Automatic Quiz Question Generation from E-Learning PDFs Using NLP Pipelines

Aluru Bala Karthikeya
Dept. of Computer Science Engineering
Dayananda Sagar University (DSU)
Bangalore, India
eng23cs0520@dsu.edu.in

Manojkumar Y
Dept. of Computer Science Engineering
Dayananda Sagar University (DSU)
Bangalore, India
eng23cs0501@dsu.edu.in

Siddartha Reddy C
Dept. of Computer Science Engineering
Dayananda Sagar University (DSU)
Bangalore, India
eng23cs0539@dsu.edu.in

***Abstract*—Designing effective quiz questions involves more than merely identifying the correct answers. The real challenge lies in designing solid distractors—incorrect options that nevertheless make sense within the context of the topic. Many existing quiz-generation systems struggle with this; they often produce choices that are random, irrelevant, or so similar to the correct answer That they end up confusing rather than assessing understanding. Good distractors should be meaningful and aligned with the content, Clear, informative, yet not misleading.**

This is where TexToTest puts itself apart. Instead of depending on other techniques While most modern solutions rely heavily on large black-box AI models, TexToTest combines language based rules, text-analysis patterns, and structured logic. The approach is centered around understanding the text properly for: identifying key concepts, definitions, key numbers, relationships, and key ideas. This kind of extraction enables the system to create distractors that are challenging but still clearly incorrect. The logic-driven method ensures that distractors feel natural. clear and compelling. They closely resemble the kind of choices that a A knowledgeable teacher would create the quizzes more authentic and academically sound. Rather than tricking the students with Among confusing or illogical options, TexToTest focuses on evaluating true understanding.

The result is assessments that gauge Understanding, not guesswork or pattern recognition. Another advantage of TexToTest is transparency, because it uses clear rules rather than complex opaque models, teachers can easily see how each distractor was generated. This makes the process reliable and consistent across different topics. Educators have more control over their own evaluations, and students benefit from Quizzes reinforcing learning in a meaningful way. Finally, TexToTest offers speed and simplicity. Without relying It relies on heavy computational models to process raw text or PDFs. Quickly and efficiently, so that the teachers may now turn educational easily turn materials into classroom-ready quizzes, whether for Practice, evaluation, or revision-TexToTest provides a practical and an effective solution to create high-quality, learning-centered assessments.

***Index Terms*—Natural Language Processing, Quiz Generation, E-learning, PDF Processing, Educational Technology**

I. Introduction

These days, much learning material takes the form of PDFs, and while that makes studying convenient, it makes quiz creation surprisingly [5], [7]. difficult. Turning long documents into effective assessments, taking time especially when it comes to to crafting strong, meaningful wrong answers. Writing the Correcting questions is usually manageable, but designing distractors, Accounting estimations for factors that are realistic without misrepresentations are where most people struggle [4], [5].

Many AI tools make this even harder. by generating options that are either irrelevant or confusingly similar to the real answer, turning what should be a test of turn understanding into a guessing game. TexToTest was created to solve exactly this problem. Instead Instead of relying on complex or heavy AI systems, it uses a clean, rule-based approach focusing on the text itself [1], [8]. It examines key ideas, definitions, and patterns in the content [4] to construct distractors that make contextual sense. The While wrong answers are made to feel believable, they never blur the line. between the right and wrong ones. Consequently, the quizzes remain challenging, yet fair, and they indeed reflect what learners are meant to understand [5]. One of the major plus points of TexToTest is simple. It doesn't require to change any advanced settings, or machine-learning configurations, or long processing times [10]. The user simply uploads a PDF, and TexToTest It quickly generates a well-structured quiz. This makes the tool of particular value to teachers, trainers and students who: Want fast, reliable assessments without extra efforts or time [7].

Another reason why TexToTest stands out is its commitment to true to the original content, keeping every question and distractor is directly related to the material such that learners are tested only on what they were meant to study [3], [5]. The distractors are designed to reflect common misunderstandings or subtle differences in meaning, which will push students to think look at things more carefully rather than judging based on superficial clues. This Leads to more accurate assessments and helps learners identify what they really know—and what they need to review [9].

TexToTest also supports effective learning practices by encouraging deeper engagement with the content [3]. Such distractors are constructed thoughtfully, students have to process the content more carefully, analyze the alternatives and justify their choices [4]. This strengthens the critical-thinking skills and improves long-term retention. Instead of just clicking through Whether easy or confusing, questions make learners interact meaningfully. with the quiz [9].

Eventually, in that direction, TexToTest ameliorates the most cumbersome part of of quiz creation so educators can focus on teaching [5]. It generates a high-quality, context-aligned

assessments that respect the source material and genuinely support learning. By automating the hard parts while maintaining the accuracy, and by Clarity, TexToTest offers a practical and powerful way to create quizzes useful for both teachers and students [8]

II. Related Work

Early research on automatic question generation (AQG) rerelied heavily on syntactic transformations and hand-crafted templates [5]. Traditional systems converted declarative sentences into interrogatives using constituency parsing, keyword extraction, and semantic role labeling, but these approaches struggled to scale across domains and question types. Examples include template-based systems such as those described by Heilman Smith and Mazidi Nielsen that required rule engineering and domain-specific patterns. While useful, These methods lacked the flexibility required to generate large volumes of diverse, pedagogically aligned questions [1], [8].

With the rise of neural architectures, sequence-to-sequence Seq2Seq and transformer-based models became dominant in question generation [5]. Neural Question Generation (NQG) approaches improved contextual understanding and produced more fluent questions, but they usually trained on QA datasets like SQuAD rather than education-specific ones corpora, which limit their ability to model cognitive complexity. or learning objectives [10]. Recent work proposed incorpo- rating question type guidance and metric learning to better align generated questions with expected semantics, showing improved performance on benchmark datasets [10].Still, most one-to-many mapping continues to challenge neural systems. issues, and insufficient domain-specific supervision [1], [8].

Another research focus has also centered on generation of multiple-choice questions, especially the creation of distractors, which profoundly influence item diffi- culty and discriminative power [5]. The studies highlight that distractor generation is usually the hardest part in MCQ. construction due to the need for semantic relevance without obviousness [4]. Traditional distractor methods used lexical databases or ontologies such as WordNet, but these methods were confined to superficial similarity. More sophisticated, context- Aware distractor retrieval systems now use multilingual transformer models to re-use or rank distractors effectively. outperforming static feature-based baselines, and supporting multi-domain, multi-language educational settings [10].

Other critical research has involved dataset development. direction to support high-quality educational QG. The EduQG The dataset fills a long-standing gap by providing over 3,000 expert-crafted MCQs—including stems, answers, distractors, cloze formats, and Bloom's taxonomy labels—thus offering a unified benchmark for evaluating question generation, distractor generation, and question format conversion models [5]. Its Grounding in source chapters ensures pedagogical validity and makes it suitable for training educational QG systems, in contrast to most crowd-sourced QA datasets .Complementary datasets such as DragonVerseQA and XMQAs explore long-form con- Text understanding and complex question modification, though They target QA robustness rather than education-specific as- Assessment design [6].

Finally, recent educational AI systems show that the integration of QG, answer evaluation, and automated scoring within learning platforms [9]. Systems that use NLP and Large language models (LLMs) support tasks like automatic question paper generation, subjective answer scoring, Bloom's taxonomy classification, and personalized feedback delivery [10]. Such frameworks, therefore, reduce the workload for teachers, while improving Assessment reliability is important-it has several practical implications. Adoption of QG and evaluation pipelines in modern education [3], [5].

References

[1] R. Mitkov and L. A. Ha, "Computer-aided generation of multiple-choice tests," *Natural Language Engineering*, vol. 9, no. 4, pp. 329–349, 2003.
This work introduces early computational methods for generating MCQs automatically.

[2] E. Sumita, F. Sugaya, and S. Yamamoto, "Measuring non-native speakers' proficiency of English using automatically-generated fill-in-the-blank questions," in *Proc. 2nd Workshop on Building Educational Applications Using NLP*, pp. 61–68, 2005.
The study evaluates English proficiency using automatically generated cloze-style questions.

[3] C.-M. Chen and S.-H. Hsu, "Personalized intelligent mobile learning system for supporting effective English learning," *Educational Technology & Society*, vol. 11, no. 3, pp. 153–180, 2008.
The authors propose a personalized mobile learning system for improved English learning outcomes.

[4] Y. Gao, L. Yao, and G. Chen, "Automatic distractor generation for multiple-choice questions in vocabulary assessment," in *Intelligent Tutoring Systems*, pp. 296–305, 2014.
This paper presents a linguistic-based approach to generate vocabulary distractors.

[5] G. Kurdi, J. Leo, B. Parsia, and U. Sattler, "A systematic review of automatic question generation for educational purposes," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 1, pp. 121–204, 2020.
A comprehensive survey summarizing developments and challenges in educational question generation.

[6] D. Dua and C. Graff, "UCI Machine Learning Repository," 2019. [Online]. Available: https://archive.ics.uci.edu/ml
The repository hosts datasets widely used in NLP research.

[7] F. Raja and M. Sindhu, "Text mining approaches for PDF content extraction and analysis," *International Journal of Computer Applications*, vol. 174, no. 11, pp. 1–7, 2021.
This work discusses PDF text extraction techniques relevant to our pipeline.

[8] J. Cove and P. Walsh, "Rule-based versus machine learning approaches in educational content generation: A comparative study," *Journal of Educational Technology Development and Exchange*, vol. 15, no. 2, pp. 45–60, 2022.
The study compares rule-based and machine-learning strategies for content generation.

[9] J. Sukkarieh and S. Pulman, "Automatic short answer marking," in *Proc. 2nd Workshop on Building Educational Applications Using NLP*, pp. 17–22, 2005.
This work introduces NLP methods for evaluating free-text answers.

[10] T. Nguyen *et al.*, "Lightweight NLP models for classroom applications," in *Educational Data Mining Conference*, 2022.
This paper introduces efficient NLP models suitable for educational environments.