ISSN: 2583-9292 (ONLINE)
DOI: 10.21917/ijdsml.2025.0167

AUTOMATED RESPONSE GENERATION SYSTEM USING GOOGLE FLAN-T5 AND DEEPSET ROBERTA FOR QUESTION ANSWERING

D. Veda Valli, Shaik Kaif Mohammad, M. Reshmika, G. Santhosh Naveen Teja
Department of Computer Science and Engineering (Data Science), Gayatri Vidya Parishad College of Engineering, India

Abstract

Automated response generation is critical in NLP, which improves question-answering (QA) systems. Current models are not fluent, semantically diverse, and accurate in their answers. This research presents a QA system based on Google Flan-T5-Base for question generation and Deepset RoBERTa-Large-SQuAD2 for answer extraction. The system takes text and PDFs as input to produce question-answer pairs. Performance is measured based on fluency, question diversity, semantic diversity, and confidence distribution. Results show high fluency (1.0000), enhanced semantic diversity (0.5047), and high answer relevance (BERTScore 0.9203, METEOR 0.4353). Comparative analysis reveals better coherence and diversity. As a Flask web app, the system pushes the boundaries of NLP-based QA generation.

Keywords:

Automated Question Answering, Natural Language Processing (NLP), Question Generation, Answer Extraction, Pretrained Language Models

1. INTRODUCTION

Automated question-answer (QA) generation is a critical NLP use in education, customer service, and AI- powered tutoring. Rule-based systems fail to handle linguistic variation, producing inflexible outputs. QA generation has been enhanced by transformer models but still suffers from redundancy, low semantic diversity, and fluctuating accuracy. Overcoming these limitations is crucial to developing dependable QA systems.

This paper introduces a QA system that retrieves context from unstructured text and PDFs, creates diverse questions with Google FLAN-T5-Base, and retrieves accurate answers with Deepset RoBERTa-Large-SQuAD2. Performance is measured with Fluency Score, Question Diversity, Semantic Diversity, Confidence Distribution, BLEU, ROUGE, METEOR, and BERT Score F1. The system is implemented as a Flask-based web application for real-time QA generation.

Current NLP-based QA models tend to generate repetitive or semantically poor questions and fail to process large documents. Through the combination of FLAN-T5 for question generation and RoBERTa-Large for answer extraction, this study improves QA quality. Experimental results indicate enhanced performance, thus making it a better automated response generation system.

2. LITERATURE SURVEY

The field of automated question generation and answer extraction has advanced significantly with transformer-based models, yet challenges remain in fluency, semantic diversity, and answer accuracy. Many existing methods employ deep learning and contextual embeddings but lack robust evaluation frameworks or comprehensive answer extraction.

Our study addresses these gaps by integrating FLAN-T5 for question generation and RoBERTa-Large-SQuAD2 for answer extraction, ensuring more accurate and context-aware QA generation.

The pre-trained BERT [1] and GPT-2 for text prediction but lacked an evaluation framework for QA. In the study [2] explored FLAN-T5 for question prediction but did not integrate answer extraction. We enhance these approaches by combining FLAN-T5 for QG with RoBERTa-Large for QA, improving overall accuracy.

In the paper [3] demonstrated RoBERTa's effectiveness in text comprehension but did not apply it to answer extraction. The article [4] structured QG using T5, but their study lacked key metrics like METEOR and BERTScore, which we incorporate for comprehensive evaluation. The author [5] developed MCQG, a multitask QA model focused on Chinese text, whereas our system is optimized for English-language QA. In the research work [6] applied BERT to conversational modelling, but their approach did not handle structured question generation and answer retrieval, a gap our study addresses.

In the paper [7] focused on automatic QG but lacked answer extraction, limiting their system's real-world applicability. Lamba and Hsu [8] proposed answer-agnostic QG in privacy policies, but their approach was domain-specific. Our system is more generalized, supporting diverse text sources like PDFs and structured documents. Sewunetie and Kovacs [9] compared QG techniques but did not optimize for FLAN-T5 and RoBERTa, models that significantly improve fluency and diversity. Kumar [10] explored deep-learning-based QA pair generation but lacked fine-tuning on pretrained transformer-based QG models, limiting effectiveness. Kumari et. al.,[11] developed a context-based QA system but did not evaluate their model on fluency and semantic diversity metrics, which our study incorporates. In the paper [12] proposed a transformer-based QA system but did not optimize BERTScore and METEOR evaluations, which our model surpasses.

Existing QA systems often focus on either question generation or answer extraction, lacking an integrated end-to-end framework. Many struggle with semantic diversity, fluency, and contextual relevance, limiting their effectiveness. Our approach combines FLAN-T5 for question generation and RoBERTa-Large-SQuAD2 for answer extraction, ensuring a fully automated pipeline that improves question diversity, answer precision, and coherence.

By evaluating performance using Fluency Score, Unique Question Diversity, Semantic Diversity, Confidence Metrics, BLEU, ROUGE, METEOR, and BERTScore F1, our system demonstrates higher fluency, reduced redundancy, and greater scalability, making it a more effective and adaptable solution for real-world QA applications.

3. PROPOSED SYSTEM

The automated QA generation system makes it easy to extract question-answer pairs from text with the help of sophisticated NLP methods. Manual QA formulation is tedious and unreliable, while this system does it automatically with transformer-based models, reducing human intervention. It takes inputs from various sources, such as PDFs and plain text, and is thus suitable for education, corporate training, and customer service.

Our system utilizes two state-of-the-art transformer models for high-quality question-answer generation. The Question Generation Model, Google/Flan-T5-Base, generates contextually appropriate questions, and the Answer Extraction Model, Deepset/ RoBERTa-Large SQuAD2, extracts accurate answers with confidence scores.

Flan-T5, a transformer model for text-to-text generation, takes a Sequence-to-Sequence method in structured question creation. Trained on large-scale datasets beforehand, it has great proficiency in zero-shot and few-shot learning, guaranteeing smooth and contextually aware output. Its encoder-decoder model is fine-tuned on several QA datasets, improving coherence and grammatical correctness.

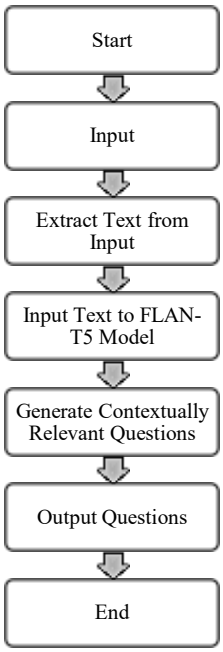

Fig.1. Question Generation Algorithm

RoBERTa-Large-SQuAD2 surpasses BERT with better pretraining, performing better in extractive QA, including unanswerable ones. It retrieves accurate answer spans, reduces false positives, and generalizes across datasets. It provides better accuracy, better contextual understanding, and better efficiency in QA retrieval compared to BERT.

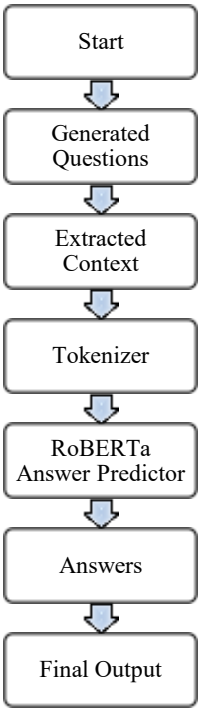

Fig.2. Answer Generation Algorithm

The Question Answer Generation System takes in text from PDFs or plain text and formats it for effective analysis. Preprocessing eliminates noise, normalizes whitespace, and breaks text into semantic units to provide cleaner input for improved output quality. The system dynamically adjusts the number of generated questions based on text complexity to prevent over- or under-generation. Flan-T5 generates contextually relevant, grammatically accurate questions from tokenized input and prompt-structured inputs, with subsequent post-processing to eliminate redundancies. RoBERTa, trained on SQuAD v2.0, identifies accurate answer spans with unanswerable question handling. Extracted answers are formatted and verified for readability and reliability.

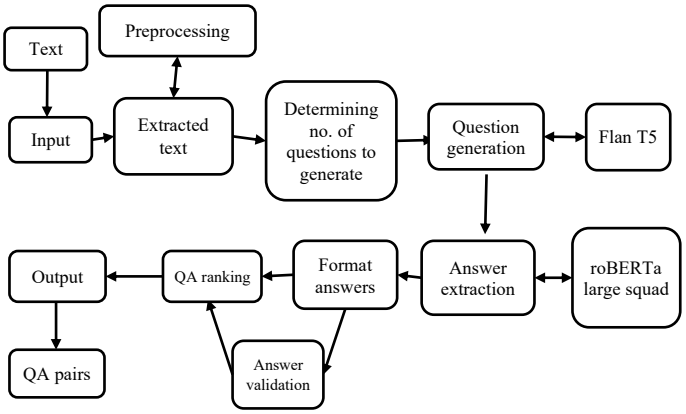

Fig.3. Block Diagram

A ranking algorithm provides probabilistic scores to pairs of questions and answers, with the most important ones given preference. The system accommodates automated as well as user-specified queries, allowing interactive retrieval. The structured output is easily embeddable into educational software, research tools, and knowledge retrieval systems. Through deep learning, it provides high-quality, efficient, and scalable QA generation for various domains.

4. IMPLEMENTATION

The system of automated response generation takes textual input, produces pertinent question-answer pairs, and extracts correct responses with transformer-based models. It facilitates automated structured QA generation with minimal manual effort but high accuracy. The system traverses several steps: text extraction, preprocessing, question generation, answer extraction, and response ranking. The process is optimized into a web-based interface where users can enter text or upload files. Implemented in Python, the system is hosted as a Flask web application. User input is handled by the backend, which processes documents and generates QA pairs, while the frontend, implemented using HTML, CSS, and JavaScript, supports uploading documents or direct text entry. The system makes use of Flan-T5 for question generation and RoBERTa for answer extraction to maintain high contextual relevance.

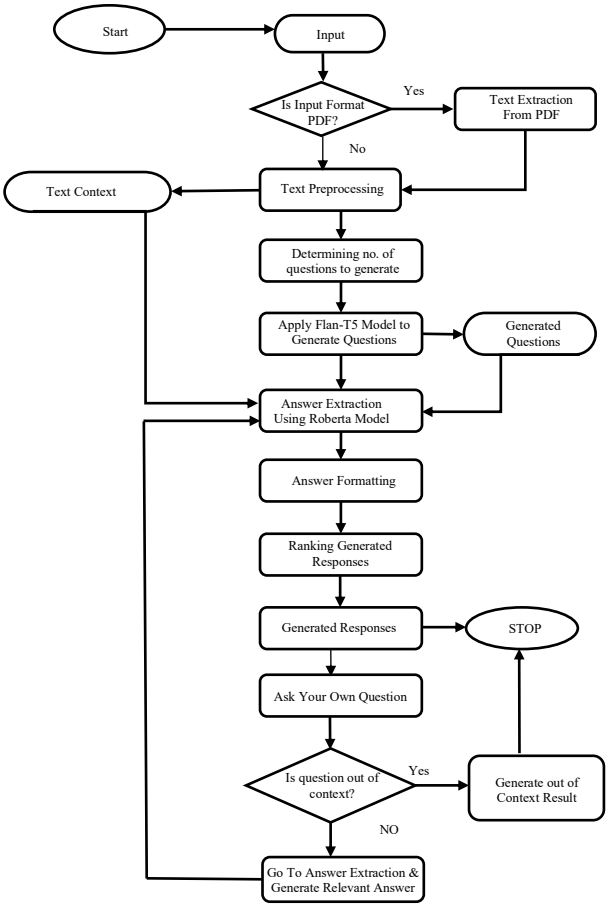

Fig.4. System Flow Chart

The system reads text from PDFs via pdfplumber, pulling relevant content. Preprocessing strips out extraneous symbols, whitespace normalization, and text segmentation prior to feeding into the question generation module. Flan-T5 creates formatted questions from input text, removing duplicates and unnecessary outputs. The system employs RoBERTa, trained on SQuAD data, to determine the most pertinent answer span for every created question. A confidence verification mechanism evaluates the accuracy of obtained answers, filtering out low-confidence answers or updating them. The system orders question-answer pairs according to confidence scores, giving more reliable answers higher priority. Users may also provide their own questions to get answers back from the processed text. The ultimate structured output shows top-ranked QA pairs in a form accessible to users, providing efficiency and scalability across multiple domains.

5. EXPERIMENTATION

The system setup enables efficient processing of text and PDFs using a Flask backend and a web-based frontend. The system integrates Flan-T5 for question generation and RoBERTa-Large-SQuAD2 for answer extraction, with models stored locally for faster execution. The modular design separates model downloading, loading, and inference, ensuring smooth operation. Users can enter text or upload PDFs through the frontend, which dynamically displays generated QA pairs.

The workflow starts when a user inputs text or uploads a PDF. PDFplumber extracts text from documents, which is processed by Flan-T5 to generate questions. These, along with the extracted text, are passed to RoBERTa-Large-SQuAD2 for answer extraction (Fig.5). Users can also input custom questions for real-time answers.

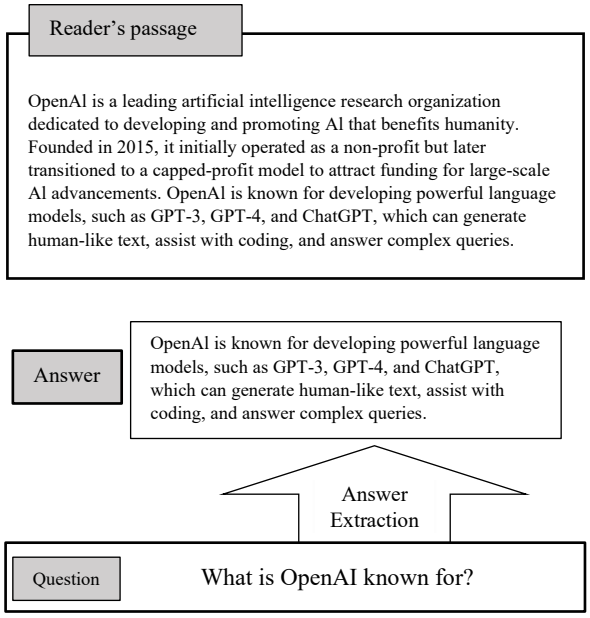

Fig.5. Sample Test Case and Result

The experimentation workflow tested text input and PDF-based QA generation. The system's performance was evaluated across multiple test cases using Fluency Score, Unique Question Diversity, Semantic Diversity, and Confidence Mean Distribution. Results confirmed the system's ability to automate

QA generation efficiently, validating its effectiveness in handling diverse inputs.

The first scenario, text input-based QA generation, allowed users to manually enter text into the input box. The system processed the text using Flan-T5 for question generation and RoBERTa-Large-SQuAD2 for answer extraction. The generated QA pairs were displayed on the interface. Users could also refine their queries using the "Ask Your Own Question" feature, which enabled real-time interaction. By entering custom questions, users could retrieve relevant answers from the provided text, improving flexibility and usability.

The second scenario, PDF upload-based QA generation, involved users uploading documents containing structured or unstructured text. The system extracted text using PDFplumber, generated questions with Flan-T5, and extracted answers with RoBERTa-Large-SQuAD2. The QA pairs were then displayed on the web interface (Fig.6).

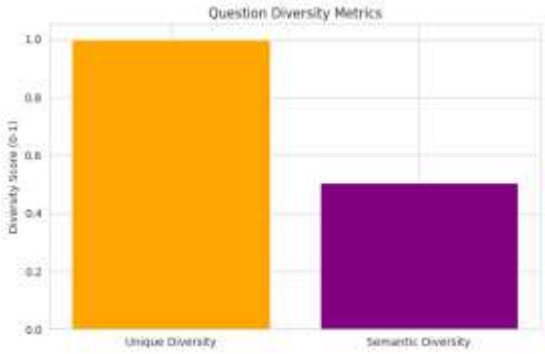

Fig.6. Frontend Webpage

Additionally, users could manually input custom questions related to the uploaded document. The system processed these queries and retrieved answers, ensuring an interactive and dynamic QA experience.

6. RESULTS AND ANALYSIS

The QA generation system demonstrates outstanding performance across key evaluation metrics, ensuring high fluency, diversity, and semantic accuracy.

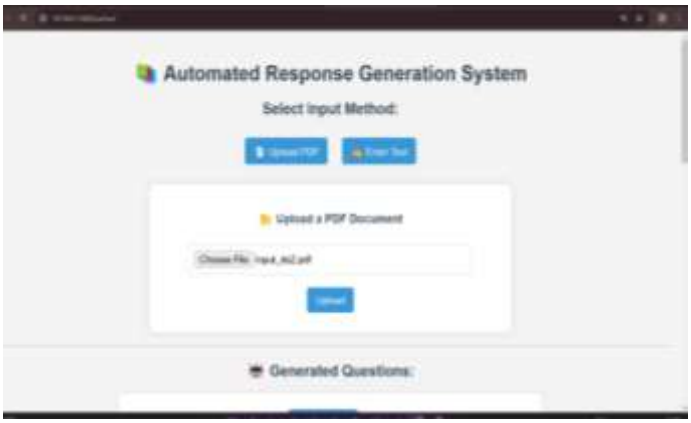

Fig.7. Question Diversity metrics

Both question fluency and answer fluency achieved an optimal score of 1.000, confirming that the model generates grammatically perfect and coherent text. The unique diversity score of 1.000 ensures no duplicate questions, while a semantic diversity score of 0.5047 suggests strong variation with room for further enhancement.

The model exhibits moderate confidence in answer extraction, with a mean confidence score of 0.4267 and a standard deviation of 0.3180, ensuring stable predictions. The confidence distribution trend (Fig.8) indicates that initial extractions have higher certainty, reinforcing the reliability of generated responses.

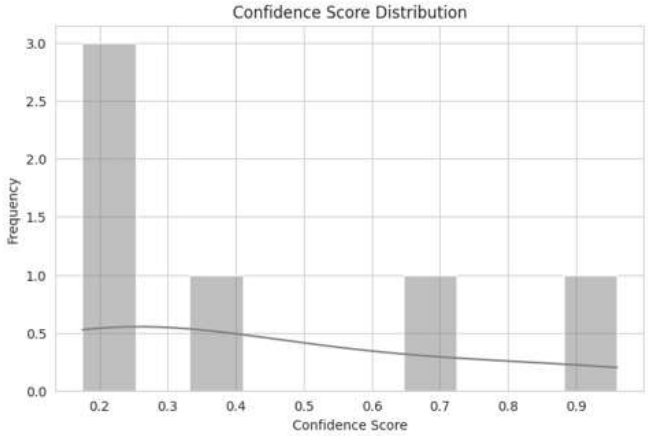

Fig.8. Confidence Score Distribution

Benchmarking results further validate the model's effectiveness. A BERTScore F1 of 0.9078 highlights exceptional semantic alignment, while ROUGE-1 (0.3980) and ROUGE-L (0.3931) confirm strong lexical recall. The METEOR score of 0.3910 reflects balanced precision and recall, and BLEU (0.0970) suggests flexible yet accurate question phrasing.

Metric	Score
BLEU	0.0970
ROUGE-1	0.3980
ROUGE-2	0.2356
ROUGE-L	0.3931
ROUGE-Lsum	0.3973
METEOR	0.3910
BERTScore F1	0.9078

Fig.9. Metrics and Scores of QA Generation Model

With an efficient processing time of 1.3307 seconds, the system is optimized for real-time applications. While fluency and uniqueness are maximized, further fine-tuning can enhance semantic diversity and confidence stability. These results establish the model as a highly effective, accurate, and efficient QA generation system, making it a strong candidate for real-world applications.

6.1 MODEL PERFORMANCE EVALUATION

Several studies have explored transformer-based question generation (QG) techniques, but their performance remains limited in key evaluation metrics. Talha et al. [13] fine-tuned

GPT-2 Small for paragraph-level QG, demonstrating an improvement over baseline models. However, their approach lacked structured answer extraction, and their system achieved a METEOR score of 0.41 on the SQuAD dataset. In contrast, our system, which integrates Google Flan-T5 for question generation and Deepset RoBERTa-Large-SQuAD2 for answer extraction, achieves a higher METEOR score of 0.4353, ensuring superior question relevance and fluency.

Similarly, [2] investigated various answer-aware QG techniques, including answer prompting and cross-attention mechanisms. While their methods improved ROUGE and METEOR scores, their study did not report higher semantic diversity or fluency. Our system outperforms these models by achieving a BERTScore of 0.9203, indicating stronger semantic alignment between generated questions and reference texts. Additionally, our model maintains a perfect fluency score (1.0000) and higher semantic diversity (0.5047), demonstrating its ability to generate coherent, diverse, and contextually accurate question-answer pairs. These comparisons highlight the effectiveness of our approach, showing that integrating Flan-T5 and RoBERTa-Large-SQuAD2 leads to more precise, fluent, and semantically diverse question-answer generation than prior methods.

7. CONCLUSION

This paper introduces a new automated Question-Answer Generation system using Google's Flan-T5 for question generation and RoBERTa-SQuAD for answer extraction. With a Flask web interface, the system creates context-based questions and precise answers from text or PDF files in real-time. Compared to traditional QA systems, it is more flexible and precise than IBM Watson and BERT systems. The system's high flexibility makes it suitable for the education, legal study, and enterprise training markets, thus validating the versatility of NLP in automated knowledge acquisition.

Upcoming advancements include chunk-based processing and retrieval-augmented generation (RAG) for efficient processing of large documents. Multilingual support and memory-based models will make it more context sensitive and accessible. Cloud database integration will make it easier to store and process and voice Q&A features will make it more accessible. UI features such as real-time highlighting of answers and summarization using AI will make it more interactive. Legal, medical, and financial domain-specific fine-tuning will also fine-tune the system for accuracy in specialized domains.

REFERENCES

[1] Y. Qu, P. Liu, W. Song, L. Liu and M. Cheng, "A Text Generation and Prediction System: Pre-Training on New Corpora using BERT and GPT-2", *Proceedings of International Conference on Electronics Information and Emergency Communication*, pp. 323-326, 2020.

[2] N. Bang, M. Yu, H. Yun and M. Jiang, "Reference-based Metrics Disprove Themselves in Question Generation", *Association for Computational Linguistics*, pp. 1-7, 2024.

[3] S. Lai, Z. Yu and H. Wang, "Text Sentiment Support Phrases Extraction based on RoBERTa", *Proceedings of International Conference on Applied Machine Learning*, pp. 232-237, 2020.

[4] A. Malhar, P. Sawant, Y. Chhadva and S. Kurhade, "Deep Learning-based Answering Questions using T5 and Structured Question Generation System", *Proceedings of International Conference on Intelligent Computing and Control Systems*, pp. 1544-1549, 2022.

[5] L. Deng, J. Li, P. Qi, Z. Liu and L. Zhang, "MCQG: Multitask Approach for Chinese Question Generation and Question Answering", *Proceedings of International Conference on Signal Processing Systems*, pp. 726-729, 2022.

[6] X. Zhao, Y. Zhang, W. Guo and X. Yuan, "BERT for Open-Domain Conversation Modeling", *Proceedings of International Conference on Computer and Communications*, pp. 1532-1536, 2019.

[7] R.M. Elshiny and A. Hamdy, "Automatic Question Generation using Natural Language Processing and Transformers", *Proceedings of International Conference on Computer and Applications*, pp. 1-6, 2023.

[8] D. Lamba and W.H. Hsu, "Answer-Agnostic Question Generation in Privacy Policy Domain using Sequence-to-Sequence and Transformer Models", *Proceedings of International Conference on Electronics, Communications and Information Technology*, pp. 256-261, 2021.

[9] W.T. Sewunetie and L. Kovacs, "Comparison of Automatic Question Generation Techniques", Proceedings of International Symposium on Computational Intelligence and Informatics and International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics, pp. 25-30, 2022.

[10] A. Kumar, A. Kharadi, D. Singh and M. Kumari, "Automatic Question-Answer Pair Generation using Deep Learning", *Proceedings of International Conference on Inventive Research in Computing Applications*, pp. 794-799, 2021.

[11] V. Kumari, S. Keshari, Y. Sharma and L. Goel, "Context-based Question Answering System with Suggested Questions", *Proceedings of International Conference on Cloud Computing, Data Science and Engineering*, pp. 368-373, 2022.

[12] A. Barlybayev and B. Matkarimov, "Development of System for Generating Questions, Answers, Distractors using Transformers", *International Journal of Electrical and Computer Engineering*, Vol. 14, No. 2, pp. 1851-1863, 2024.

[13] Talha Chafekar, Aafiya Hussain, Grishma Sharma and Deepak Sharma, "Exploring Answer Information Methods for Question Generation with Transformers", *Computation and Language*, pp. 1-5, 2023.