Text-based Q&A: Automated Question Generation and Answering for Enhanced Data Processing

Tonusree Talukder Trina, Abdul Halim Hosain, Kazi Al Refat Pranta, Nusrat Zaman Raya, Ehsanur Rahman Rhythm,Md Sabbir Hossain, and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{tonusree.talukder.trina, abdul.halim.hosain, kazi.al.refat.pranta, nusrta.zaman.raya, ehsanur.rahman.rhythm, md.sabbir.hossain1, }@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—This paper presents a system for generating insightful questions and corresponding solutions from a diverse range of subject-specific chapters. The proposed system uses natural language processing techniques to analyze the text and automatically generate questions related to the content. It is designed to handle texts from different domains and can generate questions and answers of varying word lengths, making it a versatile tool for a wide range of users. To evaluate the system's performance, we conducted experiments on a dataset of chapters from several different subjects, including history, literature, and science. The system achieved high accuracy in generating questions and providing relevant solutions, demonstrating its potential to facilitate the learning process and aid in knowledge discovery. One of the key features of the proposed system is its ability to handle questions of different word limits. Additionally, the system is capable of handling questions of varying lengths, including short and long questions. By leveraging advanced language models and techniques such as summarization and paraphrasing, the system can generate concise answers for short questions and detailed explanations for longer ones. This makes the system suitable for a variety of use cases, from quick reference to in-depth research. Overall, the proposed system has the potential to transform the way people interact with written content and provide a powerful tool for accessing and processing knowledge in a more efficient and effective manner.**

***Index Terms*—Natural Language Processing, Knowledge Discovery, Learning Process, Language Models, High Accuracy**

I. Introduction

E-learning is popular in many nations and allows students to access online resources to learn on their own. A system that generates questions and answers from a text helps students learn more effectively and assess themselves. Students will be better prepared for competitive tests and develop their knowledge of how to respond to questions in exams by practicing more questions. It aids in the preparation of exam questions for the examiners as well. The proposed method has potential uses in research, education, and knowledge management, where it can speed up users' ability to understand complex subjects and locate important information. Also, as information technology advances, a growing amount of information surfaces online, making it challenging for consumers to find what they need and make sense of the vast amounts of data. So, there is a need for a system that can generate questions for us so that we can more easily and quickly extract important information from enormous amounts of data and test our grasp of the subject through assessments.

A question-answering system and a system that generates questions and answers from a text are two distinct research projects. A question-answering system finds the answer from a given text when one is provided, whereas a question-and-answer generation system generates both possible questions and answers from a text. In comparison to Question and Answer creation in Bangla, more research has been done on Question Answering systems. The development of various NLP tools in Bangla has been the subject of numerous studies, but only a very limited number of libraries and tools are currently accessible to the general audience. Only a very small number of studies have been conducted using the established or available toolkits to generate questions and responses in Bangla.

First and foremost, this initiative is being created for the convenience of researchers and students. It is an online exercise for pupils because it will generate questions from texts. This project will be able to produce exam questions from various chapters to aid students in studying for their exams. The project will also be able to increase productivity. By analyzing and replying to large numbers of questions quickly, it will be able to quickly respond to frequently asked questions and free up extra time for more challenging tasks. This project will be adaptable, thirdly. It will be able to generate quizzes depending on any chapter modifications. Because it will be generating questions and answers from text, it will be able to operate in line with the update. Students will also gain time savings from it. It can save time by finding information more quickly than traditional search methods, which is particularly advantageous for students. In the end, it will be able to assess. Initially, this will raise inquiries. These questions will elicit responses from the students, and they will analyze their responses. The answers to those queries will then be produced. By doing so, pupils will be able to evaluate their preparation and spot any shortcomings. Also, by providing the text, researchers will be able to swiftly find the answers to their questions.

II. Literature Review

So far, the Question generation and answering method is not up to the mark to use widely around the world. There are many articles that existed related to question generation and answering. As Bangla is a low-resource language, on the paper [1] the authors had made a data set of their own called BanglaRQA, which contrasts to translated BengaliSQuAD, bn_squad, and human annotated General Knowledge, Factoid QA, BQuAD, had a variety in the types of questions [1] . Where on the paper, [2] GenQA considered one answer generative data set and three answer selection datasets of the English language which are human annotated named WikiQA, MSNLG, WQA, ASNQ [2]. In [1] RQA they used 4 models which are BanglaBERT, BanglaT5, mT5, and mBERT among them BanglaT5 worked best with an EM score of 62.42 and an F1 score of 78.11 [1]. Where in GenQA they used TANDA and 'RoBERTa' as state-of-the-art selectors [2]. As their proposed model was based on T5 of UnifiedQA. In evaluation, BanglaRQA used exact match and F1 score for multi-span answers which are based on the DROP [1]. in GenQA they used BLEU and ROUGE-L but the experiment indicated that these are poor indicators of evaluating answers as the proposed model achieves a higher score on human evaluation [2].

This paper focused on the BART transformer which is based on the Generative model. In this paper, a collection of french journals ranging over 20 years between 1966 to 1986 and a Machine Reading dataset like the Stanford Question Answering Dataset (SQuAD) is used. Lastly, The first outcomes produced with this obtained Machine Reading mode on the self-management dataset. [3].

In another paper, presented the task of SPARQL to Text Question generation which is basically natural language question generation from SPARQL query. There is 5 Question answering corpora and fine-tuning approaches like BART and T5 used. Furthermore, shows the transformer-based approaches and their limitation [4].

The Dual-Channel Reasoning Model (DCRM) is a deep learning approach that aims to improve the performance of complex question-answering systems. It uses two separate neural networks to process the surface and deep channels of a question, respectively. According to the research of Xing Cao et al., DCRM outperformed other state-of-the-art models on two publicly available datasets, SQuAD and TriviaQA. The surface and deep channels are both important for achieving high performance in DCRM, and the dual-channel approach is more effective than using either channel alone. Furthermore, incorporating a knowledge graph into the deep channel further improves performance in DCRM. These findings suggest that DCRM is an effective approach for complex question answering that utilizes both the surface and deep channels of a question to generate an answer [5].

The paper [6] proposed a neural question generation approach that focuses on generating questions based on specific chapters of a given document. The proposed model uses a recurrent neural network (RNN) to encode the input document into a fixed-length vector representation. Then, another RNN is used to generate questions based on the encoded representation of the document. The model is trained using a large corpus of documents and associated questions. The authors evaluated their approach on two large-scale reading comprehension datasets: SQuAD and CNN/Daily Mail. The results showed that their approach outperformed several other state-of-the-art models on both datasets. In particular, the approach achieved an F1 score of 52.1% on SQuAD and a BLEU-4 score of 17.8 on CNN/Daily Mail, which are both among the best-reported results at the time of publication [6].

The paper [7] presents a deep neural network-based system to generate question-answer pairs in the Tamil language for the legal domain. The system uses four preprocessing techniques and the Question and Answer Generator Module employ rules, presumptions, and pseudocode to generate grammatically correct questions and answers. The system achieved high scores in various metrics and was tested by 16 Tamil native speakers, who found 62.22% of the questions produced to be grammatically correct and meaningful. The system can be used for legal research, document summarization, and knowledge management and can promote transparency in the legal system [7].

The paper [8] describes an advanced e-learning system that utilizes natural language processing (NLP) transformers for text summarization, question and distractor generation, and questions answering. The system employs pre-trained transformer models like BERT and GPT-2 to perform the NLP tasks and improve the learning experience for students. The text summarization module generates a summary of a given text using a fine-tuned BERT model. The question and distractor generation module produces multiple-choice questions and answer options based on a given text using a fine-tuned GPT-2 model, and the question-answering module uses a fine-tuned BERT model to answer student questions based on a given text [8].

III. Dataset

Considering our project in the Bangle language, we will be working with multiple data sets. One is the BanglaRQA dataset which has 3000 passages in various fields of knowledge collected from Wikipedia and 14,889 question-and-answer pairs selected by human annotators. It also includes 3631 questions whose answers are not possible to be extracted from the respective passages. The average number of words per passage in this data set is 215 words. The questions answer pair contains questions of four types which are factoid, causal, confirmation, and list. Factoid type consists of answering what, when, who, where, which, etc using facts from the passages, causal type questions includes the question type why or how with descriptive answers, confirmations are question that has only assertive or negative answers and lastly List type includes answers with multiple keywords. Answer collection was again done by human annotators with 4 categories of answers respected to the questions.

We will have data set of Bengali Question Answering Dataset Which is translated from the English data set of SQuAD 2.0. SQuAD 2.0 includes more than 130000 questions with a context of 19029 unique values. It also has a 100 percent valid answer to the respective question ratio.

IV. Methodology

The language transformer models BanglaBert and BanglaT5, as well as mBERT and mT5, were finetuned as part of our study. To begin, we took the dataset and normalized it to meet the specifications of the model. After the first two have been successfully completed, there is a third file that is used for re-testing.

A JSON object with the keys "context," "question," and "response" is present in every file. according to the norms of All files have had their punctuation cleaned up, their white space aligned, and their article breaks eliminated. The files are then tokenized using the appropriate pretuned models. Then, we started with the model's parameters and used AdamW as our optimizer. The 5e-5 learning rate is used here. Ultimately, the learning rate is what modifies the initial model parameters and determines the size of the steps taken in the gradient descent. After that, we use the available datasets to train the models over the course of several iterations. We found that 10 epochs is the sweet spot for this model, avoiding underfitting and overfitting while still producing reliable results. To avoid bias in the training data, we randomly reorder the data and split it into eight batches.

Before starting a new batch, we reset the gradients to zero to prevent them from building up over time. The final step is to execute the model with the input data already split up into its constituent context, question, and response from the JSON dictionary format. The final result shows the expected answers to those questions and the corresponding loss amounts. After that, we utilize the loss to tell the difference between the actual responses and the prediction the model gave us. To do back propagation, the loss is used, and its gradients are compared with the initialization parameters of the model.The AdamW optimizer then decides whether or not to reduce the loss function while making other adjustments to the model's parameters.

Using py torch, we retrieve the results of each pretrained model in the same way, while simultaneously reducing the model sizes. When all the models' weights have been loaded and evaluation mode has been activated, we get a unified result. The dropout and batch normalization layers are then disabled. Predictions based on the weight are calculated, and the remaining layers are taken into account. We plan to use it for ensemble learning afterwards.

V. Evaluation

To evaluate we are currently planning to use an Exact match and for a multi-span answer, we will be using DROP to calculate the F1 score. We are also looking forward to using BLEU and ROUGE-L to determine its natural sound.

VI. Conclusion

The proposed method for selecting relevant questions and answers from selected chapters has the potential to dramatically affect how readers interact with the material delivered to them. The system can handle queries of varied lengths and the lengths of individual words, as shown by experiments conducted on a dataset consisting of chapters from a variety of fields, such as history, literature, and science. This feature distinguishes the system from similar systems in terms of functionality. The system's capacity to formulate accurate queries and provide relevant responses further enhances its ability to speed up learning and aid in information discovery. Adjustments were made to a normalized dataset using the requirements of the transformer-based language models BanglaBert, BanglaT5, mBERT, and mT5 to conduct the research presented here. We randomized the data and trained the models over 10 epochs with a batch size of 8 using the AdamW optimizer and a 5e5 learning rate to reduce the possibility of bias. Using input data in the JSON dictionary format, the model was called to return the desired response to the queries along with their computed loss. The loss was used to distinguish between the observed data and the predicted outcome. It was also used to backpropagate, compute gradients of the loss, evaluate the results in reference to the parameters of the initial model, and ultimately modify the parameters of the starting model. In conclusion, the approach described in this research offers a potent tool for improving the efficacy and efficiency of data collection and analysis. It's a versatile resource that works for everything from short lookups to in-depth studies. In addition to its superior accuracy in question generation and answer delivery, it also handles queries with a wide range of word counts and processing times. Training the transformer-based language models BanglaBert, BanglaT5, mBERT, and mT5 on a normalized dataset with the AdamW optimizer and a total of 10 epochs helped achieve excellent accuracy and show the system's potential.

References

[1] C.-C. Hsu, E. Lind, L. Soldaini, and A. Moschitti, "Answer generation for retrieval-based question answering systems," *arXiv preprint arXiv:2106.00955*, 2021.

[2] S. M. S. Ekram, A. A. Rahman, M. S. Altaf, M. S. Islam, M. M. Rahman, M. M. Rahman, M. A. Hossain, and A. R. M. Kamal, "Banglarqa: A benchmark dataset for under-resourced bangla language reading comprehension-based question answering with diverse question-answer types," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 2518–2532.

[3] F. Bechet, E. Antoine, J. Auguste, and G. Damnati, "Question generation and answering for exploring digital humanities collections," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 4561–4568.

[4] G. Lecorvé, M. Veyret, Q. Brabant, and L. M. R. Barahona, "Sparql-to-text question generation for knowledge-based conversational applications," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, 2022, pp. 131–147.

[5] X. Cao, Y. Liu, B. Hu, and Y. Zhang, "Dual-channel reasoning model for complex question answering," *Complexity*, vol. 2021, pp. 1–13, 2021.

[6] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," *arXiv preprint arXiv:1705.00106*, 2017.

[7] R. Murugathas and U. Thayasivam, “Domain specific question & answer generation in tamil,” in *2022 International Conference on Asian Language Processing (IALP)*. IEEE, 2022, pp. 323–328.

[8] S. Patil, L. Chavan, J. Mukane, D. Vora, and V. Chitre, “State-of-the-art approach to e-learning with cutting edge nlp transformers: Implementing text summarization, question and distractor generation, question answering,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.