Enhancing Question Answer Generation from PDFs: A Fusion of BERT, RAKE, T5 and DistilBERT with RQUGE Evaluation

Akshitha Mary A C[1], Prachi Acharya[2], R Rakshinee[3], Stefani Jeyaseelan[4] and
Thanvitha Nandakumar[5] , *Prakash P[6(✉)] and *Sakthivel V[7(✉)]

Vellore Institute of Technology, Chennai,
Vandalur-Kelambakkam Road, Chennai, India
School of Computer Science and Engineering,

[1]akshithamary.ac2021@vitstudent.ac.in
[2]prachi.acharya2021@vitstudent.ac.in
[3]rakshinee.r2021@vitstudent.ac.in
[4]stefani.jeyaseelan2021@vitstudent.ac.in
[5]thanvitha.nandakumar2021@vitstudent.ac.in
[6*]prakash.p@vit.ac.in
[7*]sakthivel.v@vit.ac.in

Abstract. In educational institutes worldwide, the process of formulating questions for assessments and educational materials poses a recurring challenge. Traditional methods often involve manual question creation or keyword-based searches, leading to time-consuming and often inefficient processes, especially when dealing with extensive PDF documents. Addressing this prevalent issue, the research introduces a comprehensive approach to question-answer generation (QAG) from PDF documents. The system uses the most recent Natural Language Processing (NLP) models to extract text from PDFs using pdfplumber, create document summaries using BERT, extract keywords with RAKE, and create questions using a retrained T5 model. DistilBERT is then used to generate the answers. The RQUGE metric is used to assess the system's performance. The outcomes indicate how well the suggested method generates pertinent queries and precise responses, highlighting its potential for use in information retrieval and document comprehension.

Keywords: Question Answer Generation, BERT, T5, DistilBERT, RAKE, RQUGE Metric, Natural Language Processing, Information Retrieval.

1 INTRODUCTION

The explosive growth of digital information has led to an overwhelming volume of textual data stored in diverse formats, with Portable Document Format (PDF) being a prevalent choice for document sharing and dissemination. This surge in document digitization has necessitated the development of advanced Natural Language Processing (NLP) techniques to extract meaningful information automatically. One crucial task in this domain is Question Answer Generation (QAG), which involves the automated creation of questions based on a given text and the subsequent generation of relevant answers.

The significance of QAG lies in its potential to enhance information accessibility. Traditional methods of extracting information from documents often rely on manual processes or keyword-based searches, which can be time-consuming and inefficient, especially when dealing with large datasets. QAG systems, on the other hand, automate this process by not only extracting key information but also formulating questions that prompt concise answers, mimicking the way humans seek and comprehend information.

The motivation behind this research stems from the growing need for sophisticated QAG systems, especially in handling the vast amounts of information locked within PDF documents. While serving as a popular format for preserving the original structure and layout of content, these documents pose unique challenges for automated information extraction due to their varied layouts, styles and rich textual content. Conventional information retrieval approaches struggle to provide accurate and contextually relevant results, prompting the exploration of advanced NLP models and techniques. Moreover, as the demand for automated information extraction continues to rise, there is a concurrent need for systems that not only retrieve relevant details but also present them in a structured and meaningful manner. QAG systems address this need by not only summarizing information but also by generating questions that guide the user through the core concepts of the document. This approach aligns with the natural way humans seek information, making it a valuable asset in diverse domains, from education to research and beyond.

The primary objectives of this research are rooted in addressing the limitations of existing QAG systems and harnessing the potential of advanced NLP models. The key objectives include - designing and implementing a robust pipeline encompassing various stages of QAG, from text extraction to question-and-answer generation. utilizing state-of-the-art NLP models, such as BERT, T5, and DistilBERT, to improve the system's accuracy and contextual awareness. Employing the RQUGE metric to assess the caliber of questions and answers that are created, offering a numerical representation of the system's efficacy. providing details on the real-world uses of the suggested QAG system and showcasing its adaptability to a variety of situations and use cases.

As the proposed paper delves into the intricacies of QAG from PDF documents, it is essential to underscore the broader significance of this research in the realm of information extraction and accessibility. The ability to automatically generate meaningful questions and answers from PDFs not only accelerates the process of information retrieval but also opens avenues for innovation in education, research and information-driven decision-making.

In academic settings, students and researchers often encounter extensive literature in PDF format. The QAG system can serve as a valuable companion, swiftly summarizing key concepts and generating insightful questions for a deeper understanding of the material. In professional settings, where time is often of the essence, an efficient QAG system can empower individuals to extract pertinent information swiftly, enhancing productivity and informed decision-making. Furthermore, the implications of this research extend to fields where large volumes of documentation are prevalent, such as legal and medical domains. Legal professionals dealing with extensive case files can benefit from a QAG system that extracts critical details and formulates relevant questions for further investigation. Similarly, healthcare practitioners managing voluminous medical literature can streamline their information retrieval process with the aid of an automated QAG system.

In the context of emerging technologies, the integration of advanced NLP models in the proposed architecture showcases the potential for synergy between machine learning and document analysis. The adaptability of the system to different domains highlights its versatility and applicability, positioning it as a promising tool for various industries seeking efficient solutions for information extraction from unstructured data.

The structure of the research paper adheres to a sequence that begins with related works, followed by the presentation of the proposed architecture, a detailed description of the experimental setup, the unveiling of results, concluding remarks with insights into future work and a comprehensive list of references.

2 RELATED WORKS

Numerous studies on the Automated Question Answer Generation model have been conducted in the past, and some of the most notable research being conducted currently is cited.

Kettip Kriangchaivech et al. [1] proposed a method that automatically generated questions from Wikipedia passages using transformers on SQUAD. The evaluation of the model primarily focuses on the Word Error Rate (WER) as the metric for comparing the model-generated questions with the original SQuAD questions. To learn more about the quality of the questions, the WER distribution is examined, with particular attention to various WER ranges. In addition, the evaluation takes into account how many words each question has and how frequently the model-generated questions start with a certain word in comparison to the SQuAD questions. 9.94% of the model-derived questions had a WER of less than or equal to 5, 56.38% had a WER of between 6 and 10, 26.41% had a WER of between 11 and 15, 5.81% had a WER of between 16 and 20, and 1.45% of the generated questions had a WER of more than 21. These results were obtained from the experiment.

An Automatic Question Generation System using the "Text-toText Transfer Transformer" (T5) model was proposed by Saichandra Pandraju et al. [2]. The model generates questions from texts and tables. The dataset creation process involves modifying the ToTTo dataset to include tables and their corresponding highlighted cells, along with questions generated from the descriptions of the tables. The project's goal is to develop a modified version of the ToTTo dataset called TabQGen, which will enable questions to be generated from tables. The learning rate was set to a linear schedule warmup using an AdamW optimizer of 1e-4. Using NIST, BLEU, ROUGE-L, and METEOR as metrics, the experiment's performance was evaluated.

An Automatic Question Generation Model [3] based on a Deep Learning Approach that generates Wh-questions with different difficulty levels from uploaded PDF documents. The system aims to assist in educational applications by automating the process of question generation from textual content. This model gets a BLEU-4 score of 11.3.

A Reinforcement Learning and encoder-decoder framework-based Answer-driven Deep Question Generation (ADDQG) model [7]. Deep question generation (DQG) uses reasoning across several publications to produce sophisticated queries. ADDQG beats state-of-the-art models in both automatic and human evaluations, according to extensive trials conducted on the HotpotQA dataset. Comparing the experiment's results to the SemQG model, the average improvement was 2.83, 1.27, and 1.15 points for BLEU-4, METEOR, and ROUGE-L, respectively. The main distinction between ASs2s-a and ADDQG appears to be that the model architecture is better suited for complex question generation. Both of them encode the answer and document separately, obtaining an average improvement of 4.95, 3.11, and 4.88 points in terms of BLEU-4, METEOR, and ROUGE-L, respectively.

Question Generation Pre-Training for Text Generation (QURIOUS) was developed by Shashi Narayan et al. [8].The main benefits of this approach were that: (i) it was simple to gather a large amount of data for question generation from community quality assurance platforms like Quora, Yahoo Answers, and Stack Overflow; and, more significantly, (ii) text generators that were trained to produce questions that could be answered from passages or documents would identify important terms or concepts in the input and would learn to summarize and synthesize the information. Evaluations of the trial were conducted both automatically and by humans. In terms of both automatic and human evaluations, the model produced summaries and questions that were more natural and informative when tested for answer-focused question creation and summarizing tasks.

Katherine Stasaski et al. [9] enhanced the existing causal extraction system, which extracts causes and effects from unstructured text using a set of syntactic rules, to create a pipeline to produce and assess cause-effect questions straight from a text. Two English datasets were used to assess the approach: the Textbook Question Answering (TQA) dataset, which included middle school science textbooks, and SQuAD Wikipedia articles. A transformer-based quality assurance model from the Huggingface Transformers library, which is BERT-large fine-tuned on SQuAD 2.0 with whole-word masking, was employed to evaluate the generated questions' quality. The TQA dataset's F1 score for the best QG-QA pair of clause types is 0.59, whereas the Squad Dataset's F1 score is 0.63.

To produce questions from a passage and questions that are answered from the passage, a Natural Question Generation (QG) model [10] was proposed. Using a Graph Neural Network at each stage of the decoding process, they create the

Iterative Graph Network-based Decoder (IGND) to simulate the prior generation. Based on SQuAD and MARCO datasets, experimental results show that the model performs better than the state-of-the-art models for sentence-level QG tasks. With a score of 20.33, BLEU-4 is considered the primary evaluation metric for text creation. This model produces the best outcomes. Three criteria were used to evaluate human evaluation: answerability, relevance, and fluency. The results showed that the model produced output that was of a higher caliber than what was needed.

An ASQ: Automatically Generating Question-Answer Pairs using AMRs [11] automatically mines questions and answers from a sentence using the Abstract Meaning Representation (AMR). They used the AMR Annotation Release 2.0 Corpus (39,260 sentences) and the AMR annotated Little Prince Corpus (1,562 sentences) to run ASQ to generate a question-answer pair that carried out a qualitative assessment of the outcomes from AMR 2.0. Seventy phrases were chosen at random from the AMR 2.0 test split to serve as the test set for error analysis in human evaluation. This evaluation is based on the quality assessment and error propagation from AMR parsers.

AnswerQuest: A System for Generating Question-Answer Items from Multi-Paragraph Documents was proposed by Melissa Roemmele et al. [12]. This model extracts questions directly from the given reference text. For the model's encoder and decoder layers, they use the Transformer architecture, and they improve the decoder by adding a copy mechanism. One sentence serves as the encoder input, while the question serves as the decoder output. The input sentence includes the question's answer. Regarding ratings, the outcomes mirror the automated assessment: RULEMIMIC questions receive higher ratings than STANDARD questions, and AUGMENTED questions receive higher ratings than RULEMIMIC questions. Compared to the HUMAN questions, every set of produced questions has a substantially lower rating. Overall, these findings once more demonstrate the value of adding automatically created questions to the training set. As a result, the AUGMENTED model is currently used in their demo.

3 PROPOSED ARCHITECTURES

This section discusses the proposed architecture of the model. The QAG model, as shown in Figure 2, integrates pdfplumber, BERT, RAKE, re-trained T5 and DistilBERT in an efficient workflow. Starting with text extraction, BERT summarizes the document, RAKE identifies keywords and the T5 model generates questions. DistilBERT produces accurate answers, showcasing the synergy of advanced NLP models. Evaluated with the RQUGE score, the model ensures a comprehensive assessment of question-and-answer quality.

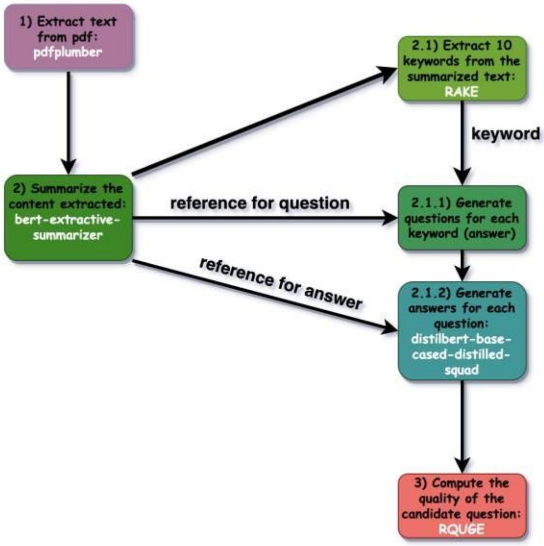
1) Extract text from pdf: pdfplumber
2.1) Extract 10 keywords from the summarized text: RAKE
keyword
2) Summarize the content extracted: bert-extractive-summarizer
reference for question
2.1.1) Generate questions for each keyword (answer)
reference for answer
2.1.2) Generate answers for each question: distilbert-base-cased-distilled-squad
3) Compute the quality of the candidate question: RQUGE

Figure 2. Workflow of the Proposed Model

The proposed Question Answer Generation (QAG) architecture, as illustrated in Figure 3, is crafted to seamlessly transform user-provided PDF files into meaningful questions and answers. Delving into open-source tools such as Hugging Face and Gradio, the paper outlines a streamlined pipeline that harnesses the power of cutting-edge NLP models. Below is a detailed overview of each step within the proposed architecture:

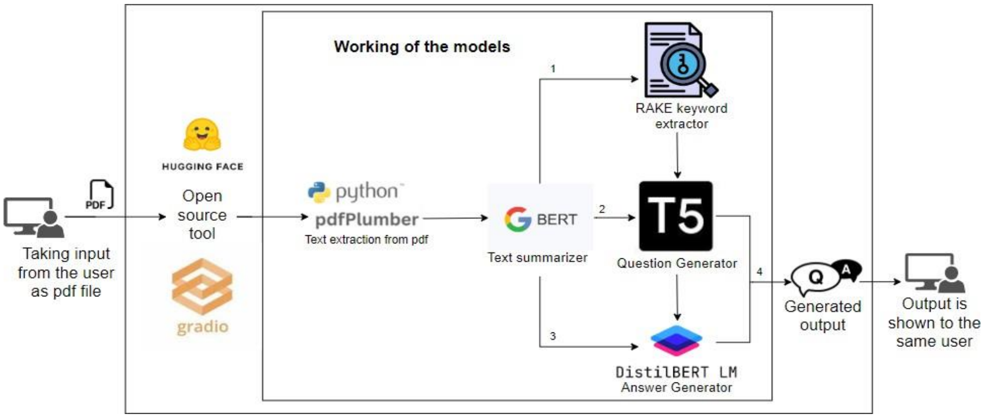

Figure 3. The Architecture of the Proposed Model

The process begins with the user providing a PDF file as input. This ensures the flexibility and user-friendliness of the system, allowing individuals to extract valuable insights from PDF documents effortlessly. Delving into the integration of state-of-the-art NLP models, two powerful open-source libraries, Hugging Face and Gradio, are leveraged to facilitate the architecture. Using the pdfplumber library, textual content is extracted from the user- provided PDF file. Pdfplumber is chosen for its reliability and efficiency in accurately converting PDF content into machine-readable text. The extracted text undergoes processing through a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model. BERT's contextual understanding enables the generation of concise and informative summaries of the input text, capturing essential information. Rapid Automatic Keyword Extraction (RAKE) is employed to identify key concepts and terms within the generated summary, enhancing the precision of subsequent question generation by pinpointing crucial information.

A re-trained T5 model is incorporated into the pipeline for question generation. The model is fine-tuned using a custom dataset and for each identified keyword, it generates questions related to the input text. This step ensures a focused and contextually relevant set of questions. Subsequently, the generated questions are paired with the original text and the DistilBERT model is employed for answer generation. DistilBERT, a more lightweight variant of BERT, efficiently produces contextually relevant answers. The final output consists of the questions generated for each keyword and their corresponding answers. This user-friendly output format ensures that individuals interacting with the system receive meaningful insights derived from the PDF document. Gradio, a user interface creation library, enhances the overall accessibility of the system. It facilitates a seamless interaction between the user and the QAG pipeline, providing a graphical interface for uploading PDF files and viewing the generated questions and answers.

4 EXPERIMENTAL SETUPS

4.1 Hyperparameter Tuning and RQUGE Evaluation

To systematically explore the impact of different hyperparameter configurations on the performance of the model for question-answering on the SQuAD dataset, a series of experiments were conducted. The key hyperparameters under investigation were the learning rate (LR), batch size (BS) and the number of training epochs (Epochs). The primary evaluation metric for assessing model performance was the RQUGE score.

The following hyperparameter combinations were explored:

Experiment	Learning Rate	Batch Size	Max Epochs	RQUGE Score
1	3e-4	4	1	4.27
2	5e-4	4	1	4.21
3	1e-4	8	1	3.79
4	1e-4	4	2	4.26
5	3e-4	4	1	4.32

Table 1. Hyperparameter Combinations

4.2 Visualization

The results of the experiments are visualized in Figure 1 below. Each bar represents a different hyperparameter configuration, with the x-axis indicating LR, BS and Epochs. The height of the bars corresponds to the achieved RQUGE score. Importantly, the blue bars represent different experimental configurations, while the green bar represents the configuration with the highest RQUGE score (Experiment 5).

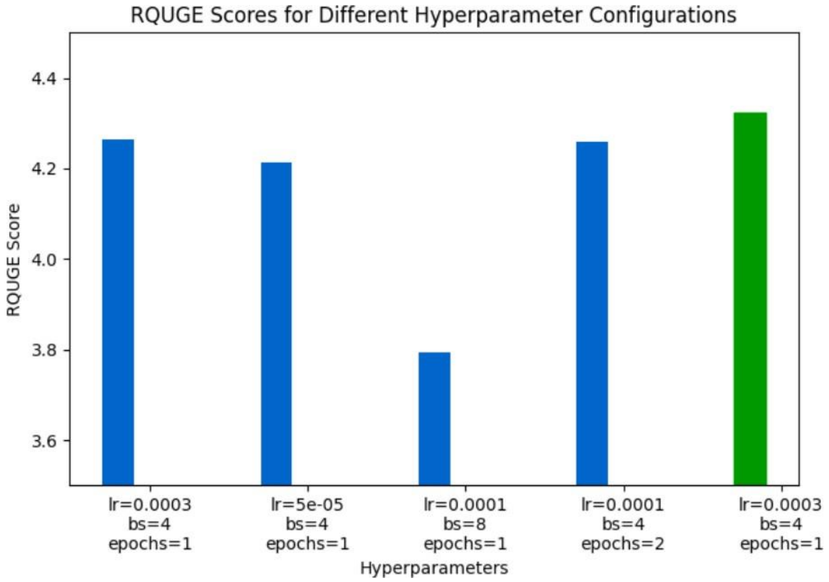

Figure 1. RQUGE Scores for Different Hyperparameter Configurations

The visualization clearly illustrates the comparative performance of each configuration in terms of the RQUGE metric. Experiment 5, indicated by the green bar, stands out as the most successful configuration with the highest RQUGE score among all experiments. These experiments and visualizations guide the understanding of the impact of hyperparameter choices on the model's performance, providing valuable insights for future iterations and improvements.

5 RESULTS

5.1 Model Comparison

In this section, the comparative results of the proposed model against an existing baseline model on the task of question-answering using the SQuAD dataset are presented. The evaluation metric utilized for this comparison is the RQUGE score, a measure of the effectiveness of the model in generating relevant and high-quality responses.

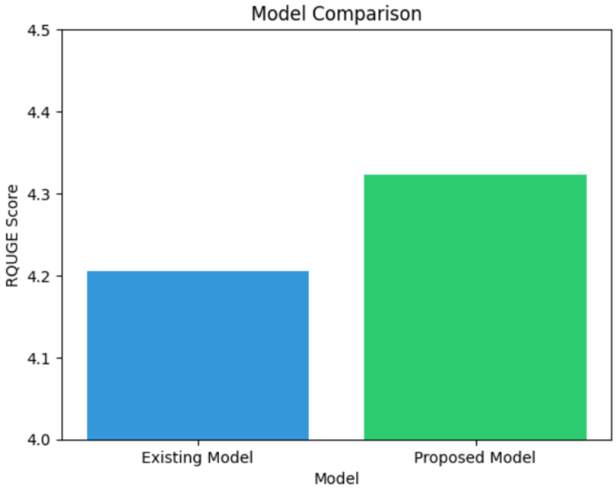

Figure 4. RQUGE Scores

In Figure 4, the bar chart vividly illustrates the comparative RQUGE scores between the existing and the proposed models. Notably, the existing model obtained a commendable RQUGE score of 4.20, while the innovative approach propelled the proposed model to a significantly higher RQUGE score of 4.32.

This marked difference in RQUGE scores accentuates the prowess of the proposed model in generating responses that not only exhibit heightened accuracy but also contextual relevance. The model's performance stands out as superior when juxtaposed with the baseline model, showcasing its capacity to excel in the intricate task of question-answering. The observed uptick in the RQUGE score serves as a robust validation of the model's potential to elevate the efficiency and efficacy of question-answering tasks.

The experiments were meticulously conducted under consistent conditions, ensuring a fair evaluation of the models' capabilities. The RQUGE scores provided in the bar chart serve as a quantitative benchmark, offering insights into the models' respective strengths. The higher RQUGE score achieved by the proposed model is indicative of its proficiency in generating responses that closely align with human-like comprehension. This alignment underscores the model's potential to contribute significantly to advancing question-answering systems.

The success of the model in surpassing the baseline establishes a compelling narrative for its prospects. The encouraging results invite further exploration and consideration for integration into real-world applications, particularly those demanding precision and relevance in responses to user queries. This achievement positions the proposed model as a promising solution for enhancing question-answering systems in diverse and practical settings.

5.2 Human evaluation

In a crucial step toward assessing the practical efficacy of the question-answering model, a comprehensive human evaluation survey was initiated. Feedback was solicited from participants tasked with rating the quality of the generated answers on a 5-point scale. The outcomes of this survey as in Figure 5 revealed an overwhelmingly positive response, signifying a unanimous vote of confidence from participants. An astonishing 46.7% of respondents bestowed the highest possible rating of 5, while an additional 43.3% expressed robust approval, assigning a rating of

4. This resounding endorsement underscores a shared consensus among users, affirming the model's capacity to deliver answers that are not only accurate but also contextually relevant, thus validating its effectiveness in real-world scenarios.

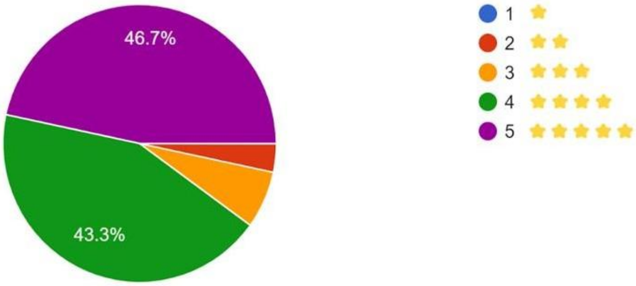

Figure 5: Pie Chart Depicting Human Evaluation Results

These survey findings, when coupled with quantitative metrics such as the RQUGE score, underscore the practical applicability of the model. The user-centric feedback gleaned from the survey provides invaluable insights into the model's performance from the perspective of those actively engaging with it. With nearly 90% of participants providing ratings indicative of high satisfaction, the model emerges as a promising solution poised to address real-world challenges in information retrieval and user interaction. These encouraging results further fortify the potential impact of the proposed model, positioning it as a catalyst for delivering heightened user experiences and reliable solutions in the realm of question-answering across diverse domains. The convergence of positive user sentiments and robust quantitative metrics underscores the model's promise and its ability to make substantive contributions to the field.

6 CONCLUSION AND FUTURE WORKS

In conclusion, the research introduces an innovative Question Answer Generation (QAG) architecture tailored for insightful information extraction from PDF documents. Open-source tools, including Hugging Face and Gradio, are utilized to seamlessly integrate advanced NLP models such as BERT, RAKE, T5 and DistilBERT into an intuitive pipeline. The user-friendly system facilitates seamless PDF uploads, providing concise summaries and delivering contextually relevant questions and answers.

The experimental evaluation, conducted on the SQuAD dataset, highlights the architecture's efficacy in accurately summarizing and extracting information from diverse PDF sources. Integration of RQUGE metrics showcases the system's capability to generate high-quality questions and answers. The user interface, powered by Gradio, enhances accessibility, ensuring the QAG system is both robust and user-friendly.

Looking ahead, the future work aims to enhance the versatility of the QAG model. The plan includes incorporating support for multi-document types to seamlessly handle diverse information sources. Additionally, expanding language capabilities to offer multi-lingual support is on the agenda, broadening the applicability of the system across global contexts. Furthermore, there are plans to diversify question types, incorporating multiple-choice questions (MCQs), filling in the blanks and exploring the complexity of questions to cater to varying educational and informational needs. Additionally, optimization efforts will be undertaken to enhance the efficiency and speed of the proposed QAG model, ensuring optimal performance across different scenarios.

7 REFERENCES

[1]. Kettip Kriangchaivech and Artit Wangperawong, “ Question Generation by Transformers”. U.S. Bank 1095 Avenue of the Americas New York, NY 10036.

[2]. Saichandra Pandraju and Sakthi Ganesh Mahalingam, Answer-Aware Question Generation from Tabular and Textual Data using T5”, International Journal of Emerging Technologies in Learning, vol.16, no.18, 2021.

[3]. Hala Abdel-Galil, Mai Mokhtar and Salma Doma, ‘AUTOMATIC QUESTION GENERATION MODEL BASED ON DEEP LEARNING APPROACH’ in International Journal of Intelligent Computing and Information Sciences.

[4].https://www.udemy.com/course/question-generation-using-natural-language-processing/learn/lecture/23905916#overview

[5]. Dataset: https://huggingface.co/datasets/squad

[6]. Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson and Marzieh Saeidi, ‘RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question’

[7]. Liuyin Wang, Zihan Xu, Zibo Lin, Hai-Tao Zheng and Ying Shen, “Answer-driven Deep Question Generation based on Reinforcement Learning”, 28th International Conference on Computational Linguistics, 2020.

[8]. Shashi Narayan, Gonc¸alo Simoes, Ji Ma, Hannah Craighead and Ryan Mcdonald, “QURIOUS: Question Generation Pretraining for Text Generation”, 2020.

[9]. Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, Marti A. Hearst, “Automatically Generating Cause-and-Effect Questions from Passages”, Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, 2021.

[10]. Zichu Fei, Qi Zhang and Yaqian Zhou,” Iterative GNN-based Decoder for Question Generation”, Conference on Empirical Methods in Natural Language Processing, 2021.

[11]. Geetanjali Rakshit and Jeffrey Flanigan,” ASQ: Automatically Generating Question-Answer Pairs using AMRs”, 2021.

[12]. Melissa Roemmele, Deep Sidhpura, Steve DeNeefe and Ling Tsou, “AnswerQuest: A System for Generating Question-Answer Items from Multi-Paragraph Documents”, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2021.