单的词向量表示方法就是 one-hot 编码.但由于其无法解决维度灾难和语义表达问题,Rumelhart 等人[30]提出分布式词表示,使用稠密的低维向量表示每个词.研究者在此理论基础上,提出众多构建词向量的方法:Word2Vec[31]、Glove[32]和 FastText[33],这些方法被广泛应用于自然语言处理领域的各项任务中.

4.1.1 ELMo

近年来,出现三大预训练模型.ELMo(Embeddings from Language Models)[34]是其中之一,它利用双向 LSTM 提取到训练数据的单词特征、句法特征和语义特征.包含 N 个词的语料$(t_1,t_2,\cdots,t_N)$,前向 LSTM 根据已知词序列$(t_1,t_2,\cdots,t_{k-1})$,求词语 t_k 的概率,如公式(1)所示.后向 LSTM 则反之,根据已知词序列$(t_{k+1},t_{k+2},\cdots,t_N)$求概率,如公式(2)所示.ELMo 就是结合前向和后向 LSTM,求取联合似然函数的最大值,见公式(3),其中 Θ 表示神经网络中的各项参数.ELMo 用于 MRC 任务时,将模型的每层输出按照权重相乘得到词向量 $ELMo_k$,再将 $ELMo_k$ 与普通词向量 x_k 或者是隐层输出向量 h_k 拼接作为模型嵌入层输入.实验表明,ELMo 使当时最好的单模型[35]在 SQuAD 数据集上 F1 值提升了 1.7%.

$$P(t_1,t_2,\cdots,t_N)=\prod_{k=1}^{N}P(t_k|t_1,t_2,\cdots,t_{k-1}) \tag{1}$$

$$P(t_1,t_2,\cdots,t_N)=\prod_{k=1}^{N}P(t_k|t_{k+1},t_{k+2},\cdots,t_N) \tag{2}$$

$$L(t)=\sum_{k=1}^{N}(\log P(t_k|t_1,t_2,\cdots,t_{k-1};\Theta)+\log P(t_k|t_{k+1},t_{k+2},\cdots,t_N;\Theta)) \tag{3}$$

4.1.2 GPT

GPT(Generative Pre-Training)[1]是生成式预训练模型,是一种结合了无监督预训练和监督微调(supervised fine-tuning)的半监督方法.在预训练阶段,使用谷歌提出的单向 Transformer[36]作为特征提取器.Transformer 依靠自注意力机制抽取特征,能力强于 LSTM,被认为是 NLP 领域效果最好的长距离特征提取器.其他方面,仍然采用标准的语言模型训练目标函数,根据已知前 k-1 个词,求取当前词概率的最大似然估计:

$$L(t)=\sum_{k=1}^{N}(\log P(t_k|t_1,t_2,\cdots,t_{k-1};\Theta)) \tag{4}$$

式(3)和式(4)比较,不难发现,GPT 只依靠上文信息进行预测,而 ELMo 则结合了上下文信息.

Radford 等人提出 GPT-2[37],是 GPT 的升级版.GPT-2 与 GPT 最大的区别在于数据规模更大,模型层数更多,高达 48 层.GPT 应用于具体 NLP 任务时,要保证任务的网络结构与 GPT 一致,最简单的做法就是在 GPT 的最后一层 Transformer 层接入 softmax 作为任务输出层,通过训练对网络参数进行微调.实验表明,GPT 应用于 RACE 数据集,使最佳模型结果提高了 5.7%.

4.1.3 BERT

考虑到 GPT 模型的不足,谷歌团队提出 BERT(Bidirectional Encoder Representations from Transformers)预训练模型[38],得到了学术界广泛关注.BERT 预训练模型在流程上,与 GPT 保持了一致,都包含了预训练阶段和微调阶段.它与 GPT 最大的不同在于其使用双向 Transformer 完成了语言模型的训练,是 GPT 模型的进一步发展.同为双向语言模型,但其与 ELMo 训练的目标函数是不同的.ELMo 分别将 $P(t_k|t_1,t_2,\cdots,t_{k-1})$和$(t_{k+1},t_{k+2},\cdots,t_N)$作为目标函数,求两者结合后的最大似然概率.而 BERT 则以 $P(t_k|t_1,t_2,\cdots,t_{k-1},t_{k+1},\cdots,t_N)$为目标函数,真正意义上表征了上下文语境特征.BERT 提出后,在 11 个 NLP 任务中均取得了最好效果.在 MRC 任务中,单个 BERT 模型在 SQuAD 数据集中较最优模型 F1 值提高了 1.5%.

[1] https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

三大预训练模型结构差异详见图 1. Trm 代表 Transformer,$(E_1,E_2,\cdots,E_N)$为预训练模型输入,$(T,T_2,\cdots,T_N)$则表示输出. ELMo 使用双向 LSTM 的输出用于下游任务,而 GPT 使用单向 Transformer,BERT 使用双向 Transformer.

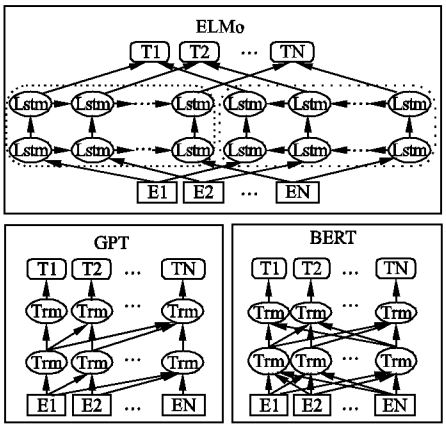

图 1 预训练模型结构对比图

Fig. 1 Comparison of pre-training model architectures

4.2 注意力机制(Attetion Mechanism)

阅读理解任务中,篇章往往较长,但与答案相关的内容只是其中的一小部分. 在传统方法中,通常利用循环网络将篇章编码成固定长度的中间语义向量,然后利用该向量指导每一步长输出. 此举既造成了信息过载,限制了模型效果,也降低了模型的运行效率. 为改变上述状况,学者们从机器翻译领域借鉴注意力机制[39],在 MRC 模型的语义交互层加入注意力机制,获取文章中与问题最相关的部分以提升效果.

4.2.1 基本概念

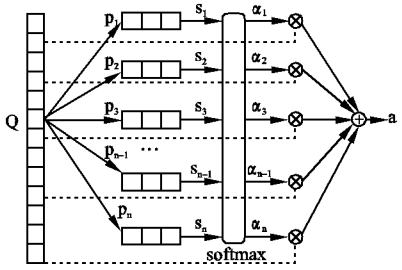

图 2 注意力机制流程

Fig. 2 Diagram of attetion mechanism

谷歌指出注意力机制就是加权求和[36]. 注意力实现机制如图 2 所示,分为两个步骤:计算注意力分布和加权平均. MRC 任务中,$P=[p_1,p_2,p_3,\cdots,p_{n-1},p_n]$代表篇章信息,向量 p_i 为篇章中的每个词的向量表示,i 表示词在文中的索引,$i\in[1,n]$;Q 为问题的向量表示. 通过打分函数 s 计算篇章中每个词与问题 Q 的相关性分数,然后经过 softmax 函数层,得到和为 1 的注意力分布 α_i,如公式(5)所示:

$$\alpha_i = softmax(s(p_i,Q)) = \frac{\exp(s(p_i,Q))}{\sum_{j=1}^{n}\exp(s(p_i,Q))} \tag{5}$$

其中函数 s 为注意力打分函数,可以是简单的计算,也可以是复杂的神经网络,常见的主要有点积运算、双线性模型、缩放点积模型和加性模型[2],分别见公式(6)-公式(9):

$$s(p_i,Q) = p_i^T Q \tag{6}$$

$$s(p_i,Q) = p_i^T W Q \tag{7}$$

$$s(p_i,Q) = \frac{p_i^T Q}{\sqrt{d}} \tag{8}$$

$$s(p_i,Q) = v^T\tanh(Wp_i + UQ) \tag{9}$$

其中 W、U、v 为神经网络模型可学习参数,d 表示篇章向量的维度. 不同的打分函数对模型的意义是不一样的,例如点积运算相较于加性模型,能更好的利用矩阵乘法,有利于训练效率地提高;双线性模型相较于点积运算,引入了非对称项,有利于信息提取等. 因此,需要根据数据和模型需要选择合适的打分函数.

注意力分布 α_i 获取了篇章中与问题强相关的部分,最后根据加权平均聚合所有篇章信息,强化相关信息,弱化甚至舍弃无关信息,用于最后的答案预测,见公式(10).

$$C = \sum_{i=1}^{n}\alpha_i p_i \tag{10}$$

4.2.2 相关模型

Attentive Reader[5]率先将注意力机制应用于机器阅读理解中,使用双向 LSTM 对文章进行编码,利用注意力机制求出每个词对应的权重,加权求和后最终表示出文章. 其中使用公式(9)作为计算注意力分布的打分函数. Stanford Attentive Reader 使用双线性项(公式(7))代替上述模型中 tanh 函数计算权重,在其基础上效果提升了 7% ~ 10%[23]. Impatient Reader[5]模型基本结构与 Attentive Reader 一致,但同时考虑了问题对文章权重的影响,因此每当从问题中获取一个词就迭代更新一次文章表示的权重. Attention Sum Reader[40]通过点积运算获取注意力权重,同时将相同词概率进行合并获取概率,得出答案. Gated-Attention[41]在 AS Reader 模型基础上,增加网络层数,并改用 Hadamard 乘法求解权重,提出新的注意力模型. Match-LSTM[42]则是第一个适用于 SQuAD 数据集的端到端神经网络模型.

之后,出现很多关于注意力的变体. 2016 年,科大讯飞提出 Attention-over-Attention Reader 层叠式注意力模型[43],在原有注意力上增加一层注意力,来描述每一个注意力的重要性. 并在其基础上衍生出交互式层叠注意力模型(Interactive AoA Reader)和融合式层叠注意力模型(Hybrid AoA Reader),在 SQuAD 数据集上均表现不俗. 针对选择题型机器阅读理解,朱海潮[44]等人提出 Hierarchical Attention Flow,通过使用词级别和句子级别注意力,将文章、问题和选项进行充分交互,在 RACE 数据集上取得优于基准模型的效果. DCN[45]利用 Co-attention 技术分别生成关于文档和问题的权重分布并结合,通过多次迭代得到答案;BiDAF[7]在交互层引入双向注意力机制 context-to-query 和 query-to-context;DFN[46]将一般模型中固定的注意力机制扩展到多策略注意力,使模型能根据问题类型动态选择出适宜的注意力机制;Reasonet[47]则将 Memory Network(见 4.3 节)和 attention 结合,动态决定阅读次数,直至能回答问题为止. 这些均为MRC任务中注意力的使用提供了新

[2] https://nndl.github.io/

思路. 具体模型在数据集上的实验结果详见表 2、表 3.

表 2 模型在 RACE 数据集上的正确率

Table 2 Accuracy on RACE datasets

模 型	RACE-M	RACE-H	RACE
Stanford AR	44.2	43.0	43.3
GA Reader	43.7	44.2	44.1
Hierarchical Attention[44]	45.0	46.4	46.0
DFN[46]	51.5	45.7	47.4
Human Performance	95.4	94.2	94.5

表 3 模型在 CNN/DailyMail 和 CBT 数据集的正确率

Table 3 Accuracy on CNN/DailyMail and CBT datasets

模 型	CNN		CBT-NE		Daily Mail	
	Val	Test	Val	Test	Val	Test
Attentive Reader[5]	61.6	63.0	70.5	69.0	–	–
Impatient Reader[5]	61.8	63.8	69.0	68.0	–	–
Stanford AR[23]	73.8	73.6	77.6	76.6	–	–
AS Reader[40]	68.6	69.5	75.0	73.9	73.8	68.6
GA Reader[41]	73.0	73.8	76.7	75.7	74.9	69.0
AoA Reader[43]	73.1	74.4	–	–	77.8	72.0
ReasoNet[47]	72.9	74.7	77.6	76.6	–	–
BiDAF[7]	76.3	76.9	80.3	79.6	–	–

4.3 记忆网络(Memory Network)

随着数据量不断增加,学者认为传统机器学习模型(如 RNN、LSTM 等)利用隐含状态记忆,容量太小,无法完整记录文本内容. 除了使用注意力机制提取与问题最相关的文章内容之外,他们提出一种可读写的外部记忆模块,与问题相关的信息保存在外部记忆中,需要时再进行读取. 并将其和推理组件联合训练,最终得到具有长期记忆推理能力的灵活记忆能力. MRC 任务中,不仅文章篇幅较长,而且还有可能需要添加先验知识,记忆网络的使用能有效改善网络容量不足、长距离依赖等问题.

4.3.1 基本概念

记忆网络的概念在 2014 年首次被提出[48]. 从某种程度上说,记忆网络是一个框架,包含输入模块、输出模块、记忆模块等. 学者可以根据自己的需要定制框架下的各个模块. 记忆网络常见模块构成如图 3 所示.

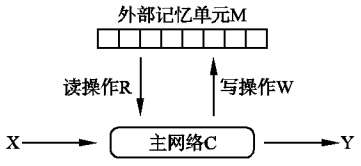

图 3 记忆网络基本组成模块

Fig. 3 Basic component module of memory network

输入模块 X:输入训练数据,包括篇章、问题等.

输出模块 Y:输出答案.

主网络 C:控制信息交互. 既包括与外界信息交互:根据输入 X,获取篇章内容 P 和问题 Q;得到预测答案后输出到 Y. 也包括与外部记忆单元交互,控制读写操作完成外部记忆单元的动态更新.

读操作 R:根据输入中的问题 Q、主网络在多次推理过程中生成的问题向量 q_t,从外部记忆单元中读取相应的信息.

写操作 W:根据主网络在多次推理过程中生成的问题 q_w、待写入的信息 a,更新外部记忆单元内的相关信息.

外部记忆单元 M:引入的外部记忆模块,用于存储信息. 存储形式可以为数组、栈、队列等,也就是说每个记忆单元都是拥有自己的地址的,读写操作都是根据寻址后完成的.

4.3.2 相关模型

Weston 等人[48]首先提出记忆网络的雏形,指出记忆网络由一个记忆模块 m 和四个组件 I(输入)、G(泛化)、O(输出)、R(回答)组成. 将上下文和问题输入之后,利用记忆模块进行存储,并能根据相关信息动态更新,最后找到与问题最相关的记忆作为答案并输出. 上述模型虽然能解决网络容量不足的问题,但模型每一层都需要监督,无法进行反向传播计算,这与现阶段端到端的模型思想相违背. 为了解决上述问题,End-to-End Memory Networks(MemN2N)[49]被提出. 它在满足基本组成模块的前提下,重新搭建模型框架,使用加权求和的方式得到输出向量,是一个端到端的反向传播记忆网络,同时支持多跳推理. Key-Value Memory Network[50]在端到端记忆网络的基础上,优化了网络结构,扩大了记忆规模,使其能更好地存储先验知识. 上述三个模型为记忆网络的发展奠定了理论基础,却一直没有运用到相关机器阅读理解数据集上,直到 MEMEN 模型的出现[51]. MEMEN 对篇章和问题采取多层次输入,包括字向量输入、词向量输入、词性输入和命名实体输入,充分融合文档和问题当中的信息,将其存储到记忆单元中. 同时使用一种新的分层注意力机制寻址记忆单元,并动态更新单元内容. 类似地还有 MAMCN[52],增加额外记忆单元,并利用 BiGRU 更新,尝试解决长距离依赖问题,实现跨文档预测答案.

表 4 模型在 SQuAD 测试数据集上的 EM 值和 F1 值

Table 4 Exact Match(EM) and F1 scores on SQuAD 1.1 test

模 型	EM	F1
Match-LSTM[42]	64.7	73.7
DCN[45]	66.2	75.9
BiDAF	68.0	77.3
ReasoNet	69.1	78.9
MEMEN[51]	70.9	80.4
MAMCN[52]	70.9	79.9
InteractiveAoA Reader	73.6	81.9
HybridAoA Reader	80.0	87.3
Bert[38]	87.4	93.2
Human Performance	82.3	91.2

从表 2-表 4 中,我们不难发现现阶段模型往往是针对特定类型数据集设计,大多数模型不具备迁移到其他类型数据集的能力. 即使迁移成功,模型也不是一成不变,相同模型在不同类型数据集上的效果也不同. 我们需要根据数据集特点对模型进行选择、设计和改进.

5 总 结

机器理解能力是机器从感知智能走向认知智能的关键,机器阅读理解近些年来取得了较快的发展. 在答案形式上,从

最初的选择题,变成单词填空,最终发展到篇章片段抽取或自主生成答案;在数据集内容上,从简单的孩童虚构故事,往依托常识、看重推理能力的真实世界人类问答发展;在关键技术上,由通过基于传统特征完成阅读理解,到如今使用深度学习技术并结合预训练模型、注意力机制、记忆网络等新型技术提升效果.近两年,注意力机制较记忆网络发展更为火热,出现很多变体.

对机器阅读理解未来发展有以下几点值得关注:

1)纵观现有阅读理解数据集,针对专业领域数据集较少,适用于通用领域的模型并不一定在特定领域有好的效果,因此,应该结合行业趋势,推出如金融、医疗领域的相关数据集.

2)从上述研究方法中,不难发现 attention 的设计与任务息息相关,如何根据任务设计合理的 attention 方法仍会是研究热点.

3)现阅读理解模型基本是在没有融合外部知识的情况下,直接从给定文档中抽取相关信息作为答案,这与人类阅读习惯有较大差异.因此,如何整合多数据源外部知识,并将其融入现有模型是值得我们关注的.

4)大多数相关中文模型依赖于英文模型,应该综合考虑中文和英文语言特点上的差异,构建更加适用于中文的模型.

5)将机器阅读理解技术与其他自然语言处理任务相结合,有利于促进自然语言处理技术整体发展.

References:

[1] Mostafazadeh N, Chambers N, He X, et al. A corpus and evaluation framework for deeper understanding of commonsense stories[J]. Computer Science, arXiv:1604.01696, 2016.

[2] Winograd T. Understanding natural language[J]. Cognitive Psychology, 1972, 3(1):1-191.

[3] Hirschman L, Light M, Breck E, et al. Deep read: a reading comprehension system[C]//Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, 1999:325-332.

[4] Schubert L K, Hwang C H. Episodic logic meets little red riding hood: a comprehensive, natural representation for language understanding[J]. Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language, 2000:111-174.

[5] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching machines to read and comprehend[C]//Advances in Neural Information Processing Systems, 2015:1693-1701.

[6] Wang S, Jiang J. Machine comprehension using mat-ch-lstm and answer pointer[J]. Computer Science, arXiv:1608.07905, 2016.

[7] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional attention flow for machine comprehension[J]. Computer Science, arXiv: 1611.01603, 2016.

[8] Xiong C, Zhong V, Socher R. Dynamic coattention networks for question answering[J]. Computer Science, arXiv: 1611.01604, 2016.

[9] Paperno D, Kruszewski G, Lazaridou A, et al. The LAMBADA dataset: word prediction requiring a broad discourse context[J]. Computer Science, arXiv:1606.06031, 2016.

[10] He W, Liu K, Lyu Y, et al. DuReader: a Chinese machine reading comprehension dataset from real-world applications[J]. Computer Science, arXiv:1711.05073, 2017.

[11] Richardson M, Burges C J C, Renshaw E. Mctest: a challenge dataset for the open-domain machine comprehension of text[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013:193-203.

[12] Lai G, Xie Q, Liu H, et al. Race: large-scale reading comprehension dataset from examinations[J]. Computer Science, arXiv: 1704.04683, 2017.

[13] Ostermann S, Roth M, Modi A, et al. SemEval-2018 Task 11: machine comprehension using commonsense knowledge[C]//Proceedings of the 12th International Workshop on Semantic Evaluation, 2018:747-757.

[14] Taylor W L. "Cloze procedure": a new tool for measuring readability[J]. Journalism Bulletin, 1953, 30(4):415-433.

[15] Hill F, Bordes A, Chopra S, et al. The goldilocks principle: reading children's books with explicit memory representations[J]. Computer Science, arXiv:1511.02301, 2015.

[16] Bajgar O, Kadlec R, Kleindienst J. Embracing data abundance: booktest dataset for reading comprehension[J]. Computer Science, arXiv:1610.00956, 2016.

[17] Cui Y, Liu T, Chen Z, et al. Consensus attention-based neural networks forchinese reading comprehension[J]. Computer Science, arXiv:1607.02250, 2016.

[18] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Aannual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002:311-318.

[19] Lin C Y. Rouge: a package for automatic evaluation of summaries[C]//Text Summarization Branches Out, 2004:74-81.

[20] Rajpurkar P, Zhang J, Lopyrev K, et al. Squad: 100,000 + questions for machine comprehension of text[J]. Computer Science, arXiv:1606.05250, 2016.

[21] Trischler A, Wang T, Yuan X, et al. Newsqa: a machine comprehension dataset[J]. Computer Science, arXiv:1611.09830, 2016.

[22] Rajpurkar P, Jia R, Liang P. Know what you don't know: unanswerable questions for SQuAD[J]. Computer Science, arXiv: 1806.03822, 2018.

[23] Chen D, Bolton J, Manning C D. A thorough examination of the cnn/daily mail reading comprehension task[J]. Computer Science, arXiv:1606.02858, 2016.

[24] Nguyen T, Rosenberg M, Song X, et al. MS MARCO: a human generated machine reading comprehension dataset[J]. Computer Science, arXiv:1611.09268, 2016.

[25] Kočiský T, Schwarz J, Blunsom P, et al. The narrativeqa reading comprehension challenge[J]. Transactions of the Association of Computational Linguistics, 2018, 6:317-328.

[26] Li Ji-hong, Yang Xing-li, Wang Rui-bo, et al. Research on rule based question answering for Chinese reading comprehension[J]. Journal of Chinese Information Processing, 2009, 23(4):3-9.

[27] Zhang Na. Research on question answering for Chinese reading comprehension based on rules[D]. Taiyuan: Shanxi University, 2008.

[28] Huang Li-wei, Jiang Bi-tao, Lv Shou-ye, et al. Survey on deep learning based recommender systems[J]. Chinese Journal of Computers, 2018, 41(7): 1619-1647.

[29] Yang Zhi-ming, Shi Ying-cheng, Wang Yong, et al. Reading comprehension model based on BiDAF and multidocument reordering[J]. Journal of Chinese Information Processing, 2018, 32(11): 117-127.

[30] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323: 533-536.

[31] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, arXiv: 1301.3781, 2013.

[32] Pennington J, Socher R, Manning C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1532-1543.

[33] Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

[34] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. Computer Science, arXiv: 1802.05365, 2018.

[35] Liu X, Shen Y, Duh K, et al. Stochastic answer networks for machine reading comprehension[J]. Computer Science, arXiv: 1712.03556, 2017.

[36] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017: 5998-6008.

[37] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[EB/OL]. OpenAI Blog, https://www.bibsonomy.org/bibtex/b926ece39c03cdf5499f6540cf63babd, 2019.

[38] Devlin J, Chang M W, Lee K, et al. Bert: pretraining of deep bidirectional transformers for language understanding[J]. Computer Science, arXiv: 1810.04805, 2018.

[39] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. Computer Science, arXiv: 1409.0473, 2014.

[40] Kadlec R, Schmid M, Bajgar O, et al. Text understanding with the attention sum reader network[J]. Computer Science, arXiv: 1603.01547, 2016.

[41] Dhingra B, Liu H, Yang Z, et al. Gated-attention readers for text comprehension[J]. Computer Science, arXiv: 1606.01549, 2016.

[42] Wang S, Jiang J. Machine comprehension using matchlstm and answer pointer[J]. Computer Science, arXiv: 1608.07905, 2016.

[43] Cui Y, Chen Z, Wei S, et al. Attention-over-attention neural networks for reading comprehension[J]. Computer Science, arXiv: 1607.04423, 2016.

[44] Zhu H, Wei F, Qin B, et al. Hierarchical attention flow for multiple-choice reading comprehension[C]//Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[45] Xiong C, Zhong V, Socher R. Dynamic coattention networks for question answering[J]. Computer Science, arXiv: 1611.01604, 2016.

[46] Xu Y, Liu J, Gao J, et al. Dynamic fusion networks for machine reading comprehension[J]. Computer Science, arXiv: 1711.04964, 2017.

[47] Shen Y, Huang P S, Gao J, et al. Reasonet: learning to stop reading in machine comprehension[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017: 1047-1055.

[48] Jason Weston, Sumit Chopra, Antoine Bordes. Memory networks[J]. Computer Science, arXiv: 1410.3916, 2014.

[49] Sukhbaatar S, Weston J, Fergus R. End-to-end memory networks[C]//Advances in Neural Information Processing Systems, 2015: 2440-2448.

[50] Miller A, Fisch A, Dodge J, et al. Key-value memory networks for directly reading documents[J]. Computer Science, arXiv: 1606.03126, 2016.

[51] Pan B, Li H, Zhao Z, et al. Memen: Multi-layer embedding with memory networks for machine comprehension[J]. Computer Science, arXiv: 1707.09098, 2017.

[52] Yu S, Indurthi S R, Back S, et al. A multistage memory augmented neural network for machine reading comprehension[C]//Proceedings of the Workshop on Machine Reading for Question Answering, 2018: 21-30.

附中文参考文献：

[26] 李济洪，杨杏丽，王瑞波，等. 基于规则的中文阅读理解问题回答技术研究[J]. 中文信息学报，2009，23(4)：3-9.

[27] 张 娜. 基于规则的阅读理解问题回答技术研究[D]. 太原：山西大学，2008.

[28] 黄立威，江碧涛，吕守业，等. 基于深度学习的推荐系统研究综述[J]. 计算机学报，2018，41(7)：1619-1647.

[29] 杨志明，时迎成，王 泳，等. 基于 BiDAF 多文档重排序的阅读理解模型[J]. 中文信息学报，2018，32(11)：117-127.