

大语言模型驱动的题库自动生成：面向PDF的跨模态解析与知识点对齐

司 彬

(萍乡市第二人民医院 信息科, 江西 萍乡 337000)

摘要: 随着在线教育的快速发展, 高质量题库的自动化生成需求日益迫切。现有题库自动生成方法主要分为两类: 基于教材内容生成的方法存在准确率不足的问题, 而基于PDF题库解析的方法由于格式多样性导致普适性较差。为此, 提出一种结合知识图谱与混合Embedding的题库生成系统。采用多模态PDF解析技术处理文本、公式等异构内容; 构建知识点图谱, 实现题目—答案—知识点的语义关联; 设计混合匹配算法, 融合结构特征与语义特征进行精准对齐。实验结果表明, 所提系统在问题—答案解析匹配任务中达到70.68%的准确率, 在题目—知识点对齐任务中达到82.95%的准确率。该系统具有良好的普适性, 与教师审核机制结合后可快速应用于CFA考试、医师资格考试等专业领域。

关键词: 题库自动生成; 大语言模型; 知识图谱; 跨模态解析; 语义匹配

DOI: 10.11907/rjdk.251489

中图分类号: TP182

文献标识码: A

文章编号: 1672-7800(2026)004-0104-06

扫描二维码阅读全文: 

Automated Question Bank Generation Driven by Large Language Models: Cross-Modal Parsing and Knowledge Point Alignment for PDF Documents

SI Bin

(Department of Information Technology, Pingxiang Second People's Hospital, Pingxiang 337000, China)

Abstract: With the rapid development of online education, the automatic generation of high-quality question banks is increasingly urgent. The existing methods for automatic generation of question bank can be divided into two categories: the method based on textbook content has insufficient accuracy, while the method based on PDF question bank analysis has poor universality due to the diversity of formats. Therefore, a question bank generation system combining knowledge map and hybrid embedding is proposed, which uses multimodal PDF parsing technology to process text, formula and other heterogeneous content; Construct a map of knowledge points to realize the semantic association of question answer knowledge points; A hybrid matching algorithm is designed to fuse structural features and semantic features for precise alignment. The experimental results show that the proposed system achieves an accuracy rate of 70.68% in the question answer analysis matching task and 82.95% in the question knowledge point alignment task. The system has good universality and can be quickly applied to professional fields such as CFA examination, physician qualification examination, etc. after being combined with the teacher review mechanism.

Key Words: automated question generation; large language models; knowledge graph; cross-modal parsing; semantic matching

0 引言

在教育数字化与在线学习迅速发展的背景下, 如何高效获取优质题库已成为教育机构和平台的重要课题。自动生成题库作为一种新兴技术手段, 能够显著提升教育资源的可及性和利用效率, 降低人工命题成本, 并支持个

性化学习。通过算法快速生成海量题目可覆盖不同难度和知识点, 满足多样化教学需求, 同时为在线教育平台和考试系统提供动态内容支持。

目前, 题库自动生成与智能出题技术备受关注, 其主要分为规则驱动、机器学习驱动和知识图谱驱动3类方法。例如, Flor等^[1]利用规则生成试题, 虽然语法正确但适应性不足。Rao等^[2]引入统计语言模型, 提升了题目多样

收稿日期: 2025-07-27

作者简介: 司彬(1988-), 男, 萍乡市第二人民医院信息科工程师, 研究方向为医院数据中心机房运维、信息安全和系统运维。

性,但缺乏语义关联。Srivastava等^[3]提升了生成流畅性,但因缺乏知识结构建模,难以覆盖核心知识点。Gabajiwala等^[4]虽改进了生成质量,但可解释性不足。知识图谱为题库自动生成提供了新思路。例如,Li等^[5]基于知识图谱生成考题,实现了较好的知识覆盖,但在多模态处理和大规模推理效率上存在瓶颈。Liu等^[6]利用领域知识图谱结合实体关系推理提升了答案准确性,但未充分利用图谱与大语言模型的协同优势。总体来说,题库自动生成技术目前仍存在以下问题:一是生成题目的准确性和逻辑性不足,尤其是复杂推理题易出现语义错误;二是对多模态内容(如图表、公式)的处理能力有限,导致STEM领域题目生成效果较差;三是缺乏统一的评估标准,难以量化题目质量与教育目标的匹配度。

针对以上不足,本文提出一种基于知识图谱的题库自动生成系统,其创新点在于:①融合多模态内容识别与结构化存储,构建保留语义逻辑的知识图谱;②结合联合嵌入与检索增强生成(Retrieval-Augmented Generation, RAG)技术增强推理,依托大语言模型与图谱数据库的高效协同,实现高准确性、可解释性的题目生成,并验证其在教育场景中的可扩展性。本文研究重点分为两个部分:一是PDF识别与格式化,包括题目、答案解析和教材;二是题目—答案解析、题目—知识点匹配。现有PDF识别研究已能处理通用场景^[7-9]、图表^[10-11]和公式^[12]。然而,由于题目和答案解析通常分属不同文档,识别后存在粘连和答案省略问题,精准分割与匹配至关重要。本文通过依据教材目录构建知识图谱来辅助匹配^[13-16];同时采用知识图谱与文本语义的混合匹配方法,以解决单纯语义匹配准确率低下的问题^[17-18]。此外,系统利用大模型应用平台Dify完成问题答案修补、知识点提炼等工作^[19-20]。

1 系统设计

如图1所示,基于知识图谱的题库答案解析自动生成系统采用分层架构设计,通过模块化组件实现从教材解析到智能出题的全流程自动化。系统首先通过输入层处理多模态教材内容,其中图片型PDF识别模块专门提取扫描版教材中的图文混合内容,并保留原始版式信息;文本型PDF识别模块则负责解析结构化文本与公式,并支持语义分段标注。数据层负责完成知识的结构化存储与增强,Markdown构建器将原始输入转化为标准Markdown格式以保留标题、列表等逻辑结构。QA文档解析模块拆分教材中的问答对并关联知识点标签;教材解析模块提取章节、图表、术语等元数据以构建知识图谱实体关系。系统的核心算法层采用知识图谱驱动的语义匹配引擎,基于知识图谱的Embedding匹配算法通过联合训练文本嵌入(如BERT)与图谱嵌入(如TransE),利用余弦相似度计算问题—答案匹配度,并支持多跳推理功能。QA增强模块则基于RAG技术,从知识库中检索相关片段注入大语言模型提示词,有效减少生成幻觉。在应用层,系统面向教育场

景提供智能出题功能,能够根据知识点权重动态生成选择题和填空题,并支持难度调节,同时自动生成解题步骤与知识拓展说明。教师审核面板提供人工干预接口,支持题目修正与反馈闭环,确保输出质量。系统依托Dify平台实现 workflow编排,通过整合大语言模型(GPT-4)与知识图谱数据库(Neo4j),确保高扩展性与多源数据兼容性。整体架构通过模块化设计实现功能解耦,各层之间通过标准化接口进行数据交互,体现了系统设计的灵活性与可维护性。该架构不仅能有效进行教材内容的结构化解析与知识提取,而且能通过智能算法实现高质量的题目生成和答案验证,为教育领域提供高效的自动化解决方案。

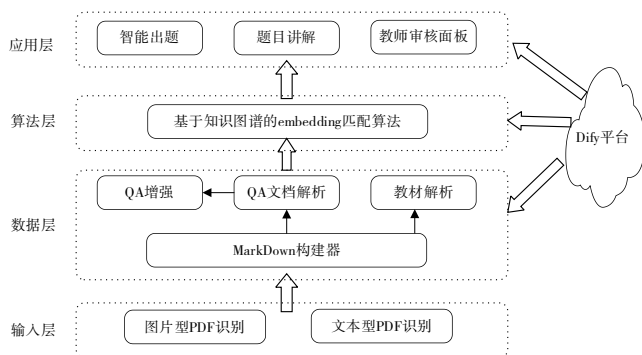


Fig. 1 An automatic generation system architecture of question bank answer analysis based on knowledge graph

图1 基于知识图谱的题库答案解析自动生成系统架构

1.1 文档输入

系统输入分为PDF文本文档和PDF图片文档两大类。从业务层面来看,输入文档包括题目PDF、答案解析PDF、教材PDF 3种,其中题目PDF可能在每道题之后都有答案解析。由于PDF文档结构化程度不高、答案省略或图片清晰度不足等原因,将这些题目与答案解析进行匹配需要花费大量精力。因此,若要使题库自动生成系统通用化,题目—答案解析匹配、题目—知识点匹配这两项工作必不可少,系统需要引入一些先进的匹配策略。

1.2 数据处理

1.2.1 Markdown构建器

为保障题库内容的丰富性与可扩展性,本文系统采用Markdown格式作为标准化存储方案。相较于原始PDF文档,Markdown格式具有显著优势:首先,该格式具备跨平台编辑能力,可在各类文本编辑器中灵活修改,展现出优异的扩展性;其次,该格式能完整保留题目中的图表等结构化数据,确保题目内容的准确性;最后,其纯文本特性支持全文检索、元数据标注及版本控制功能,相较于需要光学字符处理(Optical Character Recognition, OCR)处理的扫描版PDF文档,显著提升了内容检索效率。如图2所示,左侧为Markdown格式的题目,表格与正文合并时中间要单独加一行空行,否则表格在前端无法被正常展示;右侧为Markdown的显示效果。

在具体实现层面,系统采用差异化存储策略:对于表

An analyst collects the following information about a company and its capital structure: <table border="1"> <tr><td>Proportion of debt</td><td>40%</td></tr> <tr><td>Proportion of equity</td><td>50%</td></tr> <tr><td>Proportion of preferred shares</td><td>10%</td></tr> <tr><td>Before-tax cost of debt</td><td>8%</td></tr> <tr><td>Cost of equity</td><td>16%</td></tr> <tr><td>Cost of preferred shares</td><td>10%</td></tr> <tr><td>Marginal tax rate</td><td>25%</td></tr> </table> If interest expenses are tax deductible, the company's WACC is closest to: https://demo.cfagpt.org/oss/transfer-file/bo_d19op7v7aaic73896/	Proportion of debt	40%	Proportion of equity	50%	Proportion of preferred shares	10%	Before-tax cost of debt	8%	Cost of equity	16%	Cost of preferred shares	10%	Marginal tax rate	25%	An analyst collects the following information about a company and its capital structure: <table border="1"> <tr><td>Proportion of debt</td><td>40%</td></tr> <tr><td>Proportion of equity</td><td>50%</td></tr> <tr><td>Proportion of preferred shares</td><td>10%</td></tr> <tr><td>Before-tax cost of debt</td><td>8%</td></tr> <tr><td>Cost of equity</td><td>16%</td></tr> <tr><td>Cost of preferred shares</td><td>10%</td></tr> <tr><td>Marginal tax rate</td><td>25%</td></tr> </table> If interest expenses are tax deductible, the company's WACC is closest to: Exhibit 22: Histogram of Daily Trading Volume for a Stock for One Year	Proportion of debt	40%	Proportion of equity	50%	Proportion of preferred shares	10%	Before-tax cost of debt	8%	Cost of equity	16%	Cost of preferred shares	10%	Marginal tax rate	25%
Proportion of debt	40%																												
Proportion of equity	50%																												
Proportion of preferred shares	10%																												
Before-tax cost of debt	8%																												
Cost of equity	16%																												
Cost of preferred shares	10%																												
Marginal tax rate	25%																												
Proportion of debt	40%																												
Proportion of equity	50%																												
Proportion of preferred shares	10%																												
Before-tax cost of debt	8%																												
Cost of equity	16%																												
Cost of preferred shares	10%																												
Marginal tax rate	25%																												

Fig. 2 Markdown-based storage format for question banks and teaching materials

图2 基于Markdown的题库与教材存储格式

格类数据直接以 Markdown 原生语法存储;针对图片资源,采用“占位符+对象存储”的混合方案,即在正文中嵌入图片引用标识,实际图像文件则存储于 Minio 等对象存储服务中。该设计既保持了文档的轻量化特性,又确保了多媒体资源的可管理性。

1.2.2 QA 文档解析

图3为QA文档解析流程。以图片型PDF的识别流程为例,首先按页面对PDF文件进行拆分,其次对页面进行并行处理。每一页使用不同的工具与参数进行识别,识别后按坐标对各种素材进行排序组装。文本型PDF的识别过程与之类似,只是使用的工具不同。表1为PDF页面中每种模态的识别工具。与单一工具相比,本文采用的多模态工具组装方式具备更好的内容还原度。

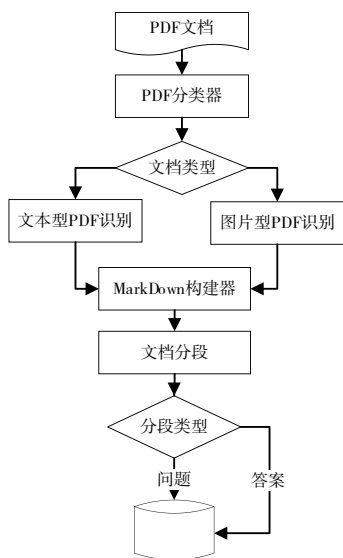


Fig. 3 QA document parsing process

图3 QA文档解析流程

Table 1 Multimodal PDF page recognition tools

表1 PDF页面多模态识别工具

PDF 类型	表格识别工具	公式识别工具	文本识别工具
文本型	pdfplumber	pdfminer	pdfplumber
图片型	OpenCv	pix2text	pytesseract

PDF 文档被识别后,使用 Markdown 构建器按照特定格式进行组织。获得 Markdown 文档后需使用常用的分段标识对文档进行分段存储,并利用 QA 增强技术对分段进行修复,使每段内容都是独立的 1 个问题或答案,最终得到 Markdown 格式的问题列表和答案列表。

1.2.3 QA 增强

QA 增强指自动化分段后的答案和问题完整性修复。基于 Dify 平台的大模型通过 RAG 技术来实现精准修复,具体流程为:将识别后的完整 Markdown 文档全部导入 Dify 平台构建知识库,建立高效的检索索引。对于每个待修复的问题或答案,系统会将其作为 Query 输入到 RAG 模块中进行语义检索,从原始题库中找出最相关的上下文片段。在修复过程中,系统会将 3 个关键要素组合后输入大模型,分别为:①RAG 检索得到的最相关 Context 片段;②待修复的原始 Query(问题或答案);③预先设计好的修复提示词模版。这种设计充分发挥了大模型的语言理解和生成能力,同时又通过原始题库文档作为主要知识来源,有效规避了通用大模型在特定专业领域可能产生的幻觉或虚构问题。这种以原始文档为主导、大模型为辅助的协同工作机制既保证了修复后内容的专业性和准确性,又提升了修复效率,实现了对不完整或存在缺陷的问题和答案的高质量修补。整个流程严格依赖原始题库提供的上下文,确保修复结果与领域知识保持高度一致。QA 增强流程见图4。

1.2.4 教材解析

如图5所示,基于目录的教材解析方法采用结构化处理流程,主要步骤为:首先对教材目录进行深度解析,将其转换为层次分明的树形数据结构。在树形结构中,每个叶子节点代表一个独立的知识点单元,这些叶子节点构成了知识体系的最小粒度;其次系统会结合目录中标注的页码信息,将完整的教材 Markdown 文档按照知识点边界进行智能分割,具体实现方式为以每个知识点对应的内容范围作为分割依据,在 Markdown 文档中精准定位起始和结束位置,确保分割后的每一段内容都完整对应一个叶子节点的知识点。完成分割后,系统会对这些结构化处理后的知

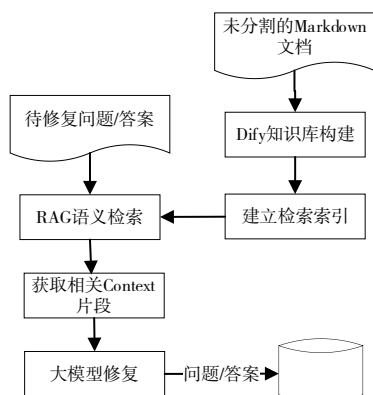


Fig. 4 QA enhancement process

图4 QA增强流程

识点数据进行持久化存储:一方面将知识点内容及其元数据(如所属章节、页码等)存入关系型数据库,便于后续精确检索与管理;另一方面将这些知识点内容同步存入RAG系统的知识库,建立高效的向量索引,为后续的语义检索和知识问答提供支持。这种目录驱动的文档解析方法实现了教材内容从非结构化到结构化的转换,既保留了原始知识的完整性,又为后续的知识检索与应用奠定了良好的数据基础。整个过程严格遵循目录的层级关系和知识点划分,确保知识组织的准确性和一致性。

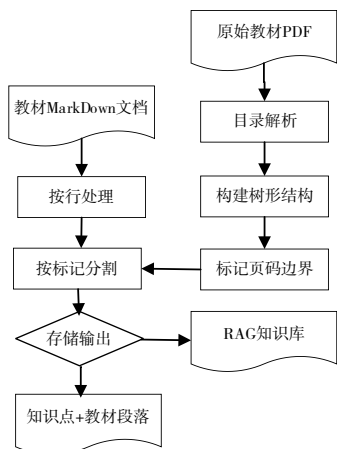


Fig. 5 Textbook parsing flow

图5 教材解析流程

1.3 基于知识图谱的Embedding匹配算法

1.3.1 Embedding

Embedding是一种将高维数据(如文本、图像、类别等)映射到低维连续向量空间的技术,其核心思想为使相似对象在向量空间中距离更近,从而方便机器学习模型理解和处理数据。Embedding能将离散的单词或文本转换成稠密的数值向量,以便进行语义计算、相似度比较等任务。Transformer模型BERT通过自注意力机制计算上下文相关的Embedding,其中自注意力的计算公式为:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

式中: Q 、 K 、 V 分别为查询、键、值矩阵; d_k 为向量维度。

BERT的掩码语言建模目标函数为 $\log P(w_i | context)$,其通过预测被掩盖的词 w_i 来学习上下文相关的Embedding。通过大规模语料训练的双向Transformer能对语言中的词和句子建模,并在下游任务中经微调后即可应用于文本分类、问答系统、命名实体识别和语言推理等。BERT的输入嵌入由以下3个部分组成:①Token Embeddings。每个输入的词被拆分为子词单元(WordPiece),使用一个可学习的lookup表将这些token映射为向量;②Segment Embeddings。由于BERT的pre-training包括NSP(Next Sentence Prediction)任务,每个token都会加上一个标识:句子A $\rightarrow$ segment 0;句子B $\rightarrow$ segment 1,用于区分两个句子;③Position Embeddings。BERT并没有使用循环神经网络,因此需要加入位置信息,每个token的位置信息由一个可学习的embedding进行编码,最终 $\text{InputEmbedding}(i) = \text{TokenEmbedding}(i) + \text{SegmentEmbedding}(i) + \text{PositionEmbedding}(i)$ 。此外,BERT使用多层(Base: 12层;Large: 24层)Transformer Encoder,每层通过Multi-head Self-Attention机制捕捉所有token之间的关系,可以获得更完整的上下文。如表2所示,与传统的Word Embedding(如GloVe、Word2vec)相比,BERT提供的Embedding可以根据任务和上下文动态变化,大幅提升了其在多项自然语言处理任务中的性能。总而言之,BERT的Embedding不再是每个词对应一个固定向量,而是每个词在不同上下文中都有不同表达。系统通过Token+Segment+Position的嵌入设计,结合深度Transformer编码器,实现了强大的语言理解能力。

Table 2 Comparison of major embedding approaches

表2 多种主流嵌入方法比较

方法	是否上下文相关	是否双向	是否可微调
Word2vec	否	否	否
ELMo	是	是(LSTM)	否(有限)
BERT	是	是	是

1.3.2 知识图谱

教材蕴含了整个垂直领域的知识图谱,本文系统利用教材知识图谱来辅助进行语义匹配。例如,教材中有“勾股定理描述了直角三角形三边的关系,而毕达哥拉斯定理是同一原理的西方命名”的描述,目的是将“勾股定理”“毕达哥拉斯定理”“直角三角形”这三者对齐,即在向量空间中使之相邻。当拿到测试文本时,可以利用知识图谱相关性更准确地匹配同一个物体但叫法不同的情况。具体实现过程为:首先使用主流的bert-base-NER模型对Markdown格式的教材进行实体识别,并储存到知识图谱数据库Neo4j中;其次利用BERT相似实体对齐机制将相似实体对齐,将每个实体三元组输入BERT模型中生成Embedding结果;最后将Embedding结果存入向量数据库Weaviate(基于Faiss)中。在实际实现中可以使用HuggingFace的Transformers库加载dslim/bert-base-NER预训练模型,也可以根据教材语言选择相应模型。

1.3.3 知识点对齐

知识点对齐包括两个部分:第一部分为题目与答案解析的匹配,第二部分为题目与知识点讲解的匹配。题目和答案解析含有的知识量相对较少,直接对其进行匹配的准确率较低。因此,本文对题目和答案解析进行2次匹配,第一次直接匹配题目和答案解析,得到分数 α ;第二次采用题目和答案解析分别获得教材知识图谱的匹配向量,此次匹配得到分数 β , α 与 β 之和最大的即为正确的匹配对。此外,本文系统还增加了一些剪枝逻辑,例如问题 q 与答案解析 a 通过知识图谱的embedding检索发现属于同一个知识点,则提高 q 与 a 匹配的权重;再如限制每个问题最多匹配一个答案解析、每个答案解析最多匹配一个问题。题目与知识点的匹配则直接取知识图谱Embedding命中的知识点,然后将该知识点对应的Markdown文本输入大模型进行简单总结。

2 验证实验

2.1 实验环境与参数设置

实验在配备NVIDIA A100 GPU的服务器上进行。嵌入模型采用轻量级句子嵌入模型all-MiniLM-L6-v2,为题目、答案及知识点文本生成384维语义向量。大语言模型采用GPT-4,用于生成知识点讲解摘要。知识图谱构建以“题目”“答案”“知识点”为实体,以“题目—答案配对”“题目—知识点归属”“知识点—子知识点”为关系,并基于上述嵌入模型为所有实体生成向量表示。采用混合匹配策略,总分 $S=(\alpha+\beta)*\omega$,其中 α 为题目与答案向量的余弦相似度; β 由二者Top-3知识点是否有交集决定(有则为1.0,无则为0);若题目与答案共享知识点,则权重系数 $\omega=1.2$,否则 $\omega=1.0$ 。每个数据集的输入PDF顺序均被随机打乱,独立运行3轮以评估鲁棒性。

2.2 数据集

使用3个不同特性的真实题库数据集进行实验,具体信息见表3。

Table 3 Information of the datasets

表3 数据集信息

数据集	题目数量	题目—答案解析 组织方式	题目—知识点 标注情况	内容描述
数据集1	1 900	分离式(题目与答案 在不同PDF文件)	已由专家标注	CFA考试题目 与解析
数据集2	900	一体式(题目后紧跟 答案于同一PDF)	未标注	软件设计师 历年真题
数据集3	598	一体式(题目后紧跟 答案于同一PDF)	已标注 (基于章节)	软件设计师 章节习题

2.3 评估指标

采用以下4个定量指标评估系统性能:①问题—答案匹配准确率。表示系统为题目正确匹配答案解析的比例,计算公式为 $QAA=\text{正确匹配的题目—答案对数}/\text{题目总数}$;②题目—知识点对齐准确率。表示系统为题目正确关联知识点的比例,计算公式为 $QKA=\text{被正确关联知识点的题目数}/\text{题目总数}$;③题目召回率。表示从PDF中成功提取题目的比例,计算公式为 $QR=\text{成功提取的题目数}/\text{PDF中实际题目总数}$;④答案解析召回率。表示从PDF中成功提取答案解析的比例,计算公式为 $AR=\text{成功提取的答案解析数}/\text{PDF中实际答案解析总数}$ 。其中,QAA与QKA为评估语义匹配效果的核心指标,QR与AR用于评估底层数据解析的质量。

2.4 评估结果与分析

系统在各数据集的评估结果见表4。可以看出,系统在所有数据集上均表现良好,题目与答案的召回率(QR、AR)均高于92%,表明数据解析层可靠。在核心语义任务上,问题—答案匹配准确率(QAA)超过70%,题目—知识点对齐准确率(QKA)超过82%。系统在数据集2和3上的QAA显著高于数据集1,主要原因在于一体式排版使系统在解析阶段即可直接配对,极大降低了匹配难度。对于最具挑战性的数据集1,其匹配失败案例主要归因于:①答案解析过短(字符数<50)。占比67.2%,文本信息过少可导致语义向量区分度不足;②题目语义高度相似。占比21.55%,同一知识点下的题目表述相近,易造成混淆;③其他因素。如知识点划分过细等。系统在数据集3的QKA高于数据集1,这是由于数据集3以“章”为知识单元,粒度较粗,而数据集1采用细粒度知识点,概念间相似性高,增加了区分难度。在为每个问题寻找最匹配答案时,可采用混合匹配策略(结合直接语义匹配 α 与知识图谱匹配 β)并应用剪枝逻辑(如强制一对一匹配、知识点权重提升等)解决上述问题。

Table 4 Overall experimental results of the system

表4 系统总体实验结果

(%)

指标	数据集1	数据集2	数据集3
QAA	70.68	99.40	97.16
QKA	82.95	/	90.61
QR	99.79	92.00	98.83
AR	98.55	92.56	97.99

随机打乱PDF输入顺序进行的3轮实验表明,系统的QAA与QKA的波动在 $\pm 1\%$ 以内,表现稳定。如表5所示,系统采用的组合解析工具在文本识别、表格还原、公式代码保留及多模态关联方面均显著优于单一OCR工具(如Tesseract)。

如表6所示,相较于传统的“从教材生成题目”的方法,本文系统能直接处理题目PDF并进行智能改编,在题目准确性、逻辑严谨性和多模态支持上更具优势。

Table 5 PDF parsing tools comparison

表 5 PDF 解析工具比较

工具	文本识别 准确率/%	表格结构 还原度/%	公式/代码保留 原生支持	多模态关联 语义关联
PDF 组合 解析工具	98.23	95.14	LaTeX/Markdown 原生支持	文本、公式、图表 语义关联
单一工具 tesseract	90.72	70.65	纯文本,符号易丢失	仅独立识别各元素

Table 6 Comparative analysis with traditional automatic question bank generation method

表 6 与传统题库自动生成方法的比较

方法	题目准确 率	题目逻辑 严谨性	多模态支 持度	知识点匹 配度	审核要求
本文系统	高	高	高	一般	一般
从教材生成题目	一般	一般	低	极高	极高

3 结 语

本文提出基于知识图谱的题库自动生成系统,实现了教材解析、知识建模与智能出题的全流程自动化。系统融合多模态识别与结构化存储技术,构建语义逻辑一致的知识图谱,并结合联合嵌入与 RAG 增强推理算法,显著提升了题目生成的准确性与可解释性。依托 Dify 平台,该系统展现出良好的扩展性与实用价值。然而,该系统尚存在不足之处:其一,知识图谱构建依赖教材质量与标注精度,对跨领域、多语种教材的适应性有待验证;其二,混合推理在大规模数据下的效率与资源消耗需优化;其三,PDF 题目特征的多样性与规则自适应机制尚不完善。后续可探索跨领域预训练与多语种处理以提升适应性,引入图神经网络与增量更新机制以优化推理效率,并结合学习者画像实现个性化出题,进一步增强系统的智能化与精准性。

参考文献:

[1] FLOR M, RIORDAN B. A semantic role-based approach to open-domain automatic question generation [C]// Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, 2018: 254–263.

[2] RAO P R, JHAWAR T N, KACHAVE Y A, et al. Generating QA from rule-based algorithms [C]// International Conference on Electronics and Renewable Systems (ICEARS), 2022: 1697–1703.

[3] SRIVASTAVA A, SHINDE S, PATEL N, et al. Questionator-automated question generation using deep learning [C]// International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020: 1–5.

[4] GABAJIWALA E, MEHTA P, SINGH R, et al. Quiz maker: automatic quiz generation from text using NLP [C]// Singapore: Proceedings of 4th International Conference on Futuristic Trends in Network and Communication Technologies, 2022: 523–533.

[5] LI D, LI Z, YANG Y, et al. Knowledge graph-enhanced large language model for domain-specific question answering systems [DB/OL]. <https://www.techrxiv.org/inst/26407>.

[6] LIU L, ZHOU Y, MA J, et al. Domain-specific question-answering systems: a case study of a carbon neutrality knowledge base [J]. Sustainability, 2025, 17(5): 2192.

[7] LIU Y, ZHANG H, WANG L, et al. DeepLayout: a multi-modal document layout analysis framework for complex PDFs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(8): 4123–4136.

[8] CHEN X, SMITH A, BROWN B. PDF-TREX: an automated pipeline for table extraction and reconstruction from PDF documents [J]. ACM Transactions on Document Systems, 2021, 15(3): 1–24.

[9] SMITH A, JOHNSON B, LEE C. A comprehensive survey on PDF document parsing: from layout analysis to knowledge extraction [J]. ACM Computing Surveys, 2023, 55(4): 1–36.

[10] CHEN R, WANG P, LI S, et al. ChartSense: automated data extraction from PDF charts via deep layout segmentation [J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(6): 2875–2888.

[11] LIU Y, CHEN Q, WU X F. Deep learning-based table recognition in PDFs: a survey [J]. Acta Automatica Sinica, 2022, 48(8): 1845–1862.

刘洋, 陈强, 吴晓峰. 基于深度学习的 PDF 表格识别技术综述 [J]. 自动化学报, 2022, 48(8): 1845–1862.

[12] LI H, ZHANG L, LIU Y. Joint extraction of cross-modal text and formulas in academic PDFs [J]. Journal of Software, 2022, 33(5): 1789–1804.

李航, 张磊, 刘洋. 学术 PDF 中跨模态文本与公式的联合提取方法 [J]. 软件学报, 2022, 33(5): 1789–1804.

[13] LI H, WANG W, ZHANG Q. Chinese textbook question answering system based on knowledge graph [J]. Journal of Chinese Information Processing, 2022, 36(4): 78–89.

李华, 王伟, 张强. 基于知识图谱的中文教材问答系统构建 [J]. 中文信息学报, 2022, 36(4): 78–89.

[14] WANG X, GAO T, ZHU Z, et al. KEPLER: a unified model for knowledge embedding and pre-trained language representation [C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020: 4175–4186.

[15] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition [C]// Proceedings of NAACL-HLT, 2016: 260–270.

[16] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of NAACL-HLT, 2019: 4171–4186.

[17] LEWIS P, PEREZ E, PIKULIK R, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [J]. Advances in Neural Information Processing Systems, 2020, 33: 9459–9474.

[18] XIONG W, HOANG T, WANG W Y. DeepPath: a reinforcement learning method for knowledge graph reasoning [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 564–573.

[19] DIFY Team. Dify: an open-source LLM application development platform [EB/OL]. <https://docs.dify.ai/>.

[20] DIFY Team. Dify community edition documentation [EB/OL]. <https://docs.dify.ai/community-edition>.

(责任编辑:尹晨茹)