智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于深度学习的机器阅读理解研究综述

杜永萍, 赵以梁, 阎婧雅, 郭文阳

引用本文:
杜永萍,赵以梁,阎婧雅,郭文阳. 基于深度学习的机器阅读理解研究综述[J]. 智能系统学报, 2022, 17(6): 1074–1083.
DU Yongping,ZHAO Yiliang,YAN Jingya,GUO Wenyang. Survey of machine reading comprehension based on deep learning[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(6): 1074–1083.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202107024

DOI: 10.11992/tis.202107024
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20221017.1752.006.html

基于深度学习的机器阅读理解研究综述

杜永萍, 赵以梁, 阎婧雅, 郭文阳
(北京工业大学 信息学部, 北京 100124)

摘 要: 机器阅读理解任务在近年来备受关注, 它赋予计算机从文本数据中获取知识和回答问题的能力。如何让机器理解自然语言是人工智能领域长期存在的挑战之一, 近年来大规模高质量数据集的发布和深度学习技术的运用, 使得机器阅读理解取得了快速发展。基于神经网络的端到端的模型结构, 基于预训练语言模型以及推理技术的应用, 其性能在大规模评测数据集上有很大提升, 但距离真正的理解语言还有较大差距。本文对机器阅读理解任务的研究现状与发展趋势进行了综述, 主要包括任务划分、机器阅读理解模型与相关技术的分析, 特别是基于知识推理的机器阅读理解技术, 总结并讨论了该领域的发展趋势。

关键词: 机器阅读理解; 自然语言处理; 深度学习; 神经网络; 端到端模型; 知识推理; 预训练语言模型; 人工智能
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2022)06−1074−10

中文引用格式: 杜永萍, 赵以梁, 阎婧雅, 等. 基于深度学习的机器阅读理解研究综述 [J]. 智能系统学报, 2022, 17(6): 1074–1083.
英文引用格式: DU Yongping, ZHAO Yiliang, YAN Jingya, et al. Survey of machine reading comprehension based on deep learning[J]. CAAI transactions on intelligent systems, 2022, 17(6): 1074–1083.

Survey of machine reading comprehension based on deep learning

DU Yongping, ZHAO Yiliang, YAN Jingya, GUO Wenyang
(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: In recent years, there has been a great deal of interest in the task of machine reading comprehension. It enables computers to learn and answer questions based on text input. One of the long-standing challenges in the field of artificial intelligence is how to make machines understand natural language. In recent years, machine reading comprehension has advanced rapidly as a result of the large-scale release of high-quality data sets and the application of deep learning technology. The use of an end-to-end model structure based on neural networks, a pre-trained language model, and reasoning technology has greatly improved their performance on large-scale evaluation data sets. However, there is still a big gap in real language understanding. This paper summarizes the research status and development trend of machine reading comprehension tasks, including division of tasks, analysis of machine reading comprehension model and related technologies, particularly machine reading comprehension technology based on knowledge reasoning, and finally discusses the development trend in this field.

Keywords: machine reading comprehension; natural language processing; deep learning; neural network; end-to-end model; knowledge reasoning; pretrained language model; artificial intelligence

认知智能是人工智能发展的最高阶段, 其目标是让机器掌握人类的语言和知识体系, 并真正理解其内在逻辑, 这意味着机器开始具备分析和思考的能力。自然语言是认知科学的一项重要研究内容, 用自然语言与计算机进行通信, 意味着要使计算机能够理解自然语言文本的意义, 以自然语言理解为核心技术的自动问答、人机对话、聊天机器人已经成为产业界和学术界的关注热点。

自动问答是语言理解的重要应用领域, 特别是机器阅读理解, 赋予了计算机从文本数据中获取知识和回答问题的能力, 它是人工智能中一项挑战性的任务, 需要深度理解自然语言并具备一定推理能力。

近年来, 机器阅读理解领域的研究进入快速发展时期, 一方面得益于大规模高质量数据集的发

收稿日期: 2021−07−13. **网络出版日期**: 2022−10−18.
基金项目: 北京市自然科学基金项目(4212013); 国家语委信息化项目(YB135-89).
通信作者: 杜永萍. E-mail: ypdu@bjut.edu.cn.

布: 包括 Facebook Children's Books Test[1]、SQuAD[2] 以及 TriviaQA[3] 等高质量数据集; 另一方面, 基于深度学习技术的模型在获取上下文交互信息方面明显优于传统模型, 例如基于双向注意力机制的 BiDAF 模型[4]、Transformer[5] 和基于注意力机制的循环神经网络 R-Net[6]。中文问答任务上, 基于注意力机制的模型如 N-Reader[7] 在中文数据集 DuReader[8] 上取得了较好的成绩。

近期, 预训练模型[9-10] 与知识推理技术[11-12] 在复杂问答任务上取得了优异的表现, 特别在多跳问答任务中, 问题的答案需要从多个篇章中获取, 模型需要通过推理才能得出答案, 图神经网络在该类任务上具有较好的适用性, Ding 等[11] 使用认知图谱与图神经网络解决复杂数据集的推理任务并取得当时 SOTA 的结果。

1　机器阅读理解任务

机器阅读理解任务, 从输入信息的角度, 可分为两种类型: 基于多模态的阅读理解任务和基于文本的阅读理解任务。

基于多模态的阅读理解任务是指使机器能够对文本、图片以及视频等多种来源的信息进行学习, 该研究任务更贴近于人类对信息获取的综合感知的学习方式, 它是新兴的具有挑战性的研究方向。目前已有一些基于多模态的阅读理解任务的评测任务和数据集, 如 RecipeQA[13] 和 TQA[14] 等。

本文主要针对基于文本的阅读理解任务进行分析, 主要分为四类: 完形填空式任务、选择式任务、片段抽取式任务和自由作答式任务。

1)完形填空式任务: 对于给定的篇章P, 从P中删去词语A。任务要求机器学习到函数F, 从$Q=P-\{A\}$中对P中缺少的词语或实体进行补全, 即$A=F(Q)=F(P-\{A\})$。

完形填空式任务的难点在于, 机器需要从不完整的文本中学习上下文语义关系, 并且不仅需要对篇章所表达的内容进行理解, 还需要把握篇章的语言表达、词语运用的习惯, 从而正确地对被删去的内容进行预测。该任务代表性数据集有 CNN/Daily Mail[15]、Facebook Children's Books Test[1] 等。

2)选择式任务: 对于给定的篇章P和问题Q, 以及问题Q的候选答案集合$A=\{A_1,A_2,\cdots,A_n\}$, 要求机器学习到函数F, 根据P、Q、A从A中选择对Q回答正确的一项, 即$A_i=F(P,Q,A)$。

选择式任务的特点在于要求数据集提供问题的候选答案集合。机器在完成选择式任务时, 需要对篇章、问题、候选答案之间的语义关系进行理解和分析, 给出正确的判断。该任务的代表性数据集有 WikiHop[16]、CommonsenseQA[17] 等。

3)片段抽取式任务: 对于给定的篇章$P=\{w_1,w_2,\cdots,w_n\}$和问题Q, 机器学习到函数F, 根据对P和Q的理解, 从P中选取连续片段A作为Q的答案, 即$A=F(P,Q),A=\{w_i,w_{i+1},\cdots,w_j\},A\in P$。

片段抽取式任务的特点在于问题的答案可以在篇章中找到, 且答案可以是词语、实体或句子等形式。构建数据集时对问题的选取有一定要求, 该任务的代表性数据集有 SQuAD[2]、NewsQA[18] 等。

4)自由作答式任务: 对于给定的篇章P和问题Q, 机器学习到函数F, 根据对P和Q的理解得出答案A, 且A不一定在P中出现, 可以为任意形式, 即$A=F(P,Q)$。

自由作答式任务在答案的选取上最为灵活, 答案的形式也无限制, 且答案范围不局限于给定篇章。这类任务往往要求机器具有一定的分析、推理能力。该任务的代表性数据集有 DuReader[8]、DROP[19] 等。

2　机器阅读理解技术

2.1　基于端到端神经网络的机器阅读理解技术

传统的机器阅读理解方法通常是基于规则或者统计学规律, 但随着该任务数据集的规模和质量的提升, 深度学习方法表现出了良好的性能。如今机器阅读理解模型的构建大多采用双向循环神经网络对问题和篇章进行编码, 并在问题–篇章交互层中使用注意力机制。

机器阅读理解模型的输入通常为问题和篇章, 最终的输出是问题的答案。常见的基于深度学习的机器阅读理解模型主要包括 4 个层次: 词嵌入层、编码层、问题–篇章交互层以及答案预测层, 如图 1 所示。

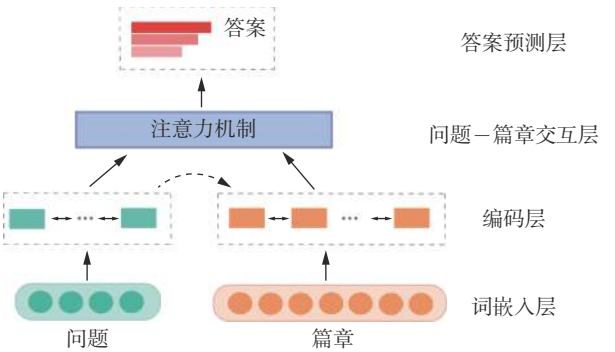


图 1　基于深度学习方法的机器阅读理解通用模型结构

Fig. 1　Generic architecture of machine reading comprehension model based on deep learning

1）词嵌入层：问题和篇章输入模型后，将输入的自然语言文字转换为定长向量。可以通过独热编码、分布式词向量表示等多种方式分别得到问题和篇章的嵌入表示。采用大规模语料库预训练得到的词表示会包含丰富的上下文信息。例如QANet[20]中使用预训练词表GloVe[21]作为词的初始化表示，为后续模型正确预测答案提供支撑。

2）编码层：词嵌入层的输出作为编码层的输入，分别对问题和篇章进行建模。一些典型的深度神经网络，例如循环神经网络，具有能够处理时间序列预测问题的特性，它通常被应用在编码层来挖掘问题和篇章的上下文信息。R-Net[6]采用多层的双向循环神经网络构建模型，并利用自注意力机制进一步捕获更加丰富的上下文信息。循环神经网络的优点是隐藏层的神经元之间可以进行交互，使得信息具有传递性。Attentive Reader[15]中的编码层部分由双向循环神经网络正向和反向的输出拼接得到篇章中第t个位置词的表示，并计算该位置词的权重。

3）问题–篇章交互层：问题和篇章之间的关联对答案的预测有着重要的作用。注意力机制被广泛应用于问题–篇章交互层中，包括单向注意力机制、双向注意力机制以及自注意力机制，用于增强与问题相关的篇章部分的表示。如图 2 所示，将问题$\boldsymbol{Q}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4,\boldsymbol{x}_5\}$融入到篇章$C$中，若要得到$C$中的词$y$的表示，首先计算$\boldsymbol{Q}$中每个词的权重$w_1,w_2,w_3,w_4,w_5=\text{softmax}(\boldsymbol{Q}^{\mathrm{T}},y)$，由$y$与$Q$中每个词点乘并使用行向softmax进行归一化得到；然后对$\boldsymbol{Q}$中每个词进行加权求和得到融入问题信息的词y的表示$\hat{\boldsymbol{y}}$，即$\hat{\boldsymbol{y}}=w_1\boldsymbol{x}_1+w_2\boldsymbol{x}_2+w_3\boldsymbol{x}_3+w_4\boldsymbol{x}_4+w_5\boldsymbol{x}_5$。以此类推，计算得到篇章$C$中每个词的新的表示，记作$A(\boldsymbol{Q},C)$。

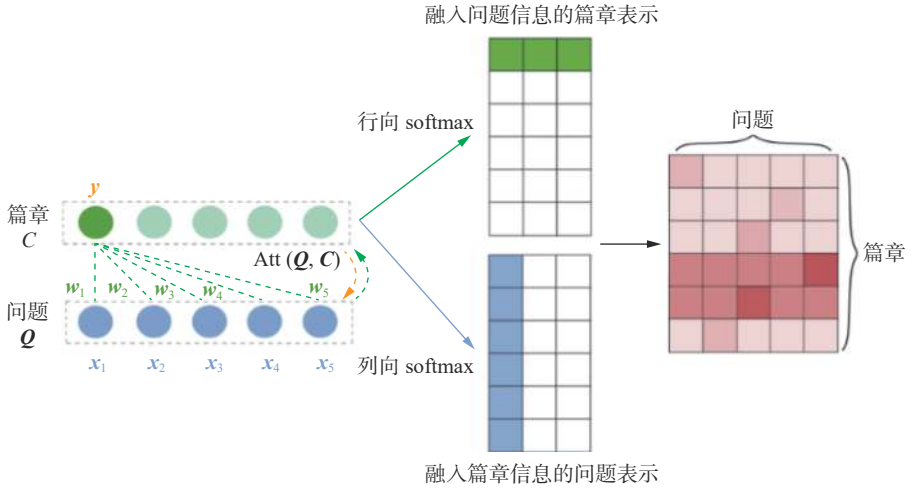


图 2　注意力机制的原理示意

Fig. 2　Structure of attention mechanism

BiDAF[4]提出的双向注意力机制不仅计算融入问题信息的篇章表示，也计算了融入篇章信息的问题表示，从而进一步提高了模型对问题和篇章的理解能力。注意力机制相比于循环神经网络，其复杂度更小，参数量也更少，解决了循环神经网络不能并行计算和短期记忆的问题。自注意力机制在机器阅读理解任务中常被用来关注篇章自身的内容，即Att(C,C)，目的是计算篇章中各个词的相似度，以学习到篇章自身词与词之间的关系，R-Net[6]、T-Reader[22]、HQACL[23]模型均采用自注意力机制提高了模型对篇章的理解能力。

4）答案预测层：答案预测层用于输出问题的答案。机器阅读理解的任务类型不同，答案形式也不同。完形填空式任务的输出是篇章中的一个单词或实体；选择式任务的输出是从候选答案中选出正确答案；片段抽取式任务需要从篇章中抽取连续子序列作为输出；对于自由作答式的任务，文本生成技术通常被用于该层来生成问题的答案。

2.2　基于预训练语言模型的机器阅读理解技术

预训练模型已经在自然语言处理的多项下游任务中取得了优秀的性能，包括 OpenAI GPT[24]、BERT[9]、XLNet[10]等，可以有效获取句法和语义信息，并进行文本表示。预训练方法通常用于机器阅读理解任务的词嵌入层，将自然语言文本编码成固定长度的向量。词的表示方法中，独热编码无法体现词与词之间的关系；分布式词向量表示方法虽然可以在低维空间中编码并通过距离度量词与词之间的相关性，但并没有包含上下文信息，为了解决这个问题，基于预训练的词表示方法被提出并应用。

Transformer[5] 是第一个完全基于注意力机制的序列生成模型, BERT[9] 提出利用双向 Transformer 预训练得到上下文级别的词表示。XLNet[10] 以自回归语言模型为基础融合自编码语言模型的优点, 克服了自回归语言模型无法对双向上下文信息进行建模的缺点。XLNet[10] 引入双流自注意力机制以解决目标位置信息融入的问题, 同时使得模型能够处理更长的输出长度。但是, 常规的预训练方法无法对文本中的实体及关系建模, ERICA[25] 框架被提出用于解决该问题, 实现深度理解, 它可以提升典型的预训练模型 BERT[9] 与 RoBERTa[26] 在多个自然语言理解任务上的性能, 包括机器阅读理解。

此外, 在面向中文的预训练语言模型中, ChineseBERT[27] 将具有中文特性的字形和拼音融入预训练过程中, 在机器阅读理解等多项中文自然语言处理任务中达到了 SOTA, 该模型在训练数据较少的情况下优于常规的预训练模型。

尽管预训练语言模型的上下文表示已经包含了句法、语义等知识, 但挖掘上下文表示所蕴含的常识的工作较少, 它对于机器阅读理解是非常重要的。Zhou 等[28] 在不同具有挑战性的测试中检验 GPT[24]、BERT[9]、XLNet[10] 和 RoBERTa[26] 的常识获取能力, 发现模型在需要更多深入推理的任务上表现不佳, 这也表明常识获取依然是一个巨大挑战。

2.3　基于知识推理的机器阅读理解技术

如何提高系统的可解释性是人工智能领域一项重要挑战, 对于机器阅读理解等自动问答任务, 特别是复杂问题回答, 机器需要具备通过推理来获取答案的能力, 而目前的深度学习方法可解释性较差是一个普遍现象, 无法将推理过程进行显示地表达。常见的基于知识推理的机器阅读理解技术包括语义蕴含推理、知识图谱推理以及基于检索的多跳推理, 如图 3 所示。

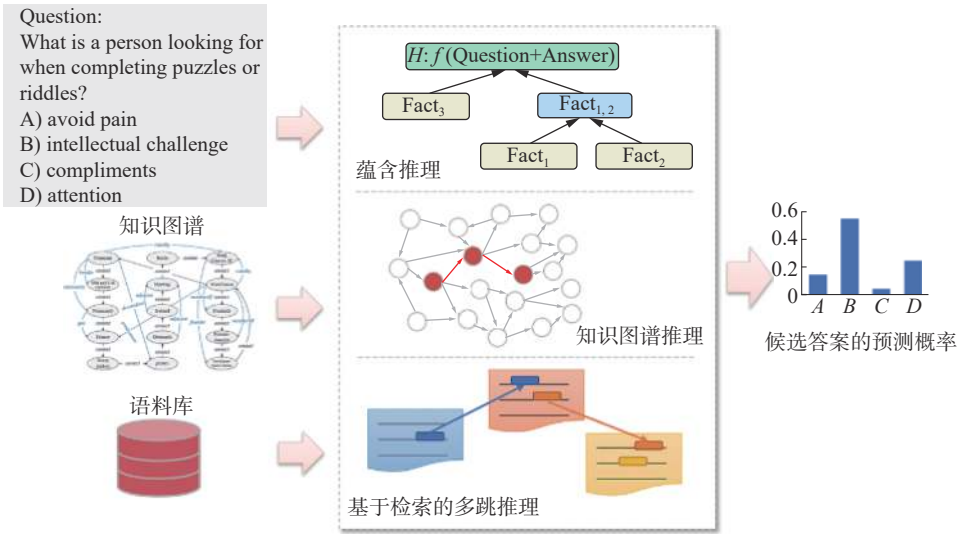


图 3　基于知识推理的机器阅读理解技术

Fig. 3　Technologies of machine reading comprehension model based on knowledge inference

基于语义蕴涵推理的问答方法: 问题回答可转换为文本蕴涵任务, 将问题和候选答案组成假设, 系统决定候选知识库是否能推出假设。Shi 等[29] 研究一种神经符号问答方法, 将自然逻辑推理集成到深度学习体系结构中, 建立推理路径, 计算中间假设和候选前提的蕴含分值, 提升模型性能并具有可解释性。Dalvi 等[30] 以语义蕴涵树的方式来生成解释, 并创建了首个包含多阶蕴涵树的数据集 EntailmentBank, 逐步从已知事实逼近由问题和答案构成的最终假设, 为自动问答任务生成更加丰富的和系统的解释, 通过一系列的推理链来支撑正确答案的获取。同时, 也出现了无效蕴涵推理等问题, 有待优化, 但该方法在进一步提高模型的可解释性方面进行了有效尝试。

基于知识推理的问答方法: 知识图谱是一种以关系有向图形式存储人类知识的资源, 与无结构的文本数据相比, 结构化的知识图谱以一种更加清晰准确的方式表示人类知识, 从而为高质量问答系统的构建带来了前所未有的发展机遇, 有代表性的大规模知识图谱包括 ConceptNet[31]、DBpedia[32]、YAGO[33] 等。常识问答数据集 CommonsenseQA[17] 是通过从 ConceptNet[31] 中抽取出具有相同语义关系的知识, 构建问题和答案。

基于知识增强的常识类问题回答中, 首先面临的问题是, 知识图谱与自然语言文本表达的异构性。Bian 等[34] 提出一种将知识转化为文本的

框架，用于为常识问答提供评测基准，在 CommonsenseQA[17] 上取得最优性能，同时也表明知识的潜力在常识问答任务上未得到充分利用，在上下文相关的高质量知识选择、异构知识的利用等方面有待继续深入。知识表达通常采用基于图的方法，但该方法关注于拓扑结构，忽略了节点和边所蕴含的文本信息。Yan 等[35] 提出基于 BERT[9] 的关系学习任务，将自然语言文本与知识库对齐进行推理，并证明了关系学习方法的有效性。

更进一步，针对生成式常识推理这一更具有挑战性的任务，现有模型很难生成正确的句子，其中一个重要原因是没有有效结合知识图谱中常识知识之间的关系信息。Liu 等[36] 研究知识图谱增强的 KG-BART 模型，结合知识图谱生成更有逻辑性更自然的句子表达，通过图注意力聚合概念语义，增强对新概念集的泛化能力。该方法的实验结果证明，结合知识图谱后，模型可以生成质量更高的语句。KG-BART 模型可以迁移到常识问答等以常识为中心的下游任务。

基于检索与知识融合的多跳推理方法：多跳问答是一项需要多层推理的挑战性任务，在实际应用中十分普遍。该任务需要从大规模语料库中发现回答问题的支撑证据，分析分散的证据片段，进行多跳推理实现对问题的回答。多跳问答通常使用实体关系进行分步推理，已有方法通过预测序列关系路径（较难优化）或汇聚隐藏的图特征进行答案推理（可解释性差）。Shi 等[37] 提出了 TransferNet，TransferNet 使用同一框架支持实体标签和文本关系的表示，推理的每一环节关注问题的不同部分，传递实体信息，取得优秀性能表现。Li 等[38] 提出新的检索目标“hop”来发现维基百科中的隐藏证据，将 hop 定义为含有超链接的文本和链接到的文档，检索维基百科回答复杂问题。

针对现有基于单跳的图推理方法会遗漏部分重要的非连续依赖关系的难题，Jiang 等[39] 定义高阶动态切比雪夫近似图卷积网络，将直接依赖和长期依赖的信息融合到一个卷积层来增强多跳图推理，在文本分类、多跳图推理等多个任务上进行实验，取得了最优性能。Feng 等[40] 提出一种适合多跳关系推理的模型 MHGRN，结合图神经网络和关系网络，通过多跳信息传递，在长度最多为 k 的关系路径上传递信息，赋予图神经网络直接建模路径的能力。

3　数据集与评价指标

3.1　数据集

大规模高质量数据集的发布是推动机器阅读理解快速发展的重要因素，根据不同任务类型，代表性数据集如表 1 所示，发布时间轴如图 4 所示。其中，完形填空式问答任务数据集中 BookTest[14] 的问题规模最大，在 2015 年以后片段抽取式的数据集规模均在万级以上。

表 1　机器阅读理解主要数据集统计

Table 1　Statistics of machine reading comprehension datasets

任务类型	名称	训练集问题数量	开发集问题数量	测试集问题数量
完型填空式	CNN/Daliy Mail[15]	297 113	13 368	11 490
	FaceBook CBT[1]	669 343	8 000	8 000
	BookTest[14]	14 140 825	10 000	10 000
选择式	MCTest[41]	1 480	320	840
	WikiHop[16]	43 738	5 129	2 451
	RACE[42]	87 866	4 887	4 934
	OpenbookQA[43]	4 957	500	500
	CommonsenseQA[17]	9 741	1 221	1 140
片段抽取式	WikiQA[44]	2 118	296	633
	SQuAD 1.1[2]	87 599	10 570	9 533
	NewsQA[18]	107 000	6 000	6 000
	SearchQA[45]	99 820	13 393	27 248
	TriviaQA[3]	138 384	17 944	17 207
	Quasar-T[41]	37 012	3 000	3 000

续表 1

任务类型	名称	训练集问题数量	开发集问题数量	测试集问题数量
	SQuAD 2.0[46]	130 319	11 873	8 862
	CoQA[47]	110 000	7 000	10 000
	CMRC 2018[48]	10 321	3 351	4 895
	HotpotQA[49]	90 564	7 405	7 405
	BeerQA[50]	134 043	14 121	14 932
自由作答式	NarrativeQA[51]	32 747	3 461	10 557
	DuReader[8]	271 574	10 000	20 000
	MS MARCO[52]	808 731	101 093	101 092
	Natural Questions[53]	307 373	7 830	7 843
	DROP[19]	77 409	9 536	9 622

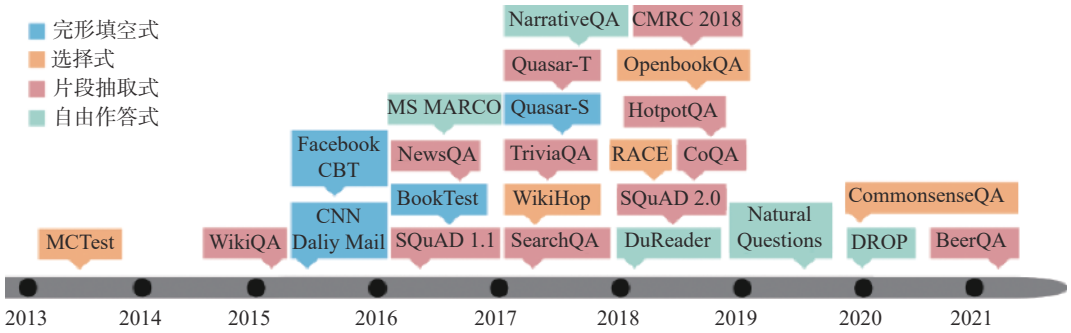


图 4　机器阅读理解数据集发布时间轴

Fig. 4　Time axis of machine reading comprehension datasets

其中, 规模较大的数据集如 SQuAD 2.0[46]、BookTest[14] 和 NewsQA[18], 推动了 BiDAF[4]、R-Net[6] 等经典模型的发展。

在片段抽取式数据集中应用广泛的数据集有 SQuAD[2], 模型 QANet[20] 与 BERT[9] 在该数据集上表现尚佳, 且 BERT[9] 在两个评测指标上首次超越了人类水平, 在此基础上预训练语言模型与微调的方法成为主流。在自由式问答数据集 DROP[19] 中, 仍然与人类 F_1 值为 0.9642 的水平存在一定距离。在需要推理的任务中, 在选择式数据集 CommonsenseQA[17] 性能表现优异的模型中用到了知识图谱和图神经网络, 评测排名第一的 DEKCOR[54] 模型还引入了辅助的篇章信息, 但与人类水平仍有差距。

3.2　评价指标

3.2.1　准确率

准确率是最常用的评价指标, 它表示机器阅读理解模型正确回答的问题占所有问题的百分比。设机器阅读理解任务包含n个问题, 其中模型正确回答了m个问题[55], 则准确率a的计算为

$$a=\frac{m}{n}\times 100\% \tag{1}$$

准确率一般用于评价完形填空式和选择式问答任务, 例如 Facebook Children's Books Test[1]、CommonsenseQA[17] 等。片段抽取式问答数据集中的 SearchQA[45] 和自由式问答数据集中的 DROP[19] 也使用了该指标。

EM(exact match)值与准确率的计算相同, EM 值要求式 (1) 中的m为所有问题中模型输出答案与正确答案完全相同的个数, 即模型输出答案与正确答案中的每个单词和位置都必须相同。在片段抽取式问答任务中, EM 值与准确度相同, 且使用 EM 值作为它们的评价指标, 例如 SQuAD[2]、TriviaQA[3]、HotpotQA[49] 等。

3.2.2　F_1 值

F_1 值评价指标, 表示数据集中标准答案与模型预测的答案之间平均单词的覆盖率, 将精确率P(precision)和召回率R(recall)折中。其中, 精确率为预测正确的答案占所有预测答案的百分比, 召回率则是预测正确的答案占所有标准答案的百分比, 而 F_1 值是将这两个指标综合在一起, 即

$$F_1=\frac{2}{P^{-1}+R^{-1}}=\frac{2PR}{P+R} \tag{2}$$

F_1 值通常是片段抽取式问答任务采用的评价指标，例如 SQuAD[2]、HotpotQA[49] 等。自由式问答任务中的 Natural Questions-Short[53] 也使用了 F_1 值。相比于 EM，F_1 值允许模型预测答案和正确答案之间有一定范围偏差，因此，数据的类别分布不平衡时，F_1 值更适用。

3.2.3　其他评价指标

ROUGE-L[56] 相比于 EM 值和 F_1 值更灵活，其值用于评价预测答案和真实答案之间的相似度，但候选答案的长度会影响 ROUGE-L 的值；BLEU[57] 最初用于机器翻译任务中，不仅可以评价预测答案和真实答案之间的相似度，还可以考察候选答案语言表达流畅性，但 BLEU 对词重复和短句现象不利。因此这两个指标通常用于不受原语境限制的任务中。一般在自由式问答中使用 ROUGE-L[56] 和 BLEU[57] 作为评价指标，例如 DuReader[8]、DR-OP[19] 等。

4　机器阅读理解领域的挑战和发展趋势

目前，在大规模高质量数据集的推动下，机器阅读理解领域的研究取得了快速发展，甚至在部分评测任务上已经超过了人类的表现。但是，在一些新提出的任务或研究方向上，机器目前的性能远未达到人类的理解水平。该领域目前的主要挑战和发展趋势概括如下。

知识驱动与推理技术提升可解释性：将知识融入机器阅读理解任务中来实现复杂的问题回答是基于人类的思考方式提出的一种策略[58]。知识驱动的阅读理解模型通过引入外部知识，辅助理解篇章内容并回答问题。大规模知识库的构建也需要考虑知识的获取方式、多模态资源中知识的获取、不同来源的知识的融合。同时，知识驱动与推理技术的运用可以较好地解决基于神经网络模型可解释性差的问题。

对话式问答任务中的语义理解：对话式问答同样是根据人类获取知识的习惯而提出的任务，让机器根据已有的一系列问答序列，对当前问题进行回答。其中，问答序列具有时序性和前后关联性，如何理解当前问题与历史问答记录的关系是该任务的一大难点。此外，指代消解技术在该任务中非常重要，机器需要根据历史问答记录，准确理解篇章、问题中的指代实体，进行补全。

机器阅读理解模型的健壮性：目前的机器阅读理解模型往往过于依赖文本表面的信息，而缺乏深入的理解。在篇章中引入干扰数据，生成对抗样本，结果表明，多数现有模型性能明显下降。如何生成有效的对抗样本，通过对抗训练提升模型的健壮性成为研究的重点。

5　结束语

机器阅读理解是自然语言处理领域的难点问题，它是评价和度量机器理解自然语言程度的重要任务。近年来基于深度学习技术的机器阅读理解模型研究发展迅速。本文介绍了机器阅读理解任务划分，对机器阅读理解相关技术进行了分析，包括端到端的神经网络模型、预训练语言模型以及知识推理等方法，并选取了各个任务中有代表性的数据集进行统计分析，介绍了不同机器阅读理解任务中常用的评价指标。目前机器的语言理解能力距离人类的理解水平还有较大差距，我们对该领域面临的挑战和发展趋势进行了分析。

参考文献：

[1] HILL F, BORDES A, CHOPRA S, et al. The goldilocks principle: reading children's books with explicit memory representations[C]//Proceedings of the 4th International Conference on Learning Representations. San Juan: OpenReview, 2016: 1−13.

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

[3] 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. Vancouver: ACL, 2017: 1601–1611.

[4] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[C]//5th International Conference on Learning Representations. Toulon: OpenReview, 2017: 1−12.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017(30): 5998–6008.

[6] WANG W H, YANG N, WEI F R, et al. Gated Self-matching networks for reading comprehension and question answering[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouve: ACL, 2017: 189–198.

[7] 梁小波, 任飞亮, 刘永康, 等. N-Reader: 基于双层 Self-attention 的机器阅读理解模型 [J]. 中文信息学报, 2018, 32(10): 130–137.

LIANG Xiaobo, REN Feiliang, LIU Yongkang, et al. N-reader: machine reading comprehension model based on double layers of self-attention[J]. Journal of Chinese information processing, 2018, 32(10): 130–137.

[8] HE Wei, LIU Kai, LIU Jing, et al. DuReader: a Chinese machine reading comprehension dataset from real-world applications[C]//Proceedings of the Workshop on Machine Reading for Question Answering. Melbourne: ACL, 2018: 37–46.

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

[10] YANG Zhilin, DAI Zihang, YANG Yiming, et al. XLNet: generalized autoregressive pretraining for language understanding[C]//Annual Conference on Neural Information Processing Systems. Vancouver: ACM, 2019: 1−18.

[11] 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. Florence: ACL, 2019: 2694–2703.

[12] YASUNAGA M, RWN H, BOSSELUT A, et al. QA-GNN: reasoning with language models and knowledge graphs for question answering[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. ACL, 2021: 535–546.

[13] YAGCIOGLU S, ERDEM A, ERDEM E, et al. RecipeQA: a challenge dataset for multimodal comprehension of cooking recipes[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018: 1358–1368.

[14] KEMBHAVI A, SEO M, SCHWENK D, et al. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 5376–5384.

[15] HERMANN K M, KOčISKỳ T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[C]//Advances in Neural Information Processing Systems. Montréal: MIT Press, 2015: 1693–1701.

[16] 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(1): 287–302.

[17] TALMOR A, HERZIG J, LOURIE N, et al. CommonsenseQA: a question answering challenge targeting commonsense knowledge[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: NAACL, 2019: 4149–4158.

[18] TRISCHLER A, WANG T, YUAN X, et al. NewsQA: a machine comprehension dataset[C]//Proceedings of the 2nd Workshop on Representation Learning for NLP. Vancouver: ACL, 2017: 191–200.

[19] DUA D, WANG Y, DASIGI P, et al. DROP: a reading comprehension benchmark requiring discrete reasoning over paragraphs[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: NAACL, 2019: 2368–2378.

[20] YU A W, DOHAN D, LUONG M T, et al. QANet: combining local convolution with global self-attention for reading comprehension[EB/OL]. (2018−04−23)[2021−07−13].https://arxiv.org/abs/1804.09541.

[21] PENNINGTON J, SOCHER R, MANNING CD. GloVe: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Qatar: ACL, 2014: 1532–1543.

[22] 郑玉昆, 李丹, 范臻, 等. T-Reader: 一种基于自注意力机制的多任务深度阅读理解模型 [J]. 中文信息学报, 2018, 32(11): 128–134.
ZHENG Yukun, LI Dan, FAN Zheng, et al. T-Reader: a multi-task deep reading comprehension model with self-attention mechanism[J]. Journal of Chinese information processing, 2018, 32(11): 128–134.

[23] DU Yongping, GUO Wenyang, ZHAO Yiliang. Hierarchical question-aware context learning with augmented data for biomedical question answering[C]//2019 IEEE International Conference on Bioinformatics and Biomedicine. San Diego: IEEE, 2019: 370–375.

[24] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving Language Understanding by Generative Pre-training[EB/OL]. [2021−07−13].https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[25] QIN Y, LIN Y, TAKANOBU R, et al. ERICA: improving entity and relation understanding for pre-trained language models via contrastive learning[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Virtual Event: ACL, 2021: 3350–3363.

[26] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[EB/OL].

(2019−07−26)[2021−07−13].https://arxiv.org/abs/1907.11692.

[27] SUN Zijun, LI Xiaoya, SUN Xiaofei, et al. ChineseBERT: Chinese pretraining enhanced by glyph and pinyin information[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Krung Thep Maha Nakhon: ACL, 2021: 2065–2075.

[28] ZHOU Xuhui, ZHANG Yue, CUI Leyang, et al. Evaluating commonsense in pre-trained language models[C]//The Thirty-Fourth AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 9733–9740.

[29] SHI Jihao, DING Xiao, DU Li, et al. Neural natural logic inference for interpretable question answering[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Virtual Event: ACL, 2021: 3673–3684.

[30] DALVI B, JANSEN P, TAFJORD O, et al. Explaining answers with entailment trees[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Virtual Event: ACL, 2021: 7358–7370.

[31] LIU H, SINGH P. Conceptnet—a practical commonsense reasoning tool-kit[J]. BT technology journal, 2004, 10(22): 211–226.

[32] BIZER C, LEHMANN J, KOBILAROV G, et al. DBpedia—a crystallization point for the web of data[J]. Journal of web semantics, 2019, 7(3): 154–165.

[33] TANON T P, WEIKUM G, SUCHANEK F. YAGO 4: a reason-able knowledge base[C]//European Semantic Web Conference. Heraklion: Springer, 2020: 583–596.

[34] BIAN Ning, HAN Xianpei, CHEN Bo, et al. Benchmarking knowledge-enhanced commonsense question answering via knowledge-to-text transformation[C]//The Thirty-Fifth AAAI Conference on Artificial Intelligence. New York: AAAI, 2021: 12574–12582.

[35] YAN Yuanmeng, LI Rumei, WANG Sirui, et al. Large-scale relation learning for question answering over knowledge bases with pre-trained language models[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Krung Thep Maha Nakhon: ACL, 2021: 3653–3660.

[36] LIU Ye, WAN Yao, HE Lifang, et al. KG-bart: knowledge graph-augmented bart for generative commonsense reasoning[C]//The Thirty-Fifth AAAI Conference on Artificial Intelligence. New York: AAAI, 2021: 6418–6425.

[37] SHI Jiaxin, CAO Shulin, HOU Lei, et al. TransferNet: an effective and transparent framework for multi-hop question answering over relation graph[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Krung Thep Maha Nakhon: ACL, 2021: 4149–4158.

[38] LI Shaobo, LI Xiaoguang, SHANG Lifeng, et al. Hopretriever: retrieve hops over wikipedia to answer complex questions[C]//The Thirty-Fifth AAAI Conference on Artificial Intelligence. New York: AAAI, 2021: 13279–13287.

[39] JIANG Shuoran, CHEN Qingcai, LIU Xin, et al. Multi-hop graph convolutional network with high-order chebyshev approximation for text reasoning[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Krung Thep Maha Nakhon: ACL, 2021: 6563–6573.

[40] FENG Yanlin, CHEN Xinyue, LIN B Y, et al. Scalable multi-hop relational reasoning for knowledge-aware question answering[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2020: 1295–1309.

[41] RICHARDSON M, BURGES C J, 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. Seattle: ACL, 2013: 193–203.

[42] LAI G, XIE Q, LIU H, et al. RACE: Large-scale reading comprehension dataset from examinations[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017: 785–794.

[43] MIHAYLOV T, CLARK P, KHOT T, et al. Can a suit of armor conduct electricity? a new dataset for open book question answering[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018: 2381–2391.

[44] YANG Y, YIH W, MEEK C. WikiQA: a challenge dataset for open-domain question answering[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015: 2013–2018.

[45] DUNN M, SAGUN L, HIGGINS M, et al. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine[EB/OL]. [2017–04–18](2021−07−13). https://arxiv.org/abs/1704.05179.

[46] 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. Melbourne: ACL, 2018: 784–789.

[47] REDDY S, CHEN D, MANNING C D. CoQA: a conver-

sational question answering challenge[J]. Transactions of the association for computational linguistics, 2019(7): 249–266.

[48] CUI Y, LIU T, CHE W, et al. A span-extraction dataset for Chinese machine reading comprehension[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019: 5883–5889.

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

[50] QI P, LEE H, SIDO T, et al. Answering open-domain questions of varying reasoning steps from text[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021: 3599–3614.

[51] KOčISKỳ T, SCHWARZ J, BLUNSOM P, et al. The NarrativeQA reading comprehension challenge[J]. Transactions of the association for computational linguistics, 2018(6): 317–328.

[52] NGUYEN T, ROSENBERG M, SONG X, et al. MS MARCO: a human generated machine reading comprehension dataset[C]//Conference and Workshop on Neural Information Processing Systems. Barcelona: MIT Press, 2016: 1−10.

[53] KWIATKOWSKI T, PALOMAKI J, REDFIELD O, et al. Natural questions: a benchmark for question answering research[J]. Transactions of the association for computational linguistics, 2019(7): 453–466.

[54] XU Y, ZHU C, XU R, et al. Fusing context into knowledge graph for commonsense question answering[EB/OL]. (2020−12−09)[2021−07−13]. https://arxiv.org/abs/2012.04808.

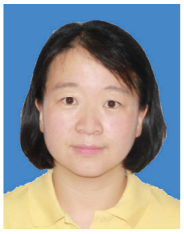
[55] 顾迎捷, 桂小林, 李德福, 等. 基于神经网络的机器阅读理解综述 [J]. 软件学报, 2020, 31(7): 2095–2126.
GU Yingjie, GUI Xiaolin, LI Defu, et al. Survey of machine reading comprehension based on neural network[J]. Journal of software, 2020, 31(7): 2095–2126.

[56] LIN C Y. ROUGE: a Package for automatic evaluation of summaries[EB/OL]. (2004−07−21)[2021−07−13]. https://aclanthology.org/W04-1013.pdf.

[57] 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. Philadelphia: ACL, 2002: 311–318.

[58] 曾帅, 王帅, 袁勇, 等. 面向知识自动化的自动问答研究进展 [J]. 自动化学报, 2017, 43(9): 1491–1508.
ZENG Shuai, WANG Shuai, YUAN Yong, et al. Towards knowledge automation: a survey on question answering systems[J]. Acta automatica sinica, 2017, 43(9): 1491–1508.

作者简介:



杜永萍, 教授, 主要研究方向为信息检索、信息抽取和自然语言处理。主持国家语委科研项目和北京自然科学基金项目 2 项。发表学术论文 50 余篇。

赵以梁, 硕士研究生, 主要研究方向为自然语言处理和自动问答。

阎婧雅, 硕士研究生, 主要研究方向为自然语言处理和自动问答。