Enriching Question Bank using Recurrent Neural Network

Harsha Anand PP, Lekha Treesa Titus, Alfeena Sulaiman, Ajumol PA

[1]Mar Athanasius College of Engineering, Kothamangalam, Kerala, India

---***---

Abstract - —The notable effectiveness of question papers crafted by human experts makes the development of Recurrent Neural Network (RNN) based automated systems a captivating area of study, despite significant advancements in technology. This project aims to create an enriching question paper-generating system, based on RNN, with added features for sorting and personalized recommendations. Automatic question generation for textual inputs is valuable in academics where answering questions helps students to learn and improve their understanding of their field of study. The proposed system automates question collection and provides personalized question papers for students based on performance which helps them to excel in academics. Providing practice question paper sets inculcates various domains of the subject familiarizing with core principles resulting in increasing curiosity and a better understanding of the subject. This comprehensive approach ensures a fair and effective exam preparation experience, fostering continuous improvement while respecting ethical considerations.

***Key Words*: Automated question paper generation system, Recurrent Neural Networks (RNN), Question Bank, LSTM, DuerQues, PyPDF**

1. INTRODUCTION

Generating question papers using Recurrent Neural Networks (RNNs) is a modern approach that leverages advanced technology to streamline the process of exam preparation in educational environments. Despite significant advancements in technology, there remains notable effectiveness in question papers crafted by human experts, which makes the development of RNNs-based automated systems a captivating area of study. In the realm of educational technology, the emergence of CBTs (Computer Based Exams) has set out for a revolutionary, in aspects of convenience, partiality-free marking system, wastage of materials like papers, etc. Recurrent Neural Networks is a deep learning model trained to process and convert a sequential data input into a specific sequential data output. It can generate questions from a sequence of keywords using a sequence-to sequence-based model. Generating question papers using Recurrent Neural Net works (RNNs) is a modern approach that leverages advanced technology to streamline the process of exam preparation in educational environments. The integration of RNNs in question paper generation holds immense potential. These systems not only automate the laborious task of manual question creation but also offer the flexibility to customize parameters such as question difficulty and topic coverage based on specific educational objectives. This personalized approach enhances the overall quality of assessments and contributes to a more engaging and effective learning experience for students and educators alike. Automated question paper generation using Recurrent Neural Networks (RNNs) faces the challenge of ensuring diversity, relevance, and appropriate difficulty levels in the generated questions, which are essential for effective assessments. This problem is addressed by sophisticated algorithms that mimic human question-crafting abilities while leveraging the capabilities of RNNs to analyze patterns, detect correlations, and adapt question difficulty based on learning objectives and student capabilities. The preference for RNN-based systems lies in their ability to automate manual question creation, maintain fairness and relevance in assessments, and provide scalable solutions that align with diverse educational contexts and standards

2. RELATED WORKS

2.1 Recurrent Neural Network

Recurrent Neural Networks (RNNs) stand out in the realm of Artificial Intelligence and Machine Learning for their ability to process sequential data. Unlike traditional feed-forward Neural Networks, RNNs possess a unique architecture that allows them to maintain a memory of past inputs, making them particularly well-suited for tasks such as time series prediction, language modeling, and speech recognition. DuerQues[1], a CNN-based intelligent system was introduced to achieve skill-oriented automatic interview question generation and recommendation. The system paved the way to check such a possibility of textual data in RNN Models. It targets and exploits skill-related knowledge from the user-generated con tent of online knowledge-sharing communities (KSCs) and the click-through behaviors of search engine queries. Built to work in large amounts of data, RNN is the recurrent connection, which enables the network to retain information about previous inputs and use it to influence the processing of subsequent inputs. This recurrent nature allows RNNs to exhibit dynamic temporal behavior, making them powerful tools for tasks that involve sequences of data points. One key characteristic of RNNs is their ability to handle inputs of varying lengths. This flexibility makes them invaluable in scenarios where the length of the input sequence may vary, such as natural language processing tasks like sentiment analysis, machine translation, and text generation. Overall, RNNs have revolutionized the field of sequential data processing, offering a powerful framework for modeling complex temporal relationships and driving advancements in areas ranging from natural language understanding to time series forecasting. In our system, we use it for question analysis and selection.

2.2 Long Short-Term Memory (LSTM)

LSTM Networks are a type of Recurrent Neural Network capable of learning order dependence in sequence prediction problems. This is a behavior required in complex problem domains like machine translation, speech recognition, and more. Understanding LSTMs and how concepts like bidirectional and sequence-to-sequence are related to the field can be challenging. LSTM has feedback connections in contrast to traditional feed-forward neural networks. It can analyze whole data sequences in addition to single data points (like photos) (such as speech or video). For instance, LSTM can be used for tasks like linked, unsegmented handwriting identification, speech recognition, and network traffic anomaly detection, or IDSs (Intrusion Detection Systems). A cell, an input gate, an output gate, and a forget gate make up a typical LSTM unit. The three gates control the information flow, and the cell retains values across arbitrary time intervals.

2.3 Natural Language Processing

Natural Language Processing (NLP) is a subfield of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. It encompasses a wide range of tasks, from basic language understanding tasks like sentiment analysis and named entity recognition to more complex tasks such as machine translation, question answering, and text summarization. One of the fundamental challenges in NLP is the ambiguity and complexity of human language. Words and phrases can have multiple meanings depending on context, making it challenging for machines to accurately interpret and generate language. NLP techniques and algorithms aim to address these challenges by leveraging machine learning models, linguistic rules, and statistical methods. As NLP technologies evolve, they hold the potential to revolutionize how we interact with and derive insights from vast amounts of textual data. In our system, we utilize NLP to perform paraphrasing.

2.4 Flask

A versatile Python framework, Flask follows a minimalist philosophy, allowing developers to choose and integrate components as needed, resulting in a flexible and customized development experience. It is designed around the concept of routes, which define the URL patterns for accessing different parts of the application. Extensibility through a wide range of extensions and libraries, covering various aspects of web development, including form handling, authentication, database integration, and more makes Flask unique and popular among web frameworks. Developers can leverage these extensions to add functionality to their Flask applications efficiently. Flask also supports the development of RESTful APIs, making it an excellent choice for building back-end services and microservices. Its lightweight nature and straightforward design make it ideal for rapid prototyping and iterating on web projects. Furthermore, Flask applications can be deployed easily to various hosting platforms, from traditional servers to cloud providers, ensuring scalability and reliability.

2.5 PyPDF

PyPDF offers a wide range of functionalities that enable users to perform various tasks such as extracting text, merging PDFs, splitting PDFs into separate pages, adding watermarks, and more. These capabilities make PyPDF a valuable tool for anyone working with PDF documents, whether for personal or professional use. The user-friendly interface makes it a widely used Python library to handle PDF documents. Programmability is another standout feature, allowing users to manipulate PDF files programmability to automate repetitive tasks or integrate PDF processing into larger applications. This level of automation can significantly improve workflow efficiency and productivity, especially when dealing with large numbers of PDF files or complex PDF-related tasks. Furthermore, PyPDF is equipped to handle encrypted PDFs, ensuring that sensitive information remains secure during processing. Additionally, the library preserves metadata within PDF files, ensuring that essential information such as authorship, creation dates, and document properties is retained throughout the manipulation process. In our system, we utilize PyPDF for extracting questions from PDF files. Users can upload question papers and using PyPDF our system extracts these questions and stores them in a CSV file.

2.6 HTML and CSS

Developed in the late 1980s, HTML is still an essential and powerful tool in the world of web development. HTML consists of a series of tags and attributes that define the structure and content of a web page. The tags in HTML provide the basic structure for a web page, such as headings, paragraphs, images, and links. The simple structure and usability make it widely acceptable. The attributes of the tags provide additional information about the elements, such as size, color, and location. HTML has evolved over the years, with new features being added to provide additional functionality, such as the ability to include multimedia content and create interactive forms. Used with other technologies of CSS and Javascript. CSS is used to separate the presentation of a web page from its content, making it easier to maintain and update. It provides a way for developers to apply styles, such as colors, fonts, and spacing, to a web page in a consistent and organized manner. CSS has several powerful features that allow developers to control the look and layout of a web page. CSS also provides a way to target specific elements on a web page and apply styles to them, allowing developers to create complex and sophisticated designs. CSS is an essential tool for anyone involved in web development and is used in conjunction with HTML to create dynamic and interactive websites. A basic understanding of CSS is essential for anyone who wants to create a website or work with web development, as it provides the means to control the look and presentation of a web page

3. PROPOSED MODEL

The proposed model checks out the possibilities of Recurrent Neural Network Models in enhancing personalized question paper generation that is efficient and adaptable. The model aims to achieve the following:

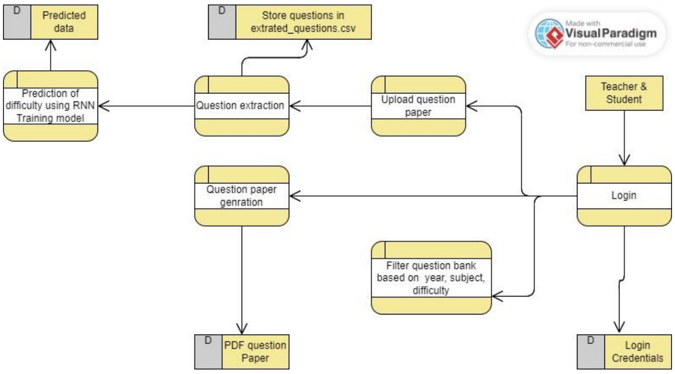

D
Predicted data
D
Store questions in extrated_questions.csv
Made with Visual Paradigm For non-commercial use
Prediction of difficulty using RNN Training model
Question extraction
Upload question paper
Teacher & Student
Question paper genration
Login
Filter question bank based on year, subject, difficulty
D
PDF question Paper
D
Login Credentials

Fig -1: Data Flow Diagram- Enriching Question Bank

1) Enriching question bank, giving users the role of contributors to add up questions by uploading a PDF question paper. It is to get free access to recently arrived question papers.

2) Automated question extraction, collecting questions along with respective course names, course codes, and marks. It reduces the herculean task of manual question entry.

3) Generation of different question paper sets with questions having given difficulty level elucidated from student's CGPA to familiarize and have a conceptual understanding of the selected course.

3.1 Data Collection and Preprocessing

Question papers set in PDF format that belong to different courses under of Bachelor of Technology domain since 2019 are collected from online sources. The dataset contains exam question papers provided by APJ Abdul Kalam Technological University (KTU). More emphasis is given to textual input. Questions along with the respective subject name, subject code and marks are extracted and stored. Frequency is updated on the arrival of questions. Perform text preprocessing steps to clean and normalize the collected data. Irrelevant information is removed.

3.2 Machine Learning Models

RNN model combined with LSTM is employed here, considering the size of data collection and efficiency of processing textual input. The model is designed such that it can figure out the word combinations and corresponding difficulty making it capable of predicting the difficulty levels of new questions (level 1 for easy, level 2 for medium, and level 3 for hard

3.3 Training and Evaluation

Split the dataset into training and testing sets for model training and evaluation, respectively. Train the selected models using the training data and optimize them to achieve the best performance. The model is saved for future purposes.

3.4 Model Integration and Deployment

Integrated the trained question selection and question generation models into a cohesive system. Developed a user-friendly interface that provides the corresponding analysis results. Ensure the system is scalable and can handle real-time or batch processing depending on the application requirements.

3.5 Continuous Improvement

The system monitors the performance of the deployed system and gathers user feedback for improvements. Continuously update the models with new data and adapt to emerging trends or language variations. Stay updated with the latest research and techniques to enhance the system's capabilities. The overall architecture and design may vary based on specific project requirements, available resources, and the complexity of the analysis. It is important to iterate and refine the design based on experimentation, evaluation, and feedback to build an effective and reliable system.

The system monitors the performance of the deployed system and gathers user feedback for improvements. Continuously update the models with new data and adapt to emerging trends or language variations. Stay updated with the latest research and techniques to enhance the system's capabilities. The overall architecture and design may vary based on specific project requirements, available resources, and the complexity of the analysis. It is important to iterate and refine the design based on experimentation, evaluation, and feedback to build an effective and reliable system.

4. DESIGN AND IMPLEMENTATION

4.1 System Architecture

The proposed system consists of the following modules:

1) User login and Verification Module

2) Question Extraction Module

3) Question Paper Generation Module Users are registered with given details of email and other credentials. Registered users are verified based on stored data. Question Extraction Module used the Python library PyPDF from which the contents of the paper are extracted. After data cleaning, each question with the respective course name, course code, and score awarded. Predictions are done on this data, and further utilized in the Question Paper Generation Module. The module ensures question papers with various combinations of questions under the corresponding course and difficulty, based on input data.

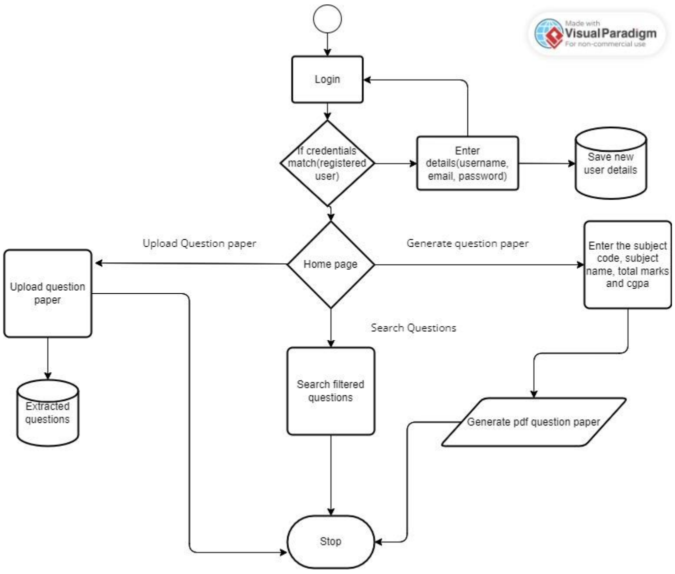

Made with Visual Paradigm For non-commercial use
Login
If credentials match(registered user)
Enter details(username, email, password)
Save new user details
Upload Question paper
Home page
Generate question paper
Enter the subject code, subject name, total marks and cgpa
Upload question paper
Extracted questions
Search Questions
Search filtered questions
Generate pdf question paper
Stop

Fig -2: Application Flow Diagram

4.2 Implementation

Being a popular medium between users and developers, the website is developed to reach the benefits to users. Users can upload PDF question papers and access generated practice set questions here. It is developed in HTML, CSS, and JavaScript, and integrated with the web framework Flask. Fig.2. provides a picturization of different modules and data flow.

Processes involved:

1. Login Process:

- Prompt the user to enter their credentials (e.g., user name and password).

- Validate the credentials against the registered users' database.

- If the credentials match, proceed to the home page. If not, display an error message and prompt for credentials again.

2. Home Page:

- Display options such as uploading a question paper, generating a question paper, and searching for questions.

- Allow the user to choose from these options.

3. Upload Question Paper:

- Provide an interface for the user to upload a PDF question paper file.

- Utilize PyPDF to extract relevant details from the uploaded question paper.

- Store the extracted details in a structured format (CSV file).

4. Extract Questions:

- Process the uploaded question paper using PyPDF to extract questions, course code, course name, mark allocation, and any other relevant information.

- Organize the extracted details into a suitable format for storage and further processing.

5. Generate Question Paper:

- Prompt the user to enter the subject code, subject name, total marks, and CGPA.

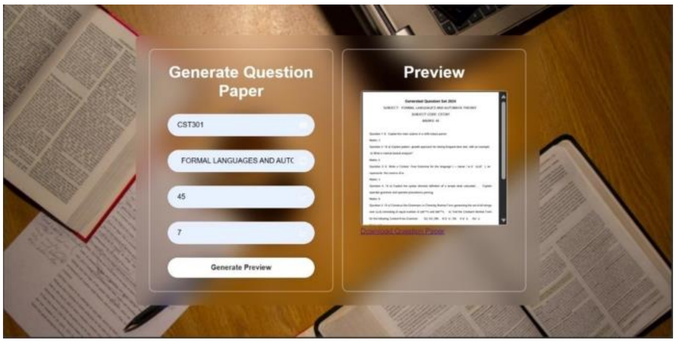
- Utilize the extracted data and user inputs to generate a new PDF question paper.

- Format the generated question paper according to standard conventions and include the necessary details.

6. Search Questions:

- Provide a search interface where users can input keywords or criteria to find specific questions.

- Utilize the stored question data to perform searches and retrieve relevant questions based on the user's input.



CST301

FORMAL LANGUAGES AND AUTC

45

7

Fig -3: Question Paper Generation

5. RESULT

The model has resulted in an accuracy of 76.23 percent, making use of available exam question paper sets since 2019 from KTU. Solely prepared to analyze the question paper set provided by the University, the system can be further enhanced and efficient updating it with incoming question papers. The question paper generated based on the user-provided course code, course name, total marks of the question paper, and attained CGPA is shown in Fig. 3.

6. CONCLUSION

In conclusion, the proposed system attempts to launch an RNN Model to predict the difficulty levels of questions and automate the generation of personalized practice test papers. The system integrates sophisticated models, including RNN and LSTM, to understand the relation between questions and difficulty, resulting in question paper generation and maintaining relevant and updated knowledge. It enhances the generation of question paper sets that are adapted to the student's performance level and learning preferences. Moreover, this system is designed to reduce the workload on educators by automating the question paper generation process and providing a structured, comprehensive question bank.

7. FUTURE SCOPE

The effectiveness of this system is can further enhanced by updating with upcoming question papers, thereby familiarizing with a lot more combinations and making the classification more efficient. The system thus can evolve with time. The app version would make the system more personalized and useful. The future scope encompasses broadening the question bank to encompass more subjects and complex question types and launching a mobile application, which is expected to increase accessibility and user engagement. The continuous integration of feedback and technological advancements will ensure that the system remains at the cutting edge of educational technology, providing significant benefits to all stakeholders involved in the educational process. Privacy and ethics considerations are essential in the future development of these systems. Stricter compliance with data protection regulations and the implementation of privacy safeguards will ensure responsible and ethical handling of user-generated content.

REFERENCES

[1] Chuan Qin, Hengshu Zhu, Dazhong Shen, Ying Sun, Kaichun Yao, Peng Wang, and Hui Xiong. 2023. Automatic Skill-Oriented Ques tion Generation and Recommendation for Intelligent Job Interviews. ACM Trans. Inf. Syst. 42, 1, Article 27 (January 2024), 32 pages. https://doi.org/10.1145/3604552

[2] A. Srivastava, S. Shinde, N. Patel, S. Despande, A. Dalvi and S. Tripathi," Questionator-Automated Question Generation using Deep Learning," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1 5, doi: 10.1109/ic-ETITE47903.2020.212.

[3] B. Pardamean, T. Suparyanto, T. W. Cenggoro, D. Sudigyo and A. Anugrahana, "AI-Based Learning Style Prediction in Online Learning for Primary Education," in IEEE Access, vol. 10, pp. 35725-35735, 2022, doi: 10.1109/ACCESS.2022.3160177.

[4] S. Amin, M. I. Uddin, A. A. Alarood, W. K. Mashwani, A. Alzahrani and A. O. Alzahrani, "Smart E-Learning Framework for Personalized Adaptive Learning and Sequential Path Recommendations Using Reinforcement Learning," in IEEE Access, vol. 11, pp. 89769-89790, 2023, doi: 10.1109/ACCESS.2023.3305584.

[5] Tao Huang, Shengze Hu, Keke Lin, Huali Yang, Hao Zhang, Houbing Song, and Zhihan Lv. 2023. Sequence Generation Model Integrating Domain Ontology for Mathematical question tagging. ACM Trans. Asian Low-Resour. Lang. Inf. Process. Just Accepted (April 2023). https://doi.org/10.1145/3593804

[6] P. M. Patil, R. P. Bhavsar and B. V. Pawar, "A Review on Natural Language Processing based Automatic Question Generation," 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2022, pp. 01-06, doi: 10.1109/ICAISS55157.2022.10010799.

[7] D. L. McGuinness, "Question answering on the semantic Web," in IEEE Intelligent Systems, vol. 19, no. 1, pp. 82-85, Jan.-Feb. 2004, doi: 10.1109/MIS.2004.1265890.

[8] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A Survey on Deep Learning for Named Entity Recognition. IEEE Trans. on Knowl. and Data Eng. 34, 1 (Jan. 2022), 50–70. https://doi.org/10.1109/TKDE.2020.298131