难度可控的词义辨析选择题自动生成

刘廷超[1]，王雨[1]，荀恩东[2†]
[1]北京语言大学 信息科学学院
[2]北京语言大学 语言科学与资源学院
liutingchao@hotmail.com, 1213546256@qq.com, edxun@126.com

摘要

针对汉语词义辨析选择题的自动生成任务，本文提出一种基于检索增强生成（RAG）技术的智能出题框架。该框架通过构建融合词汇等级、词频与句子长度的多维度难度评估模型，实现习题难度的个性化控制。研究通过整合语言要素知识库与BCC语料库，有效提升语境自然性与干扰项质量，并引入格式校验、逻辑验证与答案唯一性检测的多维校验机制，确保输出题目符合教学规范。实验结果显示，该方法在出题成功率、答案正确率与内容多样性等关键指标上显著优于传统微调模型，展现出良好的教学适配性与应用潜力，为汉语教学智能化提供新的技术路径。

关键词： 自动出题；大语言模型；检索增强生成

Difficulty-controlled automatic generation of lexical meaning discrimination multiple-choice questions

Tingchao Liu[1], Yu Wang[1], Endong Xun[2*]
[1]School of Information Science, Beijing Language and Culture University
[2]School of Linguistics and Language Resources, Beijing Language and Culture University
liutingchao@hotmail.com, 1213546256@qq.com, edxun@126.com

Abstract

This paper proposes an intelligent question generation framework based on Retrieval-Augmented Generation (RAG) technology, focusing on the automatic generation of Chinese word sense disambiguation multiple-choice questions. By constructing a difficulty estimation method that integrates vocabulary levels, word frequency, and sentence length, the system enables personalized control of question difficulty. Leveraging a linguistic element database and the BCC corpus, the framework enhances contextual naturalness and distractor quality, and incorporates a multi-dimensional validation mechanism to ensure format consistency and answer uniqueness. Experimental results show that the proposed method outperforms traditional fine-tuned models in terms of generation success rate, accuracy, and content diversity, demonstrating strong adaptability to teaching scenarios and practical value.

Keywords: Automatic Question Generation , Large Language Models , Retrieval-Augmented Generation

† 通讯作者

1 引言

词汇是语言的基本构成单元，其音、形、义构成核心属性。然而，仅掌握这些属性并不足以实现语言的流畅运用，学习者还需理解词汇在不同语境中的具体使用方式。因此，如何全面评估学习者的词汇知识成为语言教学与测试的核心问题。

针对这一问题，研究者们从不同角度对词汇知识进行了系统划分。(Nation and Nation, 2001) 在二语习得研究中，将词汇知识归纳为八个方面，强调其多维特性；(苏向丽and 李如龙, 2011) 结合汉语特点，将词汇知识细分为形式、认知、结构、语法、语用和语义六个层面；(胡韧奋, 2017) 进一步提出，汉语二语学习者的词汇知识可归纳为形式、语法、语义、语用四个核心维度，涵盖同音词、搭配关系、近义辨析和文体风格等内容，以更全面地评估学习者的词汇掌握情况。这些研究为构建科学的词汇测试体系提供了理论支撑。

在词汇测试研究中，核心问题在于如何界定测试内容。传统词汇测试长期聚焦于词义辨析与词汇量测量(Heaton, 2000; 张凯, 2002)，而(刘润清and 韩宝成, 2007) 指出，词汇测试不应局限于单一维度，应该结合语法、语用等因素，全面反映学习者的语言能力。最新研究表明，词汇运用受语音、语法等多方面的影响，单纯考察词义存在局限性，因此测试设计应涵盖多个维度，以确保其科学性和有效性(崔希亮, 2023) 。因此，词汇测试的难度应根据学习阶段的不同进行等级划分，以提高其适应性和针对性(胡韧奋, 2017) 。

作为词汇测试的重要题型，选择题通过题干与干扰项的设计实现对词义辨析能力的考察。编写单项选择题时，必须确保每题只有一个正确答案，干扰项应具备有效干扰作用，这是编写单项选择题的核心原则，同时也是最容易出错的地方(刘超英, 2022)。在实际测试中，该题型的设计存在两大核心挑战：一方面，人工出题依赖教师的经验，主观性较强，且缺乏统一标准，导致试题质量和难度控制的不稳定性；另一方面，现有的题目难度划分不够精准，不同学习者的认知水平差异较大，如何实现精准匹配仍然是一个难题。此外，传统的词汇测试大多依赖固定题库，缺乏动态调整能力，难以满足个性化学习需求。在习题设计过程中，题干和干扰项的选择至关重要，特别是干扰项的设计。优质干扰项应具备较高的混淆度但不引起歧义，这对出题者的语言敏感度和教学经验提出了更高要求。因此，如何高效且精准地生成高质量的词义辨析选择题，依然是亟待解决的问题。

近年来，随着自然语言处理技术的快速发展，智能化语言测试系统成为学术研究与教育实践的重要方向。当前，自动出题方法主要包括规则匹配、统计学习、深度学习与检索增强生成四大类。规则匹配（如预设词库、模式匹配）依赖固定规则进行试题生成，适用于结构化内容，但灵活性受限，难以处理复杂的语言现象；统计学习方法通过基于N-gram 语言模型、词频统计或TF-IDF 计算选项和填空内容，具备一定的泛化能力，但仍然难以精准控制试题难度和适应复杂语境；深度学习方法（如Seq2Seq、Transformer 模型）能够端到端自动生成习题，但由于缺乏知识约束，可能导致语法错误、答案不唯一或试题不符合教学要求；相比之下，RAG 技术融合了知识检索与文本生成的优势，能够动态调用高质量语料支持试题生成，同时结合知识库实现精准难度控制，使生成的习题更具针对性和可解释性，并支持基于不同知识点的多样化试题设计，从而显著提升测试覆盖面和适应性。

在此背景下，本研究设计习题难度计算方法，并基于RAG 技术，针对词义辨析选择题的智能生成展开深入研究，重点优化习题难度控制、质量保障，提出以下创新方案：

（1）可控难度的词义辨析习题生成任务设计 将“可控难度”作为中文词义辨析选择题自动生成的核心目标之一，明确提出习题生成不仅要符合教学需求和语言规范，还需满足设定的认知难度等级。构建了覆盖词汇等级、词频、句长与干扰项强度的综合任务设置，支持不同语言水平学习者的适配性出题需求。

（2）融合词汇等级与语境复杂度的习题难度计算方法 设计了一套结合词语等级、词频映射、句长调节与高难词惩罚项的综合难度评估模型，分别用于题干和选项的难度分析。该方法支持未登录词处理、语境语法复杂度评估，并通过线性插值计算整题难度，系统化地实现了中文词义辨析类习题的定量化难度建模，提升了个性化推荐与测评精准度。

（3）基于RAG的难度可控习题生成机制 构建了一个融合参数提取、语料检索、生成控制与质量校验的完整RAG出题流程，将RAG框架系统性引入中文词义辨析题的出题场景。通过引入语言要素库与BCC语料库支持上下文构建，结合高质量检索增强补全策略，显著提升了出题成功率、内容多样性与难度匹配度；并通过引入答案唯一性校验机制与迭代修复流程，有效解决了以往多答案、格式混乱等问题，保障习题质量。

第二十四届中国计算语言学大会论文集，第1页–第15页，济南，中国，2025年8月11日至14日。

（4）面向智能教学的习题质量校验与控制机制 提出“质量驱动型生成”策略，引入答案唯一性校验、多轮生成迭代与难度偏差调整机制，构建了一个能自主检测并优化出题结果的生成闭环，显著提升自动生成习题在教学系统中的稳定性、可靠性与评估效度。

2 背景

自动出题（Automatic Question Generation，AQG）是教育技术领域的重要研究方向，旨在利用计算机技术自动生成符合教学需求的试题，以减轻教师负担，提高测评的科学性与效率。其研究发展可以分为以下四个阶段。

基于规则匹配与模板的自动出题：最早的自动出题方法主要依赖规则匹配和模板填充，即通过预设规则或句式模板构建习题。这类方法具备一定的可控性和准确性，但灵活性和多样性受限，难以适应不同层次学习者的需求。例如，(Mitkov and others, 2003) 利用WordNet 生成多项选择题，通过规则匹配和语义网络提取干扰项，以提高测试的精准度。(Brown et al., 2005) 亦采用WordNet 进行词义知识测试，探索自动生成词义辨析题的可行性。然而，该方法严重依赖人工设定规则，难以处理复杂语言现象，导致生成的试题质量和难度控制能力有限，多样性也相对较低。

基于统计学习与知识库的智能出题：由于规则匹配方法难以灵活适应复杂语境，研究者引入了统计学习与本体知识库，以增强自动出题的泛化能力和智能化水平。例如，(Takuya et al., 2010) 采用统计机器学习方法，通过学习人工命题数据优化词句选择策略。(Liu et al., 2005) 提出结合词义消歧技术的填空题生成方法，通过搭配计算筛选干扰项，以提高词汇测试的精准度。(Papasalouros et al., 2008) 提出基于领域本体的自动出题方法，利用OWL 本体语言与自然语言生成技术提取问题和答案，但仍面临句法正确性不足的问题。

在汉语自动出题研究中，由于汉语的同音、近音、形近词现象及构词法特性，词义辨析比英语更加复杂。因此，自动出题系统需结合汉语特有的语言现象，并利用语言要素库等高质量语言资源优化干扰项生成，以提高测试的科学性和有效性。例如，(唐奇峰, 2012) 通过描述逻辑定义领域概念及其关系，并提出基于概念相似度的选项生成策略，以优化干扰项的合理性。(丁向民, 2008) 则结合知识点分类与本体干扰项生成方法，利用本体表示概念间的层级关系，实现多项选择题的自动生成。此外，(李钢, 2013) 融合教育测量理论，优化试题难度计算方法，从而提高习题难度的精准控制能力。

基于深度学习与大规模语料的智能出题：随着深度学习和自然语言处理技术的发展，自动出题研究进入基于深度学习和大规模语料库的阶段。例如，(Susanti et al., 2015) 提出了一种自动生成英语词汇测验的方法，以TOEFL 词汇题为模型，利用互联网检索文本，并结合WordNet 词典生成阅读材料、正确答案及干扰项。(Satria and Tokunaga, 2017) 提出一种基于非限制性关系从句的自动英语指代问题生成方法，通过句子拆分技术，将人类编写的文本转换为阅读理解测试题，包括目标代词、正确答案及干扰项；(Wang et al., 2024)构建了一个适用中国民法的综合性自动问答数据集，可以提供法律问题答案；(Bitew et al., 2024)研究了利用大量的答案和混淆项集合从而智能重建利用现有干扰项生成新的多项选择题的方案。

在汉语词汇测试领域，(胡韧奋, 2017) 基于二语习得理论，强调词汇知识的多维性（包括词义、搭配、用法、频率等），并采用数据驱动方法，结合自然语言处理技术和大规模语料库分析，自动生成汉语词汇测试题，以提升测试效率并降低人工出题的主观性和局限性。

基于预训练语言模型的智能出题优化：随着预训练语言模型和大规模语料库的发展，研究者开始探索基于大语言模型的智能化出题方法，以优化试题的难度控制、质量和多样性。

在难度控制方面，(Jiang and Lee, 2017) 研究了中文填空题的干扰项自动生成，基于词性、难度、拼写相似度、词共现等标准评估干扰项质量，实验表明，基于word2vec 语义相似度的干扰项更具迷惑性。(温雪峰et al., 2020) 提出基于语义相似性的选择题自动生成方法，采用图优化算法筛选题目，以减少重复问题并提高测试效率。

在多样性提升方面，(Zhang, 2022) 提出阅读理解选择题的自动生成方法，分别采用基于T5 Transformer 的端到端方法和基于序列学习的元序列表示方法，实验结果表明，这两种方法均能生成语法正确、结构合理的试题，展现了深度学习在自动出题中的应用潜力。

(Ngo et al., 2024)经过实验认为仅基于ChatGPT3.5不适合用于多项选择题的生成；(Patil et al., 2024)提出了一种结合大型语言模型与语义检索机制的自动生成专业领域问题的新方法，用于实现特定主题下的高质量问答生成。为解决教师手工出题成本高、效率低等问题，(Wang

et al., 2025)提出一种融合教师实践经验的提示词（prompt pattern）模板设计方法，构建自动出题系统（AQG），提升大语言模型（LLM）出题的专业性、准确性与教育适配性；(Hang et al., 2024)通过将大语言模型和检索增强和高级提示工程技术结合，利用广泛的外部知识库，尝试了高效生成多选题的方案。(Huang et al., 2024)尝试了利用两阶段框架并结合神经语言模型和遗传算法来生成问题组的办法，可以用来帮助教师准备课堂习题；(Fahad et al., 2024)尝试了通过微调从而给资源匮乏语言生成题目的混淆项的可能性；(Shoaib et al., 2025)利用条件生成对抗网络（cGANs），构建了一种灵活的方法，能够在不同熟练程度和主题范围内生成具备多样性与上下文关联性的选择题；(Shwe et al., 2024)提出了用于创建具有难度级别的多项选择题框架，可以提供带有难度信息的选项；(Alawwad et al., 2025)提出了一种利用检索增强技术处理跨领域场景的教科书问答（TQA）的框架，可以应对更复杂的教育场景。

当前，自动出题研究已拓展至多模态融合、个性化学习和智能反馈等方向，未来结合深度学习与强化学习，自动出题系统将增强语义理解，依据学习者特征定制个性化试题，进一步提升测评的智能化水平。

3 习题难度等级

习题难度评估是词汇教学中实现教学分级与个性化的重要基础。科学的计算方法有助于教师为不同语言水平的学习者匹配适宜内容，提升学习效果与成就感。为实现更精准的测评目标，本文提出融合题干与选项的综合难度计算方法，提升词义辨析选择题的评估精度，为自动出题与个性化教学提供支撑。

3.0.1 词语等级评估

选择题由题干与选项构成，其基本单位为词语，故词汇等级是习题难度的关键指标。本文采用教育部中外语言交流合作中心2021年7月发布的《国际中文教育中文水平等级标准》词汇表作为评估依据。在计算过程中，系统对题干与选项进行分词并等级标注，采用加权方式融合各部分难度，输出综合等级。词语提取使用THULAC 工具(Li and Sun, 2009)，等级匹配基于词汇表，未收录词汇则结合词频进行估算，以保证评估的完整性与准确性。

3.0.2 词频分析与等级映射

对于未被等级词汇表收录的词语，本研究使用wordfreq 库[1] 提供的词频信息作为依据，构建词频与等级的映射表。通过统计每一级别词语的平均词频范围，建立词频区间与等级的对应关系，将词频分布区间划分为8 个等级：1级对应频率$(6.98\times10^{-4},1]$，2级为$(1.60\times10^{-4},6.98\times10^{-4}]$，以此类推，超纲词频低于$1.73\times10^{-5}$。

为验证该方法的有效性，本研究对部分未在等级词汇表中出现的词语进行了频率提取与等级预测。以“这个”“每天”“健身房”等词为例，其词频与预测等级如表1 所示，结果显示该映射关系具有较高的合理性。

表 1: 等级映射示例

词语	频率值	预测等级
这个	2.0000×10^{-3}	1级
每天	1.5500×10^{-4}	3级
健身房	4.5700×10^{-6}	超纲

通过词语等级与词频映射机制的结合，不仅提升了习题难度计算的覆盖范围，也增强了对未登录词的处理能力，为词义辨析选择题的自动生成与难度调控提供了数据支撑。

3.0.3 题干难度计算

在词义辨析类选择题中，题干通常由一个完整的句子构成，其难度不仅取决于组成词语的等级，还受到高难词分布与句子长度等因素的影响。为更精准地量化题干难度，本研究提出一种综合计算方法，综合考虑了词语平均难度、高难词偏差与句长对总体难度的影响。其计算公

[1]Wordfreq网址：https://github.com/rspeer/wordfreq。

式如下：

$$L_{\text{stem}} = \left(L_{\text{avg}} + \lambda \cdot \frac{1}{n} \sum_{i=1}^{n} \max\left(0, L_i - L_{\text{avg}}\right) \right) \times C_n \tag{1}$$

其中，L_{stem}为题干的最终难度得分，L_{avg}为句子中词语的平均难度，λ为高难词影响系数，用于调节高难词对整体难度的影响程度，L_i为第i 个词的难度值，n为句子中的总词数，C_n为句长调整系数，用以控制句子长度对整体难度的影响$\max\left(0, L_i - L_{\text{avg}}\right)$为高难词的难度偏差，仅考虑超过平均难度的词语。

词语平均难度：计算句子中所有词语的平均难度，计算方法为将所有词语的难度求和后取平均值，作为基础难度水平。

$$L_{\text{avg}} = \frac{1}{n} \sum_{i=1}^{n} L_i \tag{2}$$

高难词偏差项：计算所有高于平均难度的词语偏差值，只考虑比平均难度更大的词（若$L_i < L_{\text{avg}}$，则该项为0），反映了难词对整体难度的额外影响。

$$D = \frac{1}{n} \sum_{i=1}^{n} \max\left(0, L_i - L_{\text{avg}}\right) \tag{3}$$

句长调整系数：依据句子长度n 调整整体难度，句子越短，难度系数较小，句子越长，理解难度越大，因此给出更高的调整系数。

$$C_n = \begin{cases} 1.0 & \text{若} 1 \le n \le 10, \\ 1.1 & \text{若} 11 \le n \le 15, \\ 1.3 & \text{若} 16 \le n \le 20, \\ 1.6 & \text{若} 21 \le n \le 30, \\ 2.0 & \text{若} n > 30 \end{cases} \tag{4}$$

3.0.4 选项难度计算

为确保选项难度与整体习题难度合理匹配，本研究对选项难度进行精细评估。考虑到选项多为短语或词语，长度较短，故采用最高词汇等级而非均值，作为选项难度，以更真实反映其对学习者的认知挑战。此外，为避免选项难度显著高于目标词造成试题失衡，公式中引入“选项-答案难度差”，对超出部分进行加权修正，从而在保证干扰性的同时，增强习题整体难度评估的科学性与合理性。

$$L_{\text{option}} = \frac{1}{N-1} \sum_{k=1}^{N-1} \left(L_{k,\max} + \max\left(0, L_{k,\max} - L_{c,\max}\right) \right) \tag{5}$$

其中，L_{option}为选项的平均难度得分，N为总选项数（包括正确答案），k为当前选项的索引（$k = 1, 2, \ldots, N-1$），$L_{k,\max}$为第k 个选项中难度最高的词语等级，$L_{c,\max}$为正确答案中难度最高的词语等级，$\max\left(0, L_{k,\max} - L_{c,\max}\right)$为选项与正确答案难度差值的正部分，确保不会因更简单的选项降低整体难度。

选项词语最高难度：若第k 个选项由多个词语组成，则以其中难度最高的词语的等级作为该选项的难度值。

$$L_{k,\max} = \max\left\{L_{k,1}, L_{k,2}, \ldots, L_{k,m}\right\} \tag{6}$$

答案词语最高难度：若正确答案包含多个词语，则取其中难度最高的词语，其难度等级作为正确答案的难度值。

$$L_{c,\max} = \max\left\{L_{c,1}, L_{c,2}, \ldots, L_{c,p}\right\} \tag{7}$$

选项与答案的难度差：仅考虑高于正确答案难度的选项，以避免较简单的干扰项对整体难度产生下调作用，如果$L_{k,\max} > L_{c,\max}$，则$\Delta L_k = L_{k,\max} - L_{c,\max}$，如果$L_{k,\max} \le L_{c,\max}$，则$\Delta L_k = 0$。

$$\Delta L_k = \max\left(0, L_{k,\max} - L_{c,\max}\right) \tag{8}$$

3.0.5 习题难度计算公式

习题难度由题干难度与选项难度的线性插值计算得到，其具体公式为：

$$L_{\text{quiz}} = \alpha \cdot L_{\text{stem}} + \beta \cdot L_{\text{option}} \tag{9}$$

其中，L_{quiz}为最终习题难度得分，L_{stem}为题干难度得分L_{option}为选项难度得分，α为题干难度权重系数，用于调整题干在总难度中的影响（建议范围：$0.5 \sim 0.7$），β为选项难度权重系数，用于调整选项在总难度中的影响（建议范围：$0.3 \sim 0.5$），且满足$\alpha + \beta = 1$。

3.0.6 习题难度计算示例

以“咱们九点从学校(　)。”为例，该习题考查动词的正确使用，选项包括“停留”“出发”“到达”和“走开”，其中“出发”为正确答案。

习题难度计算主要包括题干难度和选项难度两个部分。首先，根据题干的分词结果及其词语等级，计算平均难度，并结合高难词偏差与句长信息，确定题干难度。随后，分析选项的词语等级，计算选项的难度等级。最后，根据预设的权重比例，综合计算得出习题的最终难度。

计算题干难度，题干的分词及其对应等级如表2所示：

表 2: 题干难度等级计算示例

词语	咱们	九	点	从	学校	出发
等级	2	1	1	1	1	2

计算平均词语等级：

$$L_{\text{avg}} = \frac{2+1+1+1+1+2}{6} = 1.333$$

计算高难词偏差项：

$$\sum_{i=1}^{n} \max(0, L_i - L_{\text{avg}}) = 0.667 + 0 + 0 + 0 + 0 + 0.667 = 1.334$$

多轮实验后对权重参数进行了系统调优，最终选取$\lambda = 0.5$ 作为权重，以用于调整值的计算，表现出较优的性能指标。

$$\lambda \cdot \frac{1.334}{6} = 0.5 \times 0.222 = 0.111$$

该句长度为8，对应的句长调整系数为1.0，因此最终题干难度为：

$$L_{\text{stem}} = (1.333 + 0.111) \times 1.0 = 1.444$$

计算选项难度，选项词语及其等级如表3所示：

表 3: 选项难度等级计算示例

选项	停留	出发	到达	走开
等级	5	2	3	2

其中，正确答案“出发” 的等级为：

$$L_{c,\max} = 2$$

计算所有选项的难度累加值：

$$\sum_{k=1}^{N-1} (L_{k,\max} + \max(0, L_{k,\max} - L_{c,\max})) = (5+3) + (3+1) + 2 = 14$$

计算选项最终难度：

$$L_{\text{option}} = \frac{14}{3} = 4.667$$

计算习题难度，将题干难度和选项难度带入习题难度计算公式，得到习题难度如下：

$$\begin{aligned} L_{\text{quiz}} &= 0.7 \cdot L_{\text{stem}} + 0.3 \cdot L_{\text{option}} \\ &= 0.7 \times 1.444 + 0.3 \times 4.667 \\ &= 1.0108 + 1.4001 = 2.411 \end{aligned}$$

其中，权重系数0.7 与0.3 的设置基于以下考虑：题干作为理解题意和锁定语义焦点的核心载体，其语言复杂度对整体答题难度影响更大，因而赋予更高权重（0.7）；而选项部分尽管同样涉及识别与区分，但对语境理解的依赖相对较小，因此设置较低权重（0.3）。该比例亦参考了预实验中对答题耗时与错误率的分析结果，表现出较好的区分度与合理性。

3.0.7 习题难度等级划分

为提升习题的区分度与适配性，本文基于难度数据统计结果，将所有习题划分为6 个难度等级，旨在覆盖不同语言水平学习者的学习需求。具体难度等级划分标准见表4。

表 4: 习题难度等级标准

难度等级	分值范围	描述
L1	[0.00, 3.00)	题干1级词为主，选项1级词为主
L2	[3.00, 3.50)	题干1-2级词为主，选项2级词为主
L3	[3.50, 4.00)	题干1-3级词为主，选项3级词为主。
L4	[4.00, 4.50)	题干1-4级词为主，2-3级词比重增加，选项4级词为主。
L5	[4.50, 5.00)	题干1-4级词为主，2-3级词占比进一步提升，选项5、7-9级词为主。
L6	[5.00, 10.00]	题干1-9级词为主，选项6-9级词为主。

数据分析显示，不同难度习题在词汇层级上呈线性关系。如图1，L1–L2 题干以基础词汇为主，语言简洁；L3–L4 词汇量扩大，句式更复杂；L5–L6 引入高阶词汇与真实语境，适用于中高级测评。图2 显示，L1–L2 选项差异明显，辨析容易；L3–L4 增加近义干扰，混淆度提升；L5–L6 干扰项涵盖语义细节与语体特征，辨析难度显著增强，区分性更强。

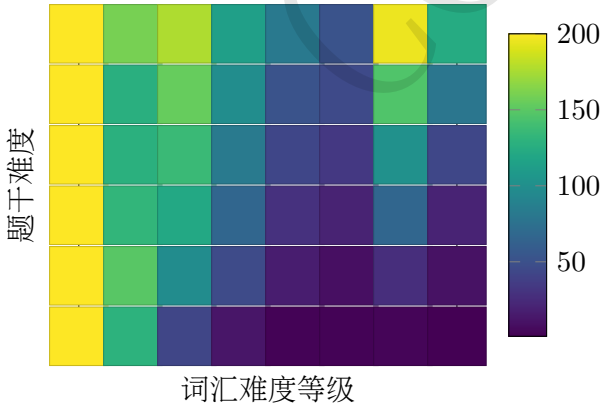

图 1: 题干词汇难度分布热力图

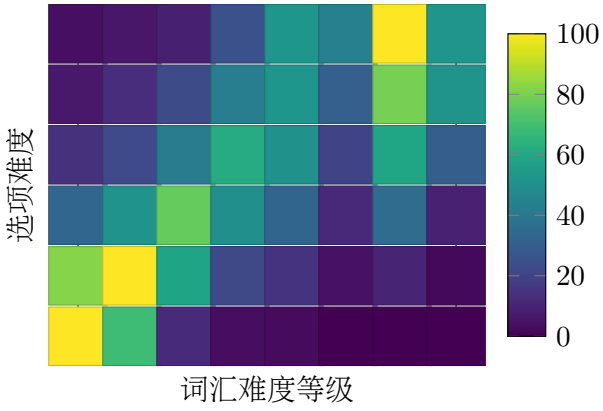

图 2: 选项词汇难度分布热力图

综上，本研究构建的难度等级体系结合词汇难度、题干复杂度与干扰项强度，实现了层次清晰、难度递进的设计目标。该体系有助于提升词汇测评的科学性、适配性与诊断价值，为教学提供精准反馈支持。

4 难度可控习题生成框架

为提升习题的整体质量与生成稳定性，引入检索增强技术，利用语言要素库和BCC 语料库动态提供语境支撑与备选词项，从而提高习题难度控制的精准度与内容生成的多样性，为保障自动生成习题的可用性与规范性，设计了多重习题校验机制，对生成内容的答案唯一性、结构完整性和格式规范性进行自动验证，从而输出符合教学要求的高质量选择题。

4.1 微调习题生成模型

模型训练流程主要包括三个阶段：数据预处理、模型微调与模型评估优化。首先，对原始习题数据进行格式转换与集划分；其次，采用LoRA 技术对预训练模型进行高效参数微调；最后，通过验证集对模型性能进行评估与持续优化，以提升习题生成的准确性与鲁棒性。

在模型微调阶段，采用LoRA 技术对Llama3.1-8B 预训练模型进行参数高效微调。LoRA 通过在部分Transformer 层插入低秩矩阵，并冻结主干参数，从而降低计算开销，提高训练效率。微调过程中，模型首先加载预训练权重，并通过PEFT 框架实现LoRA 微调。关键超参数设置如下：秩参数$r = 8$，放大因子$\alpha = 32$，随机失活比例dropout=0.1，优化器选择AdamW，学习率设定为$1e^{-5}$，批量大小为4，最大训练轮数为5。整个训练流程由Hugging Face 的Trainer 类自动管理，以确保训练的稳定性与高效性。训练完成后，从所有训练快照中选取在验证集上表现最佳的权重文件，并整合至预训练模型，形成最终的自动出题微调模型。

4.2 检索增强习题生成

尽管微调后的自动出题模型在习题生成任务中取得了一定效果，但在多样性控制和难度精度方面仍存在局限。为进一步提升生成质量和适配性，本研究引入检索增强生成技术，构建了融合知识检索与生成能力的习题生成系统。其整体架构如图3 所示。

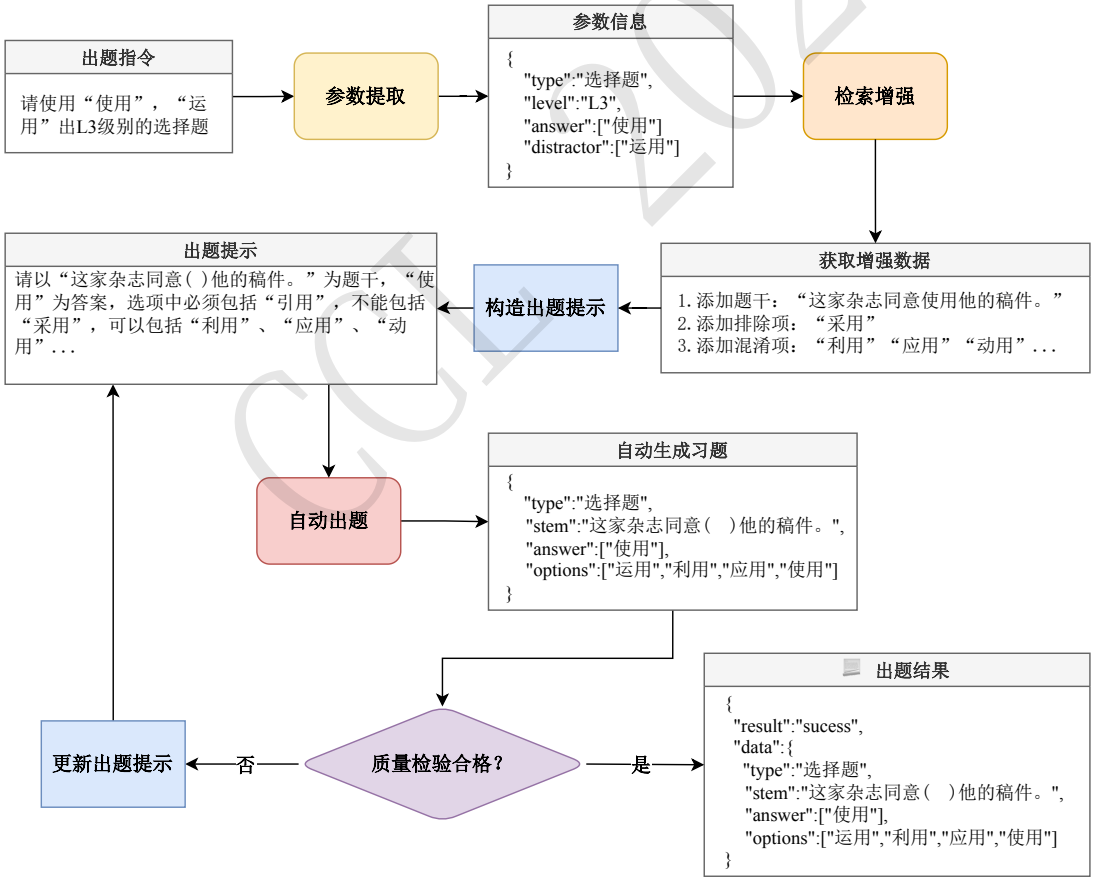

图 3: 倒排索引检索示意图

首先，系统从教师的自然语言输入中提取出题所需的关键参数，包括课程类型、难度等级、题型、题干、正确答案及干扰项等。随后，结合检索增强机制对不完整或缺失的参数进行

动态补全，确保输入信息完整、语义清晰。补全后的信息将用于构建标准化出题提示，并调用已微调的习题生成模型，从而生成符合预期难度和结构规范的高质量选择题。

4.2.1 参数提取

为实现自动化、结构化的出题流程，参数提取模块负责从教师提供的自然语言出题指令中，提取题型、目标词、题干及干扰项在内的信息。该模块基于微调的大语言模型构建，复用了习题生成模型的出题模板设计，将自然语言指令作为输入，输出标准化的结构化参数信息。

本研究选用Llama3.1-8B 作为基础模型，在指令微调阶段同时融合参数提取与习题生成任务，使微调后的模型同时具备出题意图识别能力以及习题生成能力。

为验证模型在参数提取任务中的实际表现，构建了包含1316 条样本的测试集，覆盖多种出题形式与表达方式。Llama3.1-8B 仅有2 条数据存在参数缺失，参数提取准确率高达99.85%。Llama3.1-8B 在参数提取任务中的表现优异，具备极高的准确性与鲁棒性，适合作为检索增强出题系统的核心参数提取模块，为后续习题构建与质量控制提供稳定支持。

4.2.2 检索增强

为提升习题生成的多样性与难度控制精度，本研究在出题流程中引入了检索增强模块。该模块通过动态补充和完善习题参数信息，在题干、答案和选项构建过程中引入语言要素库与BCC 语料库中的语言资源，从而确保生成的习题更加完整、精准且符合目标难度要求。

具体而言，当教师提供的出题参数有缺失时，检索增强模块会根据预设难度要求，从语料库中检索合适的句子作为题干，并匹配音近词、形近词、近义词或常见错用词作为干扰项，以提高习题的区分度和语境合理性。

该模块主要包括三个核心功能：**难度控制**：根据出题参数中的难度等级，从语言要素库和BCC 语料库中筛选匹配的词语和句子资源，确保生成的习题与学习者水平相符。**题干生成**：当题干信息缺失时，模块根据目标答案与目标难度，自动检索并补全题干，确保语境自然，语法合理。**选项生成**：模块通过多种策略筛选高质量干扰项，优先考虑与答案词义相近但不构成语义重复的词语，同时剔除可能与答案混淆的多义词或同义项，以避免生成多答案习题。

4.2.3 习题生成

在检索增强完成初步信息补全后，习题生成模块负责依据构造好的出题模板，调用大语言模型生成最终习题。与传统自动出题方式不同，本模块的输入包含排除项信息，以防止生成答案不唯一的选项组合。

系统首先接收检索增强模块返回的题干、答案、干扰项与排除项信息，然后进一步筛选或补足干扰项，最终形成完整的四选一结构。该模块的设计目标是在确保难度控制的同时，提高习题的语言自然性、结构多样性与可读性。

得益于前置检索增强模块提供的高质量语境与词汇输入，习题生成模块能有效避免生成模式化、重复度高的题目，使生成结果在内容与形式上更加丰富，增强了语言学习过程中的认知挑战性和互动性。

4.2.4 习题校验

为确保自动生成习题的质量与可用性，本研究设计了习题校验模块，围绕格式正确性、答案唯一性与难度适配性三大维度进行审核，并引入迭代优化机制，提升出题系统的稳定性与实用性。

格式校验：采用规则匹配检测题干是否含“(　)”、选项是否为四个、是否有重复项，以及答案是否包含在选项中，确保格式规范、结构完整。

唯一性校验：调用大语言模型模拟答题，比较其选择与预设答案的一致性。若不一致，则判定答案不唯一或干扰项过强，系统将剔除问题选项并重新生成，避免歧义或误选。

难度校验：通过难度模型评分判断是否在预设等级范围内。若偏离目标难度，系统将通过替换题干或调整选项进行重构，以确保题目既具挑战性又具可接受性。

迭代机制：若任一校验失败，系统触发自动迭代流程，重新生成直至通过全部校验或达到最大迭代次数，避免死循环并节约资源。

综上，该校验模块实现对生成习题的结构、难度与逻辑的全面把控，为智能出题系统的稳定运行提供保障。

5 实验

本研究探讨检索增强技术在自动出题任务中的影响，重点评估其在提升习题多样性和精准控制难度方面的作用。此外，通过具体出题案例进行对比分析，全面比较检索增强模型与微调生成模型在自动出题中的差异。

5.1 实验评测指标

为全面、客观地评估自动出题系统的性能，本研究设计了三项评测指标，涵盖出题成功率、出题正确率、习题重复率以全方位衡量系统的出题质量与适用性。

出题成功率（Question Generation Success Rate，QGSR）：衡量习题是否能够生成，并符合习题的格式要求，该指标反映模型在实际应用中的稳定性和出题能力。该指标公式定义如下：

$$S = \frac{N_s}{N_t} \tag{10}$$

其中，S为出题成功率，N_s为成功生成的习题数量，N_t为总出题次数。

出题正确率（Question Generation Accuracy，QGA）：衡量习题答案的唯一性，在格式正确的基础上，确保生成的题目不存在多个正确答案，该指标用于评估模型在答案生成上的准确性和一致性。该指标公式定义如下：

$$C = \frac{N_c}{N_s} \tag{11}$$

其中，C为出题正确率，N_c为答案唯一的习题数量，N_s为成功生成的习题数量。

习题重复率（Question Generation Redundancy Rate，QGRR）：在数据分析中，重复率R 表示习题中重复的比例，为评估习题的多样性，本研究引入习题重复率，该指标衡量在多次出题过程中，相同习题重复出现的情况。计算方法如下：

$$R = 1 - \frac{C}{N} \tag{12}$$

其中，R为习题重复率，C为不同习题个数（即唯一值的数量），N为生成的习题数量。

5.2 自动出题实验

本实验旨在评估检索增强模型在自动出题任务中的整体性能，并与微调生成模型进行对比分析，从而全面衡量两种方法在习题生成中的实际效果与适用性。为此，实验设置了三类任务，分别对应出题过程中的关键能力：题干构建、选项优化与词义辨析。

5.2.1 实验设置

在题干驱动任务中，从语言要素库中抽取200句作为输入，要求模型根据题干等级生成符合标准格式的选择题，以测试其选项筛选能力和干扰项质量控制水平。答案驱动任务则从等级词汇表中随机选取200个词作为目标答案，分别生成L1至L6等级的习题，共计1200道，重点考察模型在题干构建、语境适配与综合生成能力方面的表现。选项驱动任务中，同样从词汇表中抽取200个目标词，并配以近义词组成选项，生成1200道覆盖不同等级的习题，主要用于测试模型的词义辨析能力，确保干扰项具有辨析性而非歧义性，避免出现多个合理答案的情况。

5.2.2 实验结果

表5 展示了不同测试类型的出题成功率，该指标用于衡量模型对出题指令的遵循程度。

测试类型	微调生成模型	检索增强模型
基于题干	100.00%	100.00%
基于答案	100.00%	99.25%
基于选项	99.67%	98.25%

表 5: 不同测试类型的出题成功率

整体来看，无论是微调生成模型还是检索增强模型，出题成功率均保持在较高水平，表明两种方法在生成符合规范的习题方面具有较强的可靠性。

表6 展示了不同测试类型下的出题正确率。其中，“模型微调”表示仅使用微调后的模型出题，“检索增强”表示引入检索增强模块的正确率结果，“质量校验”表示在检索增强基础上，通过引入通义大模型[1]对出题正确性进行检验后的结果。

测试类型	模型微调	检索增强	质量检验
基于题干	68.50%	70.50%	98.50%
基于答案	67.02%	68.25%	97.52%
基于选项	50.28%	62.57%	97.08%

表 6: 不同测试类型的出题正确率

数据表明，检索增强模型相较于微调生成模型，在出题正确率上有所提升，尤其在引入质量检验功能后，大幅减少了答案不唯一的问题，提高了习题的可用性和质量保障。

具体而言，尽管检索增强模型在习题多样性方面优于微调模型，但仍存在少量答案不唯一的情况。以目标词“推荐”为例，系统生成的题干为：“我想(　)给你一个朋友。”，选项包括：“推荐”“介绍”“建议”“告诉”。质量校验模块判定其中“推荐”与“介绍”均为正确答案，存在多个正确选项。系统据此将“介绍”纳入排除列表，并触发习题重新生成流程，确保最终习题仅保留唯一正确答案。质量校验机制不仅提升了出题准确率，也增强了系统在实际应用中的可靠性，为后续实现大规模自动化出题奠定了基础。

表7 展示了不同测试类型的习题重复率。在未引入检索增强模型之前，题干的重复率较高，尤其是在同一目标词的不同难度级别习题生成过程中。

测试类型	微调生成模型	检索增强模型
基于答案	52.25%	15.37%
基于选项	54.77%	14.50%

表 7: 不同测试类型的习题重复率

例如，以“承担”为目标词分别生成L1 至L6 等级的词义辨析习题时，模型多次返回相同题干“他(　)不起这么大的责任。”，缺乏语境多样性，导致不同难度等级的习题在实际测试中区分度不足。

5.2.3 实验分析

本实验对比分析了微调生成模型与检索增强模型在自动出题任务中的表现，并从出题成功率、正确率及习题重复率三个方面进行了全面评估。实验结果表明，检索增强模型在习题质量和多样性方面表现更优，尤其是在确保答案唯一性和减少重复习题方面具有显著优势。

5.3 生成习题难度实验

本实验评估检索增强模型在习题难度控制方面的能力，验证其是否能生成与设定等级相匹配的练习内容，从而提升测评适配性与学习效果。实验分别采用“基于答案出题”和“基于选项出题”两种方式，统计习题的难度均值与标准差，结果见表8。

实验显示，两种模式下生成的习题难度均值整体接近目标等级，标准差多在1.0 以下，表明模型具备较好的难度控制稳定性。

从具体出题模式来看，基于选项出题在各等级下的难度分布更为集中，标准差普遍低于基于答案出题模式。例如，在L4、L5 和L6 级别中，基于选项出题的标准差分别为0.55、0.63和0.78，均显著小于对应的基于答案出题值，表明模型在中高难度习题区间具备更强的难度控制能力。

综上所述，检索增强模型在习题生成过程中展现出较强的难度调节能力，尤其在中高难度区间表现突出，能够有效提升习题的匹配度。相较而言，基于选项出题方式在难度稳定性方面更具优势。未来可引入更精细的调控机制，以进一步提升自动出题系统的精准性。

[1]通义网址：https://www.tongyi.com。

等级	基于答案		基于选项	
	平均值	标准差	平均值	标准差
L1	L1(2.98)	1.17	L1(2.92)	0.87
L2	L2(3.19)	1.02	L2(3.02)	0.77
L3	L3(3.68)	0.83	L3(3.50)	0.65
L4	L4(4.39)	0.86	L4(4.22)	0.55
L5	L5(4.94)	0.94	L5(4.92)	0.63
L6	L6(5.62)	1.10	L6(5.47)	0.78

表 8: 生成习题难度等级分布

5.4 案例分析与效果验证

为进一步验证检索增强技术在自动出题中的实际效果，本研究从习题多样性与难度控制两个维度，比较了“基于选项出题”与“基于答案出题”两种方式下，微调生成模型与检索增强模型的出题表现。

本实验从不同等级词汇中选取代表性词语，分别调用两类模型进行出题，并对生成习题的内容质量和难度分布进行分析，从而评估模型在内容生成与难度调控方面的能力。

5.4.1 基于选项出题

以“出来”与“出去”这组近义词为例，考察模型在语义辨析能力方面的表现，相关出题示例见表9。实验结果表明，检索增强模型在选项设计上更具辨析性，能够准确识别语义接近但语用差异明显的词项作为干扰项，从而显著提升题目的测试效度。相比之下，微调生成模型易生成缺乏区分度或语义无关的选项，影响习题的区分能力与实际应用效果。

模型	等级	习题	难度
微调生成模型	L1 ~ L6	听到妈妈叫他，小明从房间里走了(　)。 回去、进去、出来、出去	L1(2.64)
检索增强模型	L1	老师的问题她回答不(　)。 出现、出名、出来、出去	L1(2.93)
	L2	这篇文章他仅仅一个小时就写(　)了。 出现、出来、出色、出去	L2(3.07)
	L3	班长是我们大家选(　)的。 出来、出头、出现、出去	L3(3.92)
	L4	陈静装(　)一副无所谓的样子，其实心里很着急。 出去、出头、出现、出来	L4(4.47)
	L5	林英把她的喜悦用舞蹈表现(　)了。 出现、出来、出动、出去	L5(4.86)
	L6	我有什么地方做得不好，请你指(　)。 出来、出头、出去、出现	L6(5.18)

表 9: 基于选项出题示例

实验结果表明，检索增强模型在选项设计上更具辨析性，能够准确识别语义接近但语用差异明显的词项作为干扰项，从而显著提升题目的测试效度。相比之下，微调生成模型易生成缺乏区分度或语义无关的选项，影响习题的区分能力与实际应用效果。

5.4.2 基于答案出题

以“土地”为例，表10 展示了微调生成模型与检索增强模型在L1–L6 六个等级下生成的习题示例。从示例可见，检索增强模型生成的题干更具语境多样性，覆盖了更广泛的语用场景，同时在不同等级间展现出更明显的难度梯度，避免了题干重复与难度集中现象。相比之下，微调

生成模型存在重复出题的问题，部分等级使用了相同题干，且选项质量较差，易出现多个正确选项的情况，影响习题区分度与评估效度。

模型	等级	习题	难度
微调生成模型	L1 ~ L4	这块(　)是他家的。 土地、农田、地、农场	L3(3.54)
	L5	中国的(　)资源丰富，特别是煤炭。 土地、水、气候、矿产	L3(3.90)
	L6	中国是一个()大国。 土地、田地、国土、大地	L6(5.06)
检索增强模型	L1	我深深的爱着这片(　)。 土地、国家、地球、地区	L1(2.70)
	L2	这一亩(　)很适合种蔬菜。 土地、道路、地球、地区	L2(3.29)
	L3	我家人口少，承包的(　)不多。 土地、土壤、地球、地区	L3(3.57)
	L4	这片(　)养育了几代人，人们对它充满感情。 土壤、地球、地区、土地	L4(4.12)
	L5	踏上自己的(　)，心里就有一种亲切感和自豪感。 尘土、石头、土地、道路	L5(4.50)
	L6	农民们辛勤耕作，希望这片(　)能带来好收成。 土地、田地、领土、大地	L6(7.32)

表 10: 基于答案出题示例

从示例可见，检索增强模型生成的题干更具语境多样性，覆盖了更广泛的语用场景，同时在不同等级间展现出更明显的难度梯度，避免了题干重复与难度集中现象。相比之下，微调生成模型存在重复出题的问题，部分等级使用了相同题干，且选项质量较差，易出现多个合理答案，影响习题区分度与评估效度。

5.4.3 检索增强模型的优势

实验结果表明，检索增强模型在习题生成任务中整体表现优于微调模型，尤其在多样性、难度控制与选项优化方面展现出显著优势。该模型借助语言要素库与BCC 语料库，生成内容更加丰富，显著减少重复，提升习题覆盖度与灵活性。在难度控制方面，模型能够依据预设等级筛选题干与干扰项，实现精准调控，保证内容与学习者水平匹配，提升题目区分度。同时，模型在选项设计中表现出更强的语义判断能力，能够动态识别必选项与排除项，生成语义相近且混淆度高的干扰项，增强词义辨析性，有效避免答案歧义。此外，依托真实语料生成的题干与选项在语言结构和语义表达上更自然规范，整体语言质量显著提升。

6 结论

本文提出一种基于RAG 的中文词义辨析习题自动生成框架，结合参数提取、语料检索与质量校验机制，实现结构规范、难度可控的习题生成。同时，设计融合词汇等级、词频、句长与干扰项强度的难度计算模型，量化评估题干与选项难度，支持习题分级控制。实验结果显示，该方法在出题成功率、答案唯一性、多样性与难度适配性方面均优于微调模型，验证了检索增强与难度建模在智能出题中的有效性。

参考文献

Hessa A. Alawwad, Areej Alhothali, Usman Naseem, Ali Alkhathlan, and Amani Jamal. 2025. Enhancing textual textbook question answering with large language models and retrieval augmented generation. *Pattern Recognition*, 162:111332, June.

Semere Kiros Bitew, Amir Hadifar, Lucas Sterckx, Johannes Deleu, Chris Develder, and Thomas Demeester. 2024. Learning to Reuse Distractors to Support Multiple-Choice Question Generation in Education. *IEEE Transactions on Learning Technologies*, 17:375–390.

Jonathan Brown, Gwen Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 819–826.

Abdur Rahman Fahad, Nazme Al Nahian, Md Ahanaf Islam, and Rashedur M. Rahman. 2024. Answer Agnostic Question Generation in Bangla Language. *International Journal of Networked and Distributed Computing*, 12(1):82–107, June.

Ching Nam Hang, Chee Wei Tan, and Pei-Duo Yu. 2024. MCQGen: A Large Language Model-Driven MCQ Generator for Personalized Learning. *IEEE Access*, 12:102261–102273.

John Brian Heaton. 2000. *Writing English language tests*. Foreign Language Teaching and Research Press, Beijing.

Po-Chun Huang, Ying-Hong Chan, Ching-Yu Yang, Hung-Yuan Chen, and Yao-Chung Fan. 2024. EQGG: Automatic Question Group Generation. *IEEE Transactions on Learning Technologies*, 17:1994–2007.

Shu Jiang and John SY Lee. 2017. Distractor generation for chinese fill-in-the-blank items. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications*, pages 143–148.

Zhongguo Li and Maosong Sun. 2009. Punctuation as implicit annotations for chinese word segmentation. *Computational Linguistics*, 35(4):505–512.

Chao-Lin Liu, Chun-Hung Wang, Zhao-Ming Gao, and Shang-Ming Huang. 2005. Applications of lexical information for algorithmically composing multiple-choice cloze items. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP - EdAppsNLP 05*, pages 1–8. Association for Computational Linguistics.

Ruslan Mitkov et al. 2003. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing*, pages 17–22.

Ian SP Nation and ISP Nation. 2001. *Learning vocabulary in another language*, volume 10. Cambridge university press Cambridge.

Alexander Ngo, Saumya Gupta, Oliver Perrine, Rithik Reddy, Sherry Ershadi, and Daniel Remick. 2024. ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. *Academic Pathology*, 11(1):100099, January.

Andreas Papasalouros, Konstantinos Kanaris, and Konstantinos Kotis. 2008. Automatic generation of multiple choice questions from domain ontologies. *e-Learning*, 1:427–434.

Atharva Patil, Mandar Patil, Chinmay Bhosale, Vedant Koppal, and Uma Gurav. 2024. Deep learning based automated question generation for examination system. In *2024 4th International Conference on Artificial Intelligence and Signal Processing (AISP)*, pages 1–5. IEEE.

Arief Yudha Satria and Takenobu Tokunaga. 2017. Automatic generation of english reference question by utilising nonrestrictive relative clause. In *International Conference on Computer Supported Education*, volume 2, pages 379–386. SCITEPRESS.

Muhammad Shoaib, Ghassan Hasnain, Nasir Sayed, Yazeed Yasin Ghadi, Masoud Alajmi, and Ayman Qahmash. 2025. Automated generation of multiple-choice questions for computer science education using conditional generative adversarial networks. *IEEE Access*.

Lae Lae Shwe, Sureena Matayong, and Suntorn Witosurapot. 2024. The unified difficulty ranking mechanism for automatic multiple choice question generation in digital storytelling domain. *Education and Information Technologies*, 29(15):20317–20350, October.

Yuni Susanti, Ryu Iida, and Takenobu Tokunaga. 2015. Automatic generation of english vocabulary tests. In *CSEDU (1)*, pages 77–87.

Goto Takuya, Kojiri Tomoko, Watanabe Toyohide, Iwata Tomoharu, and Yamada Takeshi. 2010. Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning : an International Journal*, 2(3).

Yizhen Wang, Xueying Shen, Zixian Huang, Lihui Niu, and Shiyan Ou. 2024. cLegal-QA: A Chinese legal question answering with natural language generation methods. *Complex & Intelligent Systems*, 11(1):77, December.

Lili Wang, Ruiyuan Song, Weitong Guo, and Hongwu Yang. 2025. Exploring prompt pattern for generative artificial intelligence in automatic question generation. *Interactive Learning Environments*, 33(3):2559–2584.

Cheng Zhang. 2022. *Automatic Generation of Multiple-Choice Questions*. Ph.D. thesis, University of Massachusetts Lowell.

丁向民. 2008. 基于本体的多项选择题自动生成技术研究. 硕士, 南京航空航天大学.

刘润清and 韩宝成. 2007. 语言测试和它的方法（修订版）. 外语教学与研究出版社, 北京.

刘超英. 2022. 汉语第二语言测试与评估. 北京大学出版社, 北京.

唐奇峰. 2012. 基于本体的自动出题系统的研究. Master's thesis, 广西师范大学.

崔希亮. 2023. 语言学概论（增订本）. 商务印书馆, 北京.

张凯. 2002. 语言测试理论与实践. 北京语言大学出版社.

李钢. 2013. 基于难度组合算法的自动出题系统. 硕士, 渤海大学.

温雪峰, 崔仙姬, and 张俊星. 2020. 基于语义相似性的选择题自动生成优化方法. 计算机与数字工程, 48(12):2850–2856.

胡韧奋. 2017. 汉语作为第二语言的词汇测试自动命题研究. Ph.D. thesis, 北京师范大学.

苏向丽and 李如龙. 2011. 词价研究与汉语词汇知识的深度习得. 语言文字应用, (2):63–70.