文章编号：1003-0077(2022)10-0001-16

面向学科题目的文本分析方法与应用研究综述

黄振亚[1,2]，刘 淇[1,2]，陈恩红[1,2]，林 鑫[1,2]，何理扬[1,2]，刘嘉聿[1,2]，王士进[2,3]

(1. 中国科学技术大学 大数据分析与应用安徽省重点实验室，安徽 合肥 230027；
2. 认知智能全国重点实验室，安徽 合肥 230088；
3. 讯飞华中人工智能研究院，湖北 武汉 430058)

摘 要：分析学科题目含义、模拟人类解决问题，是当前“人工智能＋教育”融合研究的重要方向之一。近年来，智能教育系统的快速发展积累了大量学科题目资源，为相关研究提供了数据支撑。为此，利用大数据分析与自然语言处理相关的技术，研究者提出了大量面向学科题目的文本分析方法，开展了许多重要的智能应用任务，对探索人类知识学习等认知能力具有重要意义。该文围绕智能教育与自然语言处理交叉领域，介绍了若干代表性研究任务，包括题目质量分析、机器阅读理解、数学题问答、文章自主评分等，并对相应研究进展进行阐述和总结；此外，对相关数据集和开源工具包进行了总结和介绍；最后，展望了多个未来研究方向。

关键词：学科题目；题目质量分析；机器阅读理解；数学题问答；文章自主评分

中图分类号：TP391　　**文献标识码**：A

A Survey on Text Analysis Methods and Applications for Educational Questions

HUANG Zhenya[1,2], LIU Qi[1,2], CHEN Enhong[1,2], LIN Xin[1,2], HE Liyang[1,2], LIU Jiayu[1,2], WANG Shijin[2,3]

(1. Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China, Hefei, Anhui 230027, China;
2. State Key Laboratory of Cognitive Intelligence, Hefei, Anhui 230088, China;
3. iFLYTEK AI Research (Central China), Wuhan, Hubei 430058, China)

Abstract: One of the important research directions on the integration of artificial intelligence into pedagogy is analyzing the meanings of educational questions and simulating how humans solve problems. In recent years, a large number of educational question resources have been collected, which provides the data support of the related research. Leveraging the big data analysis and natural language processing related techniques, researchers propose many specific text analysis methods for educational questions, which are of great significance to explore the cognitive abilities of how human master knowledge. In this paper, we summarize several representative topics, including question quality analysis, machine reading comprehension, math problem solving, and automated essay scoring. Moreover, we introduce the relevant public datasets and open-source toolkits. Finally, we conclude by anticipating several future directions.

Keywords: educational questions; question quality analysis; machine reading comprehension; math problem solving; automated essay scoring

0 引言

让机器模拟人类解决问题的过程，从而掌握知识、培养技能，是人工智能研究的目标之一[1]。其中，理解各类学科题目(如英语题、数学题、作文题等)，并解答相应题目，是一类代表性的研究任务。相关研究涉及教育心理测量、人工智能、自然

收稿日期：2021-08-04　**定稿日期**：2022-07-18

基金项目：国家自然科学基金(62106244，U20A20229，61922073)；中央高校基本科研业务费专项资金(WK2150110021)

语言处理等多个交叉领域，长期吸引着众多来自教育学、心理学、计算机科学、脑科学等方向的研究者[2]。

解答学科题目的基础是充分分析题目数据的特点，理解各类学科题目文本的含义，并评估题目质量。在早期研究中，受限于课堂学习场景，研究者大多设计标准化测试进行实证研究[3-4]，如通过测试结果计算题目的难度等，从而评价学科题目的质量。这种研究过程具有相对严谨的流程，但其效率低，相关结论受到组织过程中多种因素的干扰（如受试者的偏差等）[3]，难以形成能够评估和解答各类问题的有效模型。近年来，伴随人工智能和自然语言处理等技术的快速发展，研究者设计模型直接阅读学科题目，评价题目质量，且通过模拟学习者解决问题的过程，可以有效地自动求解答案，具有更好的可扩展性。相关研究对探索人工智能在阅读理解、语义分析、知识推理和自主评测等方面复杂的类人认知能力具有重要意义[5-7]。本文重点探讨和总结相应代表性研究任务，包括题目质量分析[8]、机器阅读理解[5]、数学题问答[6]、文章自主评分[7]四类任务。

相比于传统领域的常见文本数据，如新闻数据、用户评论等[9-10]，学科题目的文本数据具有许多独特特点，给理解学科题目的含义带来众多挑战。首先，学科题目的编写通常具有明确的知识内涵[11]。其次，学科题目具有独特的教学质量属性，如难度、区分度等[12]。最后，学科题目之间的知识含义关联更为重要[13]。因此，题目质量分析是相关研究的基础，需要提出针对性的方法对学科题目进行深入理解和分析。围绕这一研究目标，研究者针对题目难度评估[8,14-16]、知识点预测[11]、题目表征[17]、相似度分析[12,18-20]等任务开展了大量研究，形成了一系列研究成果。

机器阅读理解任务要求模型阅读英语文章材料，依据材料内容抽取答案，回答相关问题。要解答机器阅读理解任务，需要机器阅读问题和材料，理解文本内容的语义，并从中抽取相关信息，这是研究类人语义理解能力的基础任务之一。

数学题问答任务要求模型分析数学题目，模拟人类进行必要的数学推理和计算（如数学表达式），给出答案。数学题问答任务需要机器在语义理解的基础上，应用一定的数学知识进行形式化推理，从而进一步探索类人知识运用和逻辑推理等能力。

文章自主评分任务要求模型模仿人类专家的评测标准，对给定文章进行自动打分。文章自动评分任务需要机器能够对文章进行自主综合评价，例如从语法正确性、文章表达结构与内容扣题程度等多个不同的维度对文章进行评估，对智能算法有更高的要求。

上述代表性研究具有重要的实际应用价值。首先，通过对学科题目进行分析，可以帮助各类智能教育系统为学习者提供众多学习服务，例如个性化推荐等[21]。其次，研究成果对于多个教育领域的传统研究（如认知诊断等）产生了积极作用[22]。此外，上述研究任务也是当前“人工智能＋教育”亟需解决的重要问题，有望推动交叉领域技术的发展。

下文中，首先介绍典型学科题目数据；接着分节总结题目质量分析、机器阅读理解、数学题问答和文章自主评分等四个代表性研究任务的研究进展；然后介绍相关任务的开源工具包；最后对未来研究方向进行展望。

1 数据集

智能教育系统收集并积累了大量的学科题目数据，为相关研究提供了数据基础。目前，代表性的公开数据集主要包括三类学科题目，即英语阅读题、数学问答题、文章写作题。

1.1 英语阅读题数据集

英语阅读题是一类基础的学科题目，可以支持多个研究与应用任务。按题型划分，英语题目主要包括两大类：完型填空（Cloze）和阅读理解（Reading Comprehension）。公开的完型填空数据集主要包括CNN & Daily Mail[23]、CBT（Children's Book Test）[24]和CliCR[25]等。

在阅读理解问答题数据集中，根据获取答案的方式，可以划分为：多项选择、片段抽取和自由回答三种类型。其中，代表性的多项选择型数据集主要包括MCTest[26]和RACE[27]，片段抽取型数据集主要包括SQuAD[28]、NewsQA[29]、TriviaQA[30]、DuoRC[31]和CMRC2018[32]等，自由回答型数据集主要包括MS MARCO[33]、NarrativeQA[34]、SearchQA[35]、

DuReader[36]。目前，应用最为广泛的是 SQuAD，数据示例如图 1(a)所示，主要包含三个部分：一段上下文文章(Context)，一个问题(Question)，以及一个来自文章中某个片段的答案(Answer)。三类数据集具有相近的数据结构，仅答案来源有所区别。其中，多项选择题要求从多个候选项中选取正确答案；片段抽取题要求抽取文章中一段文本(单词或词组)以回答问题；自由回答型问题的答案根据词汇表生成，包含不在文章中出现的单词或词组。

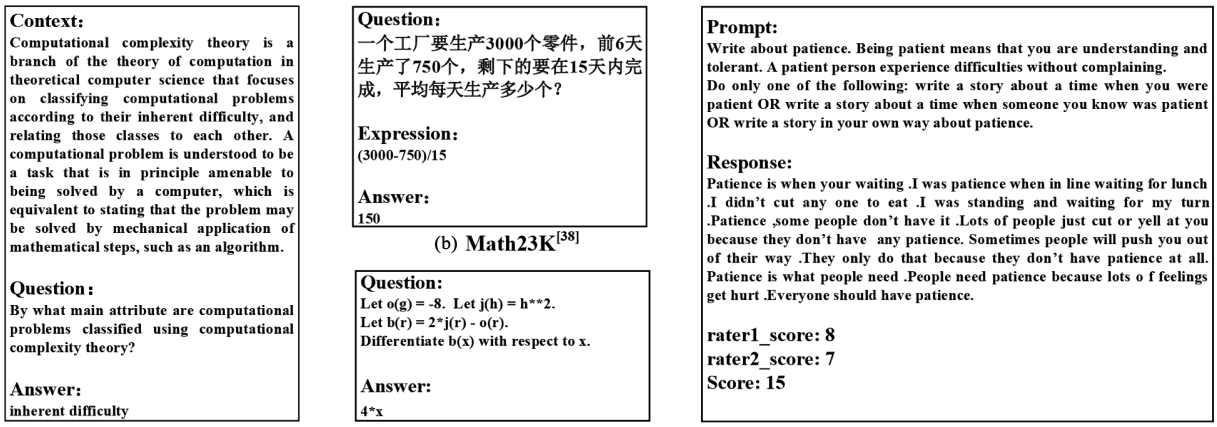

图 1　代表性数据集样例

(a)英语阅读题；(b)、(c)数学问答题；(d)文章写作题

1.2　数学问答题数据集

常见的数学问答题主要包括两类，即数学应用题和数学简答题。其中，数学应用题(Math Word Problem，MWP)通常面向小学基础教育学习，是一类最为基础的题目。目前，应用较为广泛的数学应用题数据集包括 Dolphin[6,37]、Math23K[38]、MAWPS[39]、MathQA[40] 和 ASDiv[41]。其中，Dolphin、MAWPS、MathQA 和 ASDiv 是英语数据集，Math23K 为中文数据集。图 1(b)展示了 Math23K 中的数据样例，包含三个部分：题目问题描述(Question)，数学表达式(Expression)和答案(Answer)。其中，数学表达式是用于回答该问题的数学运算式，通常由数字和 6 个基本运算符号(加、减、乘、除、求余、幂)组成。

数学简答题是一类面向初高中学生的数学问答题。相较于数学应用题，数学简答题较为复杂，题目描述不再局限于自然语言文本，还包括明确的数学公式等。据笔者所知，目前开源的唯一大规模数学简答题数据集由 DeepMind 团队发布[42]，涉及 8 个数学知识领域(如 Algebra、Arithmetic、Calculus 等)。图 1(c)展示了 Calculus 领域问答题样例，包含：问题(Question)和答案(Answer)。详细描述可参考文献[42]。

1.3　文章写作题数据集

文章(Essay)写作题数据在机器自动评分任务中使用较多。目前，代表性数据集主要包括 CLC-FCE[43]，ASPS①，SemEval-2013[44]。图 1(d)展示了 ASPS 中的数据样例，包括题目要求(Prompt)、一篇文章(Response)、评分(Score)。其中，文章由学生根据要求撰写，评分是一位或者多位专家教师对文章的打分总和。另外有一些数据的打分用等级表示，例如，TOEFL11[45]数据中使用低、中、高三个等级。

2　题目质量分析

高质量的学科题目对于保证教学活动和算法研究的效果至关重要。因此，精准评价题目质量具有重要意义。首先，学习者阅读题目内容，理解题目含义是其学习掌握知识、运用知识解决问题的前提。因此，模拟学习者分析、理解题目的能力是智能应用(如机器阅读理解、数学题问答、文章自主评分等)的基础[46]。其次，精准分析题目可以帮助高效构建并

① https://www.kaggle.com/c/asap-aes

管理智能教育系统的资源库，减少人工管理工作量，提供众多智能服务（如个性化推荐等），提高学习者的学习效率[17]。

相比于传统领域的常见文本数据，如新闻数据、用户评论、商品描述等，学科题目数据具有以下特点。首先，学科题目的编写具有严谨的知识逻辑和明确的知识内涵[11]，例如计算题与代数知识相关，几何证明题更关注几何图形知识等。此外，教育心理学研究[12]表明学科题目具有重要的属性，如难度、区分度、信度、效度等。这些属性对于衡量一个题目的质量具有重要意义。针对这些特点，研究题目质量评估任务是一个重要的方向。围绕这一目标，研究者提出了面向学科题目的分析方法，在难度评估[14-16]、知识点预测[11]、题目表征[17]、相似度分析[12,19-20]等具体任务中取得了阶段性成果。本节将对相关研究进展进行介绍。

在教育心理学研究中，学科题目的难度评估是对于保证教育公平性和教育质量具有重要意义，已有较长研究历史[3,14-16,47-50]。早期的研究基于标准化测试，提出经典测量理论（Classic Test Theory，CTT）[47]，定义题目难度表示为测试题目的通过率，即通过题目的人数与总人数的比例，通过率越高，测试题目的难度越低。与此同时，用相关分析工作探索与题目难度相关的可能因素。例如，文献[50]发现包括题目类别、知识结构深度等因素与其难度属性相关。因此，有经验的专家教师可以依赖专业知识背景对学科题目的难度进行标注[3]。此外，在一些重要的标准化测试（如TOEFL，GRE等）中，基于项目反应理论（Item Response Theory，IRT），可以利用测试结果评估题目的难度。IRT的相关介绍可以参考文献[48]。然而，上述方案需要花费大量的时间和人力成本，且对参与人员（如标注教师，测试组织者）的专业知识经验要求较高。此外，其评估方式较为主观，难度标准难以统一，因此，难度评估结果容易出现不一致的现象[14]，难以大规模使用。为此，近期的研究工作希望能够直接分析学科题目的文本，自动预测题目的难度。

2017年，Huang等人[14]针对英语阅读理解题目，首次提出一种数据驱动的解决方案，即TACNN（Test-aware Attention-based Convolutional Neural Network）模型，利用历史的测试结果和阅读问答题的文本，自动预测题目的难度属性。图2显示了该模型的框架，包含四个部分，即输入层、语句理解层、语义关联层和难度预测层。

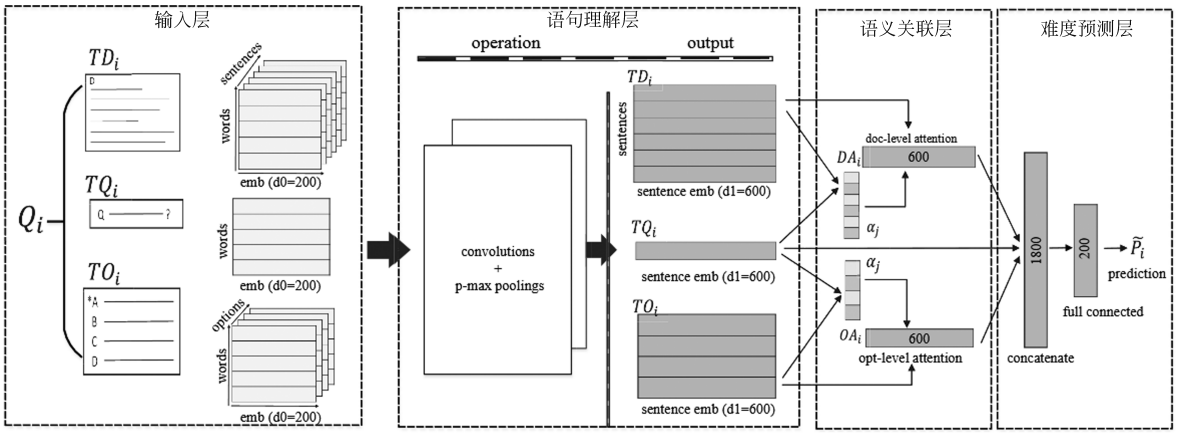

图2 TACNN难度预测模型

图片引自文献[14]

其中，输入层将英语阅读题文本，包括阅读篇章（图2中TD）、问题（TQ）和选项（TQ），划分成一组语句序列，利用词表征将语句中的单词映射到嵌入向量的空间中。接着，语句理解层利用卷积神经网络（CNN）从局部到全局逐步学习题目语句的语义信息。然后，语义关联层用注意力网络衡量问题对阅读材料和选项内容的关联程度，捕获关键的语义信息。最后，在难度预测层中，考虑到历史记录中不同测试群体的差异性，设计了测试依赖的模型训练方式，消除了不同测试结果带来的误差，预测题目难度。该模型充分学习历史数据的经验，可以直接基于题目的文本内容自动预测难度。文献[14]中的实验表明，该模型的预测精准度和稳定性均取得了领先结果。此后，Qiu等人[15]面向多选题，进一步考虑英语阅读题中选项之间的关联信息，改进模型难度预测的效果。借鉴上述研究经验，大量研究者开

展针对学科题目中不同质量属性的分析研究。文献[11]利用知识点标签之间的层级结构，提出层次依赖网络自顶向下预测题目的知识点。Liu 等人[12]结合学科题目的异构信息（含题目文本、几何图形、知识点标签等），设计了多模态注意力网络捕捉题目中“语义-知识”和“语义-图形”中的语义关联，预测题目对之间的相似度。可以看出，题目质量分析的基础是从题目内容中捕获尽可能多的语义知识与逻辑信息。

在上述研究中，研究方案大多基于端到端的有监督模型，其结果依赖题目属性的标注（即难度、知识点、相似度等）质量。然而，获得高质量的题目标注是困难的，依赖于标注者的专业知识。因此，智能教育系统中收集的数据存在大量属性标注缺失的现象。为此，预训练方法是解决属性缺失问题的一种有效方法，它通过在大规模题目语料上预训练优化语言模型参数，使模型能够有效建模文本语义，再在少量标签数据上微调即可达到较好的效果。然而，现有预训练语言模型主要面向通用语料，旨在捕获文本中的语义信息，而学科题目文本的建模则更侧重题目的知识与逻辑含义，因此现有预训练方法难以直接应用于题目文本分析。为此，Yin 等人[17]提出面向学科题目（以数学选择题为例）的预训练模型 QuesNet。相较于经典的预训练模型，如 BERT[51]等，该模型基于学科题目自身特点，从语义理解逻辑和知识推理逻辑两个层面分别设计了自监督训练目标，提高题目表征效果，具体方式如图 3 所示。

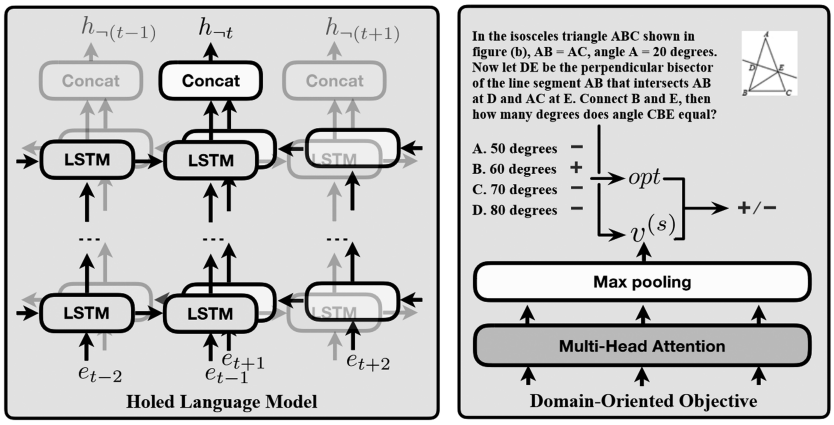

图 3　QuesNet 的自监督学习策略

左：基于语义理解逻辑的学习目标

右：基于知识推理逻辑的学习目标（图片引自文献[17]）

首先，语义逻辑的目标是让模型能够基于学习到的内容语义信息（即每一步题目字词）预测下一步出现的题目内容（两个学习方向），该内容不止包括题目字词，还包括特有的题目元素（如数学公式等）。其次，知识逻辑的目标是利用模型学习的题目表征直接预测题目选项是否为正确答案。该学习过程关注题目与选项在知识层面的关联。文献[17]在难度评估和知识点预测等多个题目分析任务中取得了进步。进一步，Huang 等人[52]提出解耦的无监督题目表征模型 DisenQNet，在大规模题目语料上进行预训练，将题目的知识点等共性信息与难度等个性信息解耦分离并分别建模其表征向量，再通过最大化互信息等方式将预训练得到的表征模型应用于下游任务，有效提高了难度估计、相似性分析等下游任务的效果。

通过题目质量分析任务，模型可以模拟学习者的阅读和分析能力，可以在更复杂的应用任务中探索众多更为高级的类人认知能力。在接下来的各节中，本文将介绍机器阅读理解、数学题问答和文章自主评分等代表性智能应用研究。

3　机器阅读理解

机器阅读理解（Machine Reading Comprehension，MRC）是“人工智能＋教育”研究中利用自然语言处理等技术解决学科问答题的重要课题之一，旨在模拟学习者语言理解和语义分析等认知能力。相关研究可以追溯到 20 世纪 70 年代[53]。近年来，随着大型数据集的发布和自然语言处理技术的成熟，MRC 研究快速发展。尤其是斯坦福大学发布的 SQuAD[28]数据集，为 MRC 研究提供了一个优秀的研究和评测平台。2018 年 1 月阿里巴巴和 MSRA 的研究团队在 SQuAD 1.1 数据集上首次在

EM指标上超过人类。2019年3月哈工大讯飞联合实验室在SQuAD 2.0数据集上首次在精准匹配率(Exact Match,EM)和F_1值两个指标上都超过人类。截止目前,机器模型的效果已经大大超越人类结果①。

基于1.1节的介绍,MRC任务可以分为四个类型:完型填空,多项选择,片段抽取和自由回答[5]。通常来说,MRC任务要求模型阅读英语问答题上下文(输入),对提出的问题(输入)做出回答(输出)。其中,完型填空要求从候选单词表中选取问题中缺失的正确单词;多项选择要求判断每个候选项是否是问题的正确回答;片段抽取要求模型从文章中抽取一个连续的片段(即词组、短语等);自由回答要求模型根据词汇表生成片段作为问题回答。目前,相关的研究工作可以大致分为四类:基于规则的方法[26,53-54]、基于机器学习的方法[55-58]、基于深度学习的端到端方法[23,59-64]和预训练方法[51,65]。

基于规则的方法基于文本的语言学特征,人工设计大量规则解决英语问答题。例如,文献[53]设计了QUALM系统,考虑上下文语境,基于规则脚本与计划模拟人类理解故事的方式。文献[54]设计了Deep Read系统,基于传统自然语言处理技术(词干提取、词性识别、指代消解、词袋模型)的语言学特征设计规则检索文章中包含问题正确答案的句子。Richardson等人[26]在MCTest数据集上设计了两种基于规则的基线方法,即启发式滑动窗口方法和基于文本蕴涵系统的方法。基于机器学习的方法将MRC任务建模为监督学习问题,将题目的上下文文章和问题作为输入、答案作为标签,希望模型学习从文章和问题到答案的映射关系。研究者基于手工设计的语言学特征(如词性标签、指代关系、句法依赖等),在简单的最大间隔监督学习框架上建立模型[55-57]。Berant等人[58]将问题映射为一种正式查询,并使用大量人工设计的特征设计机器学习模型预测流程结构执行查询。

SQuAD等大型英语阅读数据集的出现满足了深度学习算法对数据的需求,促使大量基于深度学习的模型应用于MRC任务中[5]。此类方法不依赖已有工具或人工设计的特征,具有更高的准确性和泛化能力。目前,相关工作大致包括端到端方法[23,59-64]和预训练方法[51,65]。

端到端方法接收题目文章和问题作为输入,预测所有候选项上的概率分布或生成文本片段作为输出。例如,Hermann等人[23]发布大规模监督数据集CNN & Daily Mail,并在该数据集上设计基于注意力机制的长短时记忆网络(Long Short-Term Memory Network,LSTM)模型Attentive Reader,其性能相较传统方案有较大提升。Chen等人[59]在SQuAD数据集上设计了Stanford Attentive Reader,结合双向LSTM和注意力机制,基于题目中单词间的相似性预测答案位置,并将其扩展到其余三类MRC任务中。此后,BiDAF[60]从问题和文章的两个映射方向query-to-context和context-to-query上提高效果。AoA Reader[61]在双向注意力机制的基础上,建模注意力权重的重要性。R-Net[62]结合文章和问题间的注意力匹配和文章内的自注意力匹配对英语问答题进行建模,预测答案位置。MEMEN[63]利用记忆网络改善循环神经网络长距离依赖上的缺陷。文献[64]提出了一个多粒度框架来建模文档的结构特征。

预训练方法[51,65]首先通过掩码语言模型等预训练任务在大规模语料上训练一个表达能力强的语言模型,学习题目语义,再根据特定任务在规模较小的数据集上微调,提高模型在多任务上的效果。GPT[65]和BERT[51]是代表性预训练模型。其中,GPT是一个生成式预训练模型,使用Transformer模型[66]的解码器建模输入的前文信息。BERT使用Transformer模型的编码器建模输入的前后文信息。经过微调,GPT和BERT可以直接应用于MRC任务,显著提高效果。此后,Li等人[67]通过多任务学习的方法将领域掩码语言模型、自然语言推断和段落排序任务作为机器阅读理解的辅助任务对预训练语言模型进行微调,并集成多个预训练语言模型提高方法在机器阅读任务上的效果。为了解决跨语言的MRC任务,Cui[68]等提出了Dual BERT建模源数据和目标语言之间的关系。围绕长文本机器阅读理解问题,针对预训练语言模型只能接收固定长度输入的缺陷,Gong等人[69]提出一种基于强化学习的方法将文本动态分段,将每一段文本依次输入BERT模型获得答案。进一步,Luo等人[70]基于预训练语言模型,提出了一种问题生成与问题回答的协同学习框架,在少量标注数据上预训练模型之后,在大规模无标签文本语料上自动生成问题和回答问题,进一步提高模型的性能。

① https://rajpurkar.github.io/SQuAD-explorer/

表 1　部分代表性 MRC 模型的性能对比

数据集	CNN	Daily Mail	RACE	SQuAD	
指标	ACC	ACC	ACC	EM	F_1
文献[23]	63.0	69.0	—		—
	63.8	68.0	—	—	—
文献[71]	69.5	73.9	—	—	—
文献[59]	72.7	76.0	43.3	70.0	79.0
文献[72]	77.9	80.9	44.1	—	—
文献[61]	74.4	—	—	—	—
文献[73]	—	—	47.4	—	—
文献[74]	—	—	—	64.7	73.7
文献[75]	—	—	—	66.2	75.9
文献[60]	76.9	79.6	—	68.0	77.3
文献[76]	74.7	76.6	—	69.1	78.9
文献[63]	—	—	—	71.0	80.4
文献[77]	—	—	—	71.0	79.9
文献[62]	—	—	—	71.3	79.7
文献[51]	—	—	—	87.4	93.2

表 1 总结了部分代表性机器阅读理解模型的性能。总结而言，基于规则的方法和基于机器学习的方法具有较强的可解释性，但这些方法依赖大量人工设计规则和特征工程，依赖已有语言处理工具，模型的准确率有限，且难以泛化到大规模数据集中[78]。此外，如何从语言学特征中人工构建对 MRC 任务有效的规则或特征是一个巨大的挑战。基于深度学习的方法解决了模型准确性和泛化能力不足的问题，是目前常用的机器阅读理解方法，但基于深度学习的方法可解释性较低，且需要大量的标注数据进行训练，在 SQuAD 等大型数据集出现之后才逐渐兴起。预训练方法使用大规模无标签语料预训练，对标注数据的需求大大降低，且方法的准确率和泛化能力进一步提高，预训练语言模型能够捕获语料中的一些常识和领域知识，提高机器阅读理解任务的效果，目前已经成为机器阅读理解的主流方法之一。

除此之外，近期研究基于现实场景扩展 MRC 任务需求，模拟学习者更为复杂的语义分析能力，相关工作可以参考文献[5]。

4　数学题问答

数学题问答任务是“人工智能＋教育”融合研究的重要任务之一，相较于第 3 节介绍的 MRC 任务，要求模型进一步模拟学习者知识表达和逻辑推理等方面的认知能力，运用已掌握的知识推理出正确的答案[6]。

在相关研究中，数学应用题问答（Math Word Problem，MWP）是关注度最高的一类任务。该任务基于 1.2 节介绍的数学应用题数据集，要求模型阅读数学应用题题目文本（输入），推理生成相应的数学表达式（输出），计算答案。这个过程要求建立人类能够理解的自然文本与计算机能够理解的逻辑符号表达式的关联，需要模型具备（文本）语义理解、（数字）信息抽取、（符号）逻辑推导和（表达式）生成等方面的性能。数学应用题问答任务的研究历史较长，最早可以追溯到 20 世纪 60 年代。目前，相关研究方法可以大致分为三类：基于规则的方法[79-81]、基于语义解析的方法[82-85]和基于深度学习的方法[38,46,86-90]。

基于规则的方法是早期的研究方法。该方法依赖人工定义的模板匹配问题文本，并根据人工设计的规则通过简单的计算获得问题答案[79-80]。基于语义解析的方法将原始题目文本映射为特定的结构化逻辑形式，如语义解析树，再通过传统统计学习方法从逻辑形式中抽取数值变量并推理答案[82-84]。例如，Roy 等[84]提出表达式树的方法，将求解表达式的逐步推导转换为等价树结构的自底向上构建。

近年来，研究者借鉴深度学习在多个自然语言处理研究上的经验，将数学应用题问答表示为一类特殊的翻译任务。由于 seq2seq（Sequence-to-Sequence）方法具有很强的推理与生成新模板的能力，2017 年，腾讯公司的研究者[38]在 EMNLP 2017 会议上提出 DNS 模型，将 seq2seq 方法应用于 MWP 任务上。模型包括编码器与解码器两个模块，编码器通过循环神经网络将应用题文本自动编码为一个特征向量，解码器逐步将特征向量自动解码为数学表达式求解题目。在此基础上，最新的研究主要从增强问题理解能力（编码器模块）与表达式推理能力（解码器模块）两个方向进行改进。

在问题理解方面，为挖掘数学应用题文本的深层次信息，如数值语义、句子结构等，相关研究设计大量改进的方法。例如，文献[87]通过多种注意力

模型从题目文本中抽取不同类型的上下文信息，建模问题文本中不同句子间的关系。文献[89]提出的Graph2Tree模型挖掘问题文本中数值与词语、数值之间大小等关系，丰富对数值信息的理解。考虑到学习者阅读题目遵循逐句分析的层次化阅读习惯，文献[46]提出层次化模型HMS，在编码阶段将应用题划分成“字词—分句—问题”层次进行理解，并基于应用题的语法依赖结构增强语义。文献[90]额外引入常识知识图谱，挖掘融合知识的文本表征。为了融合不同编码器的优势，文献[91]提出了Multi-E/D模型，将通过基于序列的编码器挖掘得到的文本序列特征与通过基于图的编码器挖掘得到的语义结构特征进行结合，进而提高对问题的理解能力。

在表达式推理方面，文献[86]提出的T-RNN模型将表达式的解码分为两个阶段，第一阶段通过seq2seq方法生成仅包含数值的表达式框架，并转换为等价的表达式树。第二阶段根据每个缺失运算符的运算数，通过树结构的递归神经网络，生成每个缺失的运算符。文献[88]模拟学习者求解数学应用题过程中的目标分解过程，提出基于目标驱动的树结构的神经网络模型GTS，根据目标分解的过程自顶向下构造表达式树，在保证推理表达式合理性的同时具有较好的可解释性。文献[46]则在解码阶段设计了层次化树结构的指针网络，区分了数学表达式中不同类型符号的推理过程。与GTS不同，文献[92]提出的Seq2DAG模型采用自底向上的表达式树构建顺序，能够有效利用推理过程的中间步骤与子表达式，并实现满足交换律的运算过程。此外，为了增强模型对符号约束、对数值信息的利用，文献[93]和文献[94]引入了预测问题文本数值个数、数值位置、数值大小关系等额外任务，提高解题模型的推理能力。

此外，随着大规模预训练模型的发展，研究者将BERT[95-96]、RoBERTa[97]、BART[98]等预训练语言模型用于加强对应用题文本的理解能力。Huang等人[95]使用BERT初始化问题表征与问题类比模块中Transformer层的参数。Kim等[97]基于预训练的RoBERTa模型获得问题中词语的表征。Shen等[98]先将问题文本输入基于文本去噪任务预训练的BART模型获得表达式，再基于该模型对表达式的合理性进行评估，从而在这两种任务上对预训练模型进行微调。

表2总结了部分代表性MWP方法的性能对比。总结而言，早期的基于规则与基于语义解析的方法的答案生成过程具有更好的可解释性，但是需要大量人工构建的模板、规则、形式语言等，泛化能力不佳，应用范围有限。而深度学习模型具有较好的文本特征自动抽取能力与复杂表达式推理生成能力，但其求解过程难以解释，且缺乏对数理逻辑规则的运用。

表2 部分代表性MWP模型的性能对比

数据集	Dolphin	MAWPS	Math23K
指标	ACC	ACC	ACC
文献[84]	26.11	—	—
文献[99]	28.78	—	—
文献[85]	30.06	60.25	—
文献[38]	—	59.5	58.1
文献[86]	—	66.8	68.7
文献[87]	—	76.1	69.5
文献[88]	—	78.6	75.6
文献[46]	—	80.3	76.1
文献[89]	—	83.7	77.4
文献[90]	—	—	76.3
文献[92]	—	—	77.1
文献[93]	—	—	75.67
文献[94]	—	—	78.1
文献[91]	—	—	78.4
文献[95]	—	—	82.3
文献[98]	—	84.0	85.4
文献[97]	—	89.4	—

尽管MWP任务已经得到了长足的进展，但研究指出其任务的复杂程度仍处于初级阶段，仅符合小学基础教育的数学推理要求。2019年，DeepMind团队发布了一个数据集[42]，记录了大量数学简答题数据（如1.2节介绍），开始研究模型在更为高级和复杂的数学问答题上的求解能力。在这个任务中，除了传统MWP任务需要的能力之外，模型还需要具备（公式）理解、（变量）关联、（过程）记忆等复杂认知能力。目前该任务的研究进展较少，DeepMind团队尝试了多个基础的seq2seq模型。除此之外，Huang等人[100]考虑了简答题中多个公式的结构依赖关系，提出了融合图神经网络的求解模型，做出了一定的尝试。目前的研究证明求解该

数学简答题是一个更为困难的任务。

5　文章自主评分

文章自主评分（Automated Essay Scoring，AES）任务是智能教育研究的另一个重要任务，旨在模拟专家教师对文章进行打分，从而模拟人类对长篇文章的自主评测等认知能力。相较于上述研究任务，AES 对模型具有更高的要求。本节重点介绍针对主观题作文文章的自动评分。

基于 1.3 中介绍的相关数据集，该任务要求模型阅读题目要求和给定的文章，通过分析文章在整体或多个维度上的情况，给出相应的评分[7]。这个过程包括多个重要挑战，首先需要对词汇和语法等语句正确性进行检查；其次，需要对文章表达在连贯性、清晰度和说服力等维度进行评估。此外，还需要对文章在是否紧扣题目要求等相关度上进行检测。因此，AES 要求模型具备类人综合评价的复杂认知能力，吸引了国内外大量研究人员[7,101-102]。文章自主评分研究最早可以追溯到 20 世纪 60 年代[103]。近期的研究主要分为两类：基于特征构造的方法[104-106]和基于深度学习的模型设计[107-111]。

在 AES 中，大量研究关注如何构造有效的评测特征。主要想法是根据文章特点提取出对于分数评估有利的特征。例如，考虑到对复杂单词类别的运用可以体现出文章写作水平，文献[104]和文献[105]使用单词列表或字典将不同的单词分配到具体的词汇、句法或语义类别来构造单词分类特征。文献[106]使用语法树的深度来进行特征构造，从而评估句法的复杂程度。

近年来，研究者更加关注基于深度学习的 AES 方法。据笔者所知，文献[107]首次将深度神经网络应用到 AES 任务。模型首先将文章单词的独热码向量作为输入，使用一个卷积层来获得 n 元语法层级的特征，然后将这些特征输入 LSTM 中，最后拼接每一个语义特征向量，输出作文的评分。进一步，Dong 等[108]考虑到文章在词汇级和句子级上的层级结构特征，使用两个卷积-池化网络依次对不同级别的特征进行处理。考虑到在一个文章中不同的单词或者句子的重要性不同，Dong 等[109]对前期工作[108]进行了改进，使用注意力池化层代替最大池化层或平均池化层，增强语义关联，提高打分结果。此外，Tay 等[110]认为文章的连贯性和文章的整体分数有比较重要的相关性，从直觉上来说，连贯的句子间应该有比较强的相似性，因此，该文中使用一个全连接层网络，将从不同时间步长收集的 LSTM 两个位置的输出作为输入，并计算每对这样的位置输出的相似性。另外，考虑到数据集的限制，Lun 等[111]提出了一种数据增强的文章自主评分策略。此外，目前在 AES 研究中的一个趋势是使用 BERT[51]等预训练模型在相关的任务中进行微调。Liu 等[112]提出了两阶段的自动评分方式。在第一个阶段使用 BERT 模型获得语句的表征，并输入到一个循环神经网络中，在第二阶段加入手工特征增强效果。考虑到文章中语句的结构和通顺性的问题，Nadeem 等[113]设计了基于语句感知的辅助预训练任务。进一步，Yang 等[114]提出的 R2BERT 模型针对 AES 系统中的场景，采用多任务损失的方式来对 BERT 模型进行微调。近期，Wang 等人[115]认为在真实的评估场景中，教师通常会从多个角度来对文章进行评估，因此作者基于 BERT 模型学习多尺度特征；其次，考虑到训练数据稀少的问题，作者使用迁移学习的策略从非当前领域的文章中学习领域相关的知识。

表 3 总结了部分代表性 AES 模型的性能对比。总的来说，基于特征构造的方法具有更强的可解释性，但是需要对手工特征进行精心设计来获得较好的表现。而与之相对的，基于深度学习的模型设计能缓解手工特征构造时的困难，但是缺乏对结果的可解释性。

表 3　部分代表性 AES 模型的性能对比

数据集	ASAP	SemEval-2013		
指标	Avg QWK	ACC	M-F_1	W-F_1
文献[107]	0.761	—	—	—
文献[108]	0.734	—	—	—
文献[109]	0.764	—	—	—
文献[110]	0.764	—	—	—
文献[111]	—	0.828	0.823	0.827
文献[112]	0.773	—	—	—
文献[114]	0.794	—	—	—
文献[115]	0.791	—	—	—

目前，AES 研究正逐渐从对文章的整体评测转向对某些特定维度的评测，例如连贯性、说服力或者是否符合主旨等，以及研究如何解决数据标注稀缺问题带来的挑战，具有广阔的前景。

6 开源工具与代码

本节将对上述四类研究任务涉及的重要开源工具或代表性模型代码进行介绍，如表4所示。

表4 代表性开源工具简介

任务	名称	简介
题目质量分析	EduNLP	专注学科题目分析，具有题目文本分析、结构识别、公式解析、题目表征等功能，且提供多种预训练模型
MRC	SogouMRCToolkit 文献[116]	提供多种已发布的经典MRC模型，及测评结果
MWP	MWPToolkit 文献[117]	模块化MWP求解流程，提供多种已发布的MWP模型与测评结果
AES	EASE	提供多种特征构造和回归模型选择
	ESCRITO 文献[118]	提供多层API测评写作得分的NLP工具包

首先，EduNLP[①]是题目质量分析基础任务的工具包，专注于学科题目的语法语义分析，包含题目结构识别、题目分词、公式解析、语义向量化等功能，并提供多种预训练模型。

在MRC任务中，代表性的开源代码包括Attentive Reader[23]、BiDAF[60]、RCM[69]、文献[64]等。此外，MRC任务的常用工具包之一是SogouMRCToolkit[116]。该工具包提供BiDAF、R-Net等多种已发布的经典MRC模型，及其在三个数据集(SQuAD 1.0，SQuAD 2.0，CoQA)上的测评结果。此外，该工具包提供读取数据集、处理数据、构造模型的相关接口，方便开发者快速有效地开发机器阅读模型。

在MWP任务中，代表性开源代码包括GROUP-ATT[87]、GTS[88]、HMS[46]、Graph2Tree[89]、KA-S2T[90]、NS-Solver[93]、NumS2T[94]、Multi-E/D[91]、REAL[95]、Generate & Rank[98]。此外，MWPToolkit[117]是MWP任务的开源工具库，其包含了经典的DNS、Seq2Tree、Graph2Tree，以及基于预训练的BERTGen、GPT-2等共17个模型，并测试了它们在6个常见数据集上的结果。MWPToolkit将现有MWP求解模型解耦为高度可重用的模块，从而能够支持开发者进行数据读取、数据处理、模型构造、超参数搜索、模型评估等操作。

针对AES任务，文献[107]、文献[109]、文献[113]、文献[115]等提供了文章自主评分方法的开源实现。此外，EASE[②]是AES任务中常见的开源系统之一，提供多种手工特征构造的方法以及多个回归函数的选择，例如支持向量回归(Support Vector Regression，SVR)和贝叶斯线性岭回归(Bayesian Linear Ridge Regression，BLRR)。开发者可以在此基础上实现自己的AES模型。此外，ESCRITO[118]是一个评测学生写作能力的NLP工具包，包含了供教师使用的高层封装API以及供开发者使用的基础开发API。

7 未来研究方向

利用自然语言处理技术模拟人类学习过程，是“人工智能+教育”的重要研究方向。目前，通过研究题目质量分析、机器阅读理解、数学题问答和文章自主评分等智能应用任务，模型在题目理解、语义分析、知识推理和自主评测等认知能力上取得了重要进展。然而，相比于人类的教育学习过程，现有研究的类人学习过程仍较为简单，性能仍有待改进。本节将对未来可行的研究方向进行简单介绍。

7.1 多模态学科题目理解

学科题目数据是一种多模态数据，大部分学科题目中除自然语言文本外，还包含图像等多模态数据和公式、表格等异构数据，例如数学几何题目包括题目文本、几何图形等异构信息，这些数据需要利用多模态分析技术进行处理。此外，智能教育系统提供了大量与学科题目相关的课程视频等教育资源，有助于对学科题目的准确分析和理解。因此，结合图像、公式、课程视频等多模态数据，融合相关的课程知识，可以对学科题目资源和学习者的教育学习活动做更加完整的分析。目前，已有研究针对多模态题目数据进行了初步探索。例如，文献[119]融合课程分析与题目分析结果，为学科题目寻找匹配的课程片段内容。然而，现有研究尚未直接探索人工智能算法在学习课程和解决问题等视觉与思维能力融合方面的机理。

7.2 教育知识图谱

教育学习过程离不开对知识的分析、记忆、归纳

① https://github.com/bigdata-ustc/EduNLP

② https://github.com/edx/ease

和推理。在这个过程中，学习者可以形成一个属于自己的知识库，为运用知识解决相关问题提供了基础。因此，如何从课程、课本、题目等教育资源中构造教育知识图谱，具有重要的学术和应用价值。首先，教育知识图谱可以帮助计算机有效管理知识，可以提高多个智能教育任务（如机器阅读、智能问答等）的效果[120]。其次，教育知识图谱的构造对于模拟人类知识库的形成、研究类人知识推理能力具有积极意义。目前已有研究[121-122]分别从课程资源和多源文本资源（课本、题目和网页百科等）构造了包含先序关系和相关关系等多关系教育知识图谱。然而，相关研究仍处于初级阶段，具有研究前景。

7.3　可解释性与因果分析

本文介绍的相关研究内容大多关注于提升具体任务的效果，例如，在机器阅读理解任务中，模型在 EM 和 F_1 值两个指标上的效果已经超过人类。然而，研究仍表明相关模型对结果的解释性较差。当前，研究模型的可解释性已经成为相关研究的重点。其中，融合教育学理论，探索人类学习规律，指导模型学习过程，对模型结果进行因果分析，是现有研究有待突破的难点和重要的研究方向。

8　结束语

智能教育系统积累了大量学科题目数据，为“人工智能＋教育”融合方向的研究提供了可能，相关研究对于探索为模型赋能人类复杂智慧有积极意义。本文介绍了面向学科题目的文本分析方法与应用，重点对题目质量分析、机器阅读理解、数学题问答和文章自主评分等任务的研究进展进行简述。此外，本文还对相关数据集和开源工具等进行了介绍。最后，本文展望了未来研究方向。

参考文献

[1] 焦李成. 下一代人工智能的挑战与思考[J]. 智能系统学报，2020，15(06)：1185-1187.

[2] Roll I，Wylie R. Evolution and revolution in artificial intelligence in education[J]. International Journal of Artificial Intelligence in Education，2016，26(2)：582-599.

[3] Fuchs L S，Fuchs D，Hamlett C L，et al. Effects of expert system consultation within curriculum-based measurement：Using a reading maze task[J]. Exceptional Children，1992，58(5)：436-450.

[4] Hontangas P，Ponsoda V，Olea J，et al. The choice of item difficulty in self-adapted testing[J]. European Journal of Psychological Assessment，2000，16(1)：3.

[5] Liu S，Zhang X，Zhang S，et al. Neural machine reading comprehension：Methods and trends[J]. Applied Sciences，2019，9(18)：3698.

[6] Zhang D，Wang L，Zhang L，et al. The gap of semantic parsing：A survey on automatic math word problem solvers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2019，42(9)：2287-2305.

[7] Ke Z，Ng V. Automatedessay scoring：A survey of the state of the art[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence，2019：6300-6308.

[8] 黄振亚. 面向个性化学习的数据挖掘方法与应用研究[D]. 合肥：中国科学技术大学博士学位论文，2020.

[9] Wu F，Qiao Y，Chen J H，et al. Mind：A large-scale dataset for news recommendation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics，2020：3597-3606.

[10] Xu G，Meng Y，Qiu X，et al. Sentiment analysis of comment texts based on BiLSTM[J]. IEEE Access，2019，7：51522-51532.

[11] Huang W，Chen E，Liu Q，et al. Hierarchical multi-label text classification：An attention-based recurrent network approach[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management，2019：1051-1060.

[12] Liu Q，Huang Z，Huang Z，et al. Finding similar exercises in online education systems[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2018：1821-1830.

[13] Liu Q，Tong S，Liu C，et al. Exploiting cognitive structure for adaptive learning[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2019：627-635.

[14] Huang Z，Liu Q，Chen E，et al. Question difficulty prediction for READING problems in standard tests[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence，2017：1352-1359.

[15] Qiu Z，Wu X，Fan W. Question difficulty prediction for multiple choice problems in medical exams[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management，2019：139-148.

[16] Xue K，Yaneva V，Runyon C，et al. Predicting the difficulty and response time of multiple choice questions using transfer learning[C]//Proceedings of the

15th Workshop on Innovative Use of NLP for Building Educational Applications, 2020: 193-197.

[17] Yin Y, Liu Q, Huang Z, et al. Quesnet: A unified representation for heterogeneous test questions[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019: 1328-1336.

[18] 黄仔. 基于多模态学习的试题建模方法与应用研究[D]. 合肥: 中国科学技术大学博士学位论文, 2019.

[19] Pelánek R. Measuring similarity of educational items: An overview[J]. IEEE Transactions on Learning Technologies, 2019, 13(2): 354-366.

[20] Rihák J, Pelánek R. Measuring similarity of educational items using data on learners' performance[C]//Proceedings of the 10th International Conference on Educational Data Mining, 2017: 16-23.

[21] 刘淇, 陈恩红, 朱天宇, 等. 面向在线智慧学习的教育数据挖掘技术研究[J]. 模式识别与人工智能, 2018, 31(01): 77-90.

[22] Liu Q, Huang Z, Yin Y, et al. Ekt: Exercise-aware knowledge tracing for student performance prediction[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(1): 100-115.

[23] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching machines to read and comprehend[J]. Advances in Neural Information Processing Systems, 2015, 28: 1693-1701.

[24] Hill F, Bordes A, Chopra S, et al. The goldilocks principle: Reading children's books with explicit memory representations[J/OL]. arXiv preprint arXiv:1511.02301, 2015.

[25] Suster S, Daelemans W. CLICR: A dataset of clinical case reports for machine reading comprehension[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 1551-1563.

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

[27] Lai G, Xie Q, Liu H, et al. RACE: Large-scale reading comprehension dataset from examinations[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 785-794.

[28] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAd: 100,000+ questions for machine comprehension of text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016: 2383-2392.

[29] Trischler A, Wang T, Yuan X, et al. NewsQA: A machine comprehension dataset[C]//Proceedings of the 2nd Workshop on Representation Learning for NLP, 2017: 191-200.

[30] Joshi M, Choi E, Weld D S, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1601-1611.

[31] Saha A, Aralikatte R, Khapra M M, et al. DuoRC: Towards complex language understanding with paraphrased reading comprehension[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1683-1693.

[32] Cui Y, Liu T, Che W, et al. Aspan-extraction dataset for Chinese machine reading comprehension[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 5883-5889.

[33] Bajaj P, Campos D, Craswell N, et al. MSMARCO: A human generated machine reading comprehension dataset[J/OL]. arXiv preprint arXiv: 1611.09268, 2016.

[34] 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.

[35] Dunn M, Sagun L, Higgins M, et al. SearchQA: A new q&a dataset augmented with context from a search engine[J/OL]. arXiv preprint arXiv: 1704.05179, 2017.

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

[37] Huang D, Shi S, Lin C Y, et al. How well do computers solve math word problems? Large-scale dataset construction and evaluation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 887-896.

[38] Wang Y, Liu X, Shi S. Deep neural solver for math word problems[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 845-854.

[39] Koncel-Kedziorski R, Roy S, Amini A, et al. MAWPS: A math word problem repository[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 1152-1157.

[40] Amini A, Gabriel S, Lin S, et al. MathQA: Towards interpretable math word problem solving with operation-based formalisms[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 2357-2367.

[41] Miao S Y, Liang C C, Su K Y. Adiverse corpus for evaluating and developing English math word problem solvers[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 975-984.

[42] Saxton D, Grefenstette E, Hill F, et al. Analysing mathematical reasoning abilities of neural models[J/OL]. arXiv preprint arXiv:1904.01557, 2019.

[43] Yannakoudakis H, Briscoe T, Medlock B. A new dataset and method for automatically grading ESOL texts[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011: 180-189.

[44] Dzikovska M O, Nielsen R, Brew C, et al. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge[C]//Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics: Proceedings of the 7th International Workshop on Semantic Evaluation, 2013: 263-274.

[45] Blanchard D, Tetreault J, Higgins D, et al. TOEFL11: A corpus of non-native English[J/OL]. ETS Research Report Series, 2013,(2): 1-15. https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/j.2333-8504.2013.tb02331.x[2022-07-18].

[46] Lin X, Huang Z, Zhao H, et al. Hms: A hierarchical solver with dependency-enhanced understanding for math word problem[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021: 4232-4240.

[47] Alagumalai S, Curtis D D. Classical test theory[M]. Dordrecht: Springer, 2005: 1-14.

[48] DiBello L V, Roussos L A, Stout W. Review of cognitively diagnostic assessment and a summary of psychometric models[J]. Handbook of Statistics, 2006, 26:970-1030.

[49] Beck J, Stern M, Woolf B P. Using the student model to control problem difficulty[C]//Proceedings of the User Modeling, 1997: 277-288.

[50] Kubinger K D, Gottschall C H. Item difficulty of multiple choice tests dependant on different item response formats-an experiment in fundamental research on psychological assessment[J]. Psychology Science, 2007, 49(4): 361.

[51] Devlin J, Chang M W, Lee K, et al. Bert: Pretraining of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.

[52] Huang Z, Lin X, Wang H, et al. DisenQNet: Disentangled representation learning for educational questions[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021: 696-704.

[53] Lehnert W G. The process of question answering[D]. PhD diss., New Haven: Yale University, 1977.

[54] 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, 1999: 325-332.

[55] Sachan M, Dubey K, Xing E, et al. Learning answer-entailing structures for machine comprehension[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 239-249.

[56] Narasimhan K, Barzilay R. Machine comprehension with discourse relations[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 1253-1262.

[57] Wang H, Bansal M, Gimpel K, et al. Machine comprehension with syntax, frames, and semantics[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 700-706.

[58] Berant J, Srikumar V, Chen P C, et al. Modeling biological processes for reading comprehension[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014: 1499-1510.

[59] Chen D, Fisch A, Weston J, et al. Reading wikipedia to answer open-domain questions[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1870-1879.

[60] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional attention flow for machine comprehension[J/OL]. arXiv preprint arXiv:1611.01603, 2016.

[61] Cui Y, Chen Z, Wei S, et al. Attention-over-attention neural networks for reading comprehension[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 593-602.

[62] Wang W, Yang N, Wei F, et al. Gated self-matching networks for reading comprehension and question answering[C]//Proceedings of the 55th Annual Meet-

ing of the Association for Computational Linguistics, 2017: 189-198.

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

[64] Zheng B, Wen H, Liang Y, et al. Document modeling with graph attention networks for multi-grained machine reading comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 6708-6718.

[65] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J/OL]. OpenAI Blog, 2019, 1(8): 9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf[2022-07-18].

[66] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[67] Li H, Zhang X, Liu Y, et al. D-net: A pre-training and fine-tuning framework for improving the generalization of machine reading comprehension[C]//Proceedings of the 2nd Workshop on Machine Reading for Question Answering, 2019: 212-219.

[68] Cui Y, Che W, Liu T, et al. Cross-lingual machine reading comprehension[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 1586-1595.

[69] Gong H, Shen Y, Yu D, et al. Recurrent chunking mechanisms for long-text machine reading comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 6751-6761.

[70] Luo H, Li S W, Yu S, et al. Cooperative learning of zero-shot machine reading comprehension[J/OL]. arXiv preprint arXiv:2103.07449, 2021.

[71] Kadlec R, Schmid M, Bajgar O, et al. Text understanding with the attention sum reader network[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 908-918.

[72] Dhingra B, Liu H, Yang Z, et al. Gated-attention readers for text comprehension[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1832-1846.

[73] Xu Y, Liu J, Gao J, et al. Dynamic fusion networks for machine reading comprehension[J/OL]. arXiv preprint arXiv:1711.04964, 2017.

[74] Wang S, Jiang J. Machine comprehension using match-lstm and answer pointer[J/OL]. arXiv preprint arXiv:1608.07905, 2016.

[75] Xiong C, Zhong V, Socher R. Dynamic coattention networks for question answering[J/OL]. arXiv preprint arXiv:1611.01604, 2016.

[76] 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, 2017: 1047-1055.

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

[78] Chen D. Neural reading comprehension and beyond[D]. PhD diss., San Francisco: Stanford University, 2018.

[79] Fletcher C R. Understanding and solving arithmetic word problems: a computer simulation[J]. Behavior Research Methods, Instruments, & Computers, 1985, 17(5): 565-571.

[80] Yuhui M, Ying Z, Guangzuo C, et al. Frame-based calculus of solving arithmetic multi-step addition and subtraction word problems[C]//Proceedings of the 2nd International Workshop on Education Technology and Computer Science, 2010: 476-479.

[81] Mukherjee A, Garain U. A review of methods for automatic understanding of natural language mathematical problems[J]. Artificial Intelligence Review, 2008, 29(2): 93-122.

[82] Hosseini M J, Hajishirzi H, Etzioni O, et al. Learning to solve arithmetic word problems with verb categorization[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014: 523-533.

[83] Liang C C, Wong Y S, Lin Y C, et al. A meaning-based statistical English math word problem solver[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 652-662.

[84] Roy S, Roth D. Solving general arithmetic word problems[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 1743-1752.

[85] Wang L, Zhang D, Gao L, et al. Mathdqn: Solving arithmetic word problems via deep reinforcement learning[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018, 5545-5552.

[86] Wang L, Zhang D, Zhang J, et al. Template-based math word problem solvers with recursive neural networks[C]//Proceedings of the 33rd AAAI Confer-

ence on Artificial Intelligence，2019：7144-7151.

[87] Li J，Wang L，Zhang J，et al. Modeling intra-relation in math word problems with different functional multi-head attentions[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics，2019：6162-6167.

[88] Xie Z，Sun S. A goal-driven tree-structured neural model for math word problems[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence，2019：5299-5305.

[89] Zhang J，Wang L，Lee R K W，et al. Graph-to-tree learning for solving math word problems[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics，2020：3928-3937.

[90] Wu Q，Zhang Q，Fu J，et al. A knowledge-aware sequence-to-tree network for math word problem solving[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2020：7137-7146.

[91] Shen Y，Jin C. Solving math word problems with multi-encoders and multi-decoders[C]//Proceedings of the 28th International Conference on Computational Linguistics，2020：2924-2934.

[92] Cao Y，Hong F，Li H，et al. A bottom-up DAG structure extraction model for math word problems[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence，2021：39-46.

[93] Qin J，Liang X，Hong Y，et al. Neural-symbolic solver for math word problems with auxiliary tasks[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing，2021：5870-5881.

[94] Wu Q，Zhang Q，Wei Z，et al. Math word problem solving with explicit numerical values[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing，2021：5859-5869.

[95] Huang S，Wang J，Xu J，et al. Recall and learn：A memory-augmented solver for math word problems[C]//Proceedings of the Association for Computational Linguistics，2021：786-796.

[96] Liang Z，Zhang J，Shao J，et al. Mwp-bert：A strong baseline for math word problems[J/OL]. arXiv preprint arXiv:2107.13435，2021.

[97] Kim H，Hwang J，Yoo T，et al. Improving a graph-to-tree model for solving math word problems[C]//Proceedings of the 16th International Conference on Ubiquitous Information Management and Communication，2022：1-7.

[98] Shen J，Yin Y，Li L，et al. Generate and rank：A multi-task framework for math word problems[C]//Proceedings of the Association for Computational Linguistics，2021：2269-2279.

[99] Roy S，Roth D. Unit dependency graph and its application to arithmetic word problem solving[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence，2017：3082-3088.

[100] Huang Z，Liu Q，Gao W，et al. Neural mathematical solver with enhanced formula structure[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval，2020：1729-1732.

[101] 付瑞吉，王栋，王士进，等. 面向作文自动评分的优美句识别[J]. 中文信息学报，2018，32(06)：88-97.

[102] 何屹松，孙媛媛，张凯，等. 计算机智能辅助评分系统定标集选取和优化方法研究[J]. 中国考试，2020(01)：30-36.

[103] Page E B. The imminence of grading essays by computer[J]. The Phi Delta Kappan，1966，47(5)：238-243.

[104] Yannakoudakis H，Briscoe T. Modeling coherence in ESOL learner texts[C]//Proceedings of the 7th Workshop on Building Educational Applications Using NLP，2012：33-43.

[105] Farra N，Somasundaran S，Burstein J. Scoring persuasive essays using opinions and their targets[C]//Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications，2015：64-74.

[106] Chen H，He B. Automated essay scoring by maximizing human-machine agreement[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2013：1741-1752.

[107] Taghipour K，Ng H T. A neural approach to automated essay scoring[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2016：1882-1891.

[108] Dong F，Zhang Y. Automatic features for essay scoring：an empirical study[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2016：1072-1077.

[109] Dong F，Zhang Y，Yang J. Attention-based recurrent convolutional neural network for automatic essay scoring[C]//Proceedings of the 21st Conference on Computational Natural Language Learning，2017：153-162.

[110] Tay Y，Phan M，Tuan L A，et al. Skipflow：incorporating neural coherence features for end-to-end automatic text scoring[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence，2018：

5948-5955.

[111] Lun J, Zhu J, Tang Y, et al. Multiple data augmentation strategies for improving performance on automatic short answer scoring[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020: 13389-13396.

[112] Liu J, Xu Y, Zhu Y. Automated essay scoring based on two-stage learning[J/OL]. arXiv preprint arXiv: 1901.07744, 2019.

[113] Nadeem F, Nguyen H, Liu Y, et al. Automated essay scoring with discourse-aware neural models [C]//Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications, 2019: 484-493.

[114] Yang R, Cao J, Wen Z, et al. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking [C]//Proceedings of the Association for Computational Linguistics, 2020: 1560-1569.

[115] Wang Y, Wang C, Li R, et al. On the use of BERT for automated essay scoring: joint learning of multi-scale essay representation [J/OL]. arXiv preprint arXiv:2205.03835, 2022.

[116] Wu J, Yang Y, Deng C, et al. Sogou machine reading comprehension toolkit [J/OL]. arXiv preprint arXiv:1903.11848, 2019.

[117] Lan Y, Wang L, Zhang Q, et al. Mwptoolkit: An open-source framework for deep learning-based math word problem solvers[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence, 2022: 13188-13190.

[118] Zesch T, Horbach A.Escrito: An NLP-enhanced educational scoring toolkit [C]//Proceedings of the 11th International Conference on Language Resources and Evaluation, 2018: 2310-2316.

[119] Wang X, Huang W, Liu Q, et al. Fine-grained similarity measurement between educational videos and exercises[C]//Proceedings of the 28th ACM International Conference on Multimedia, 2020: 331-339.

[120] Chen P, Lu Y, Zheng V W, et al. Knowedu: A system to construct knowledge graph for education [J]. IEEE Access, 2018, 6: 31553-31563.

[121] Pan L, Li C, Li J, et al. Prerequisite relation learningfor concepts in MOOCS[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1447-1456.

[122] Huang X, Liu Q, Wang C, et al. Constructing educational concept maps with multiple relationships from multi-source data [C]//Proceedings of the IEEE International Conference on Data Mining, 2019: 1108-1113.

黄振亚(1992—),博士,副研究员,主要研究领域为数据挖掘、文本挖掘、知识推理、教育大数据分析。
E-mail: huangzhy@ustc.edu.cn

陈恩红(1968—),博士,教授,主要研究领域为机器学习与数据挖掘、社会网络、个性化推荐等。
E-mail: cheneh@ustc.edu.cn

刘淇(1986—),通信作者,博士,教授,主要研究领域为数据挖掘与知识发现、机器学习方法。
E-mail: qiliuql@ustc.edu.cn