Unleashing the Power of Hybrid NLP Models: Reading Comprehension-Based Question Answering through Text-based QA Systems.*

Antara Firoz Parsa, Abdullah Khondoker, S M Ishtiak mahmud, Md Iftekhar Islam Tashik, Enam Ahmed Taufik
Abid Hossain, Md Humaion Kabir Mehedi, and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{antara.firuz.parsa, abdullah.khondoker, sm.ishtiak.mahmud, iftekhar.islam.tashik, enam.ahmed.taufik,
abid.hossain, humaion.kabir.mehedi}@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—The paper titled "Unleashing the Power of Hybrid NLP Models: Advancements in Reading Comprehension-Based Question Answering through Text-based QA Systems" presents novel approaches and datasets in the field of question answering for social science studies and the Bengali language. The study introduces a BART Transformer-based generative model with semantic constraints for question generation, improving question quality in digitized archive collections. Evaluation on various corpora demonstrates the effectiveness of the approach, particularly on the challenging ARCHIVAL dataset. Additionally, the paper introduces BanglaRQA, a reading comprehension-based question-answering dataset for Bengali, along with the evaluation of Transformer models. The results highlight the need for further improvement in performance across different question-answer types. The research also introduces the task of Multi-Question Generation, addressing the limitation of generating only one question per answer. Proposed evaluation metrics and future work considerations emphasize the potential for enriching educational resources through diverse question wordings.**

***Index Terms*—BART, BanglaRQA, Natural Language Processing, Hybrid Model**

I. Introduction

The continuous evolution of Natural Language Processing (NLP) techniques is transformative, especially in the domain of Question Answering (QA). This study embarks on a novel journey to harness the capabilities of hybrid NLP models, focusing particularly on the Bengali language—a realm previously underrepresented in modern QA systems.

Central to the paper's objective is the creation of a tailored dataset, finely curated to cater to specific goals. The dataset, meticulously composed of 6,000 passage-question-and-answer pairs, draws from the educational foundation of Bangla language, encompassing classes six to ten. These pairs, expertly selected by human annotators in consultation with authoritative NCTB textbooks, form the bedrock of a comprehensive QA resource. The dataset is meticulously diversified, accommodating distinct question types.

To process this expansive dataset, we adopted a systematic methodology employing the power of the RoBERTa model and the Hugging Face Transformers library. Through rigorous tokenization, preprocessing, and training, we endeavored to sculpt a QA system optimized for Bengali language comprehension.

In the subsequent sections, we delve deeper into the dataset's nuances, elucidate the methodological approach, present our evaluation outcomes, and wrap up with a comprehensive conclusion, highlighting the implications and future prospects of our findings. Join us in exploring the intricacies of Bengali reading comprehension-based question answering through text-based QA systems.

II. Literature review

In the first paper [1], a model was developed for a question-answering system from Reading Comprehension(RC). The aim of these systems is to provide more accurate answers to users. The author developed a dataset containing 3636 reading comprehension. [1] In their experiments, they use deep neural network architectures for their dataset training such as LSTM (Long Short-Term Memory), Bi-LSTM (Bidirectional LSTM) with attention, RNN (Recurrent Neural Network), ELECTRA, and BERT (Bidirectional Encoder Representations from Transformers) and use transformer-based models for their work. Among all these BERT performs a satisfactory outcome, the testing accuracy is 87.78 percent and the training accuracy is 99 percent. One limitation of this study is the relatively small size of the dataset employed. The limited size of the dataset may have constrained the generalizability of the results and the ability to capture the full variability within the research domain. [1]

The paper [2] addressed overconfidence and over-sensitivity issues in current RC models. Their experiment demonstrated that it improves the robustness of reading comprehension models. The paper develops a method that includes outer knowledge to impose various linguistic constraints, including entity constraint, lexical constraint, and predicate constraint. [2] So the model can produce more accurate predictions for both semantic different and semantic equivalent adversarial

examples. The author also presents posterior regularization into RC models. By applying posterior regularization and implementing linguistic constraints, the method increases the robustness of the base RC models and it can successfully integrate these constraints into the learning process.

This paper [3] aims to explore question-answering in digitized archive collections for Social Science studies. The study introduces a novel approach using a BART Transformer-based generative model with semantic constraints for question generation. [3] Experiments are conducted on three corpora: FQUAD, CALOR-QUEST, and ARCHIVAL. Results demonstrate that incorporating semantic annotations improves question quality. However, evaluating MRQA performance on CALOR-QUEST and ARCHIVAL using models trained on automatically generated questions shows no significant improvements. Nevertheless, the overall approach performs well on the challenging ARCHIVAL dataset. The analysis emphasizes the distinction between expert and crowdsourced questions. The study contributes to designing practical evaluation settings for language understanding systems in archive collections. [3]

In another paper [4], to address the lack of a Bangla Question/Answer Dataset, introduces BanglaRQA, a reading comprehension-based question-answering dataset for Bangla. It contains 3,000 context passages and 14,889 question-answer pairs, covering answerable and unanswerable questions across four question categories and three answer types. The paper [4] also evaluates four Transformer models on BanglaRQA, with the best model achieving 62.42 percent EM, and 78.11 percent F1 scores. However, further analysis reveals variation in performance across question-answer types, indicating room for improvement. The paper [4] demonstrates the effectiveness of BanglaRQA as a training resource by achieving strong results on the bn$_s quad dataset$.

This paper [5] explores the application of NLP techniques, specifically neural language models, for generating question/answer exercises from English texts. The aim is to support ESL teaching to children by generating beginner-level exercises. The proposed approach involves a four-stage pipeline: pre-processing, answer candidate selection, question generation using the T5 transformer-based model, and post-processing. Evaluation on benchmark datasets demonstrates comparable results to previous works. However, limitations are identified, such as imperfect co-reference resolution and errors in question generation. Future work involves refining the system by exploring other language models, expanding the evaluation corpus, and fine-tuning the model for specific English proficiency levels. The developed tool has potential for integration into an educational platform for English language teaching. [5]

Another paper [6] investigates the development of a deep learning-based question answering system in Bengali, aiming to overcome the limitations and lack of progress in this field. By leveraging state-of-the-art transformer models, the research focuses on training a QA system using a synthetic reading comprehension dataset translated from SQuAD 2.0. Furthermore, a human annotated QA dataset sourced from Bengali Wikipedia is utilized for evaluating the models. Comparative analysis with human children provides valuable insights and establishes a benchmark score. The research [6] emphasizes the importance of addressing the challenges in low-resource language settings, particularly in the context of reading comprehension-based question answering.

The paper [7] introduces the task of Multi-Question Generation, aiming to generate diverse questions assessing the same concept. It addresses the limitation of existing systems that generate only one question per answer. The paper proposes an evaluation framework based on desirable question qualities and presents results comparing different question generation approaches. The authors highlight the issue of word overlap between generated questions and input passages and propose metrics to measure question answerability, semantic similarity, and distinct wordings. Future work includes exploring human evaluation metrics, reinforcement learning objectives, and advanced paraphrase systems. The paper suggests incorporating teacher evaluation to define desirable question properties and evaluates the educational impact of diverse question wordings. The publicly-released pipeline holds potential for enriching educational resources at scale. [7]

III. Dataset

In the context of our project's focus on the Bangla language, a pivotal component involves the creation of a dataset tailored to our objectives. Emphasizing the essence of customization, we undertook the task of dataset construction, marking a foundational stride in the development of an effective Bangla question answering system. Our dataset is a result of a meticulous curation process involving 6,000 question-and-answer pairs. These pairs, meticulously selected by human annotators consulting NCTB textbooks from classes six to ten, offer a contextual and informative foundation. Each passage in the dataset, averaging 200 words, provides substantial context for the subsequent question answering. Diversifying our dataset, we incorporated different question types. These encompass an array of queries, from factual inquiries and cause-and-effect exploration to confirmations and inquiries involving multiple keywords. Importantly, human annotators meticulously collected responses for each question type, ensuring the dataset's reliability and relevance. The primary objective driving this endeavor is the development of a proficient Bangla question-answering system. To this end, we painstakingly organized the dataset into training and validation subsets, with each subset elegantly encapsulated within CSV files. These files harmoniously interweave multiple passages with their corresponding questions and expertly annotated answers. By taking the reins in dataset creation, we lay the groundwork for a question-answering system deeply rooted in precision, contextual understanding, and linguistic nuance intrinsic to the Bangla language.

IV. Methodology

The present research methodology entails a systematic and comprehensive comparative analysis of distinct question answering (QA) models, specifically RoBERTa. At First, we normalized our dataset and sought to explain the complex relationships between multiple embeddings.

The AutoTokenizer provided by the Hugging Face Transformers library was employed for tokenization. It tokenizes both the questions and contexts, breaking them down into subword units and mapping them to corresponding token IDs. Additionally, attention masks were generated to indicate which tokens should receive attention during training.

After tokenization, the data was preprocessed to generate inputs that the model could process. These inputs included tokenized sequences, attention masks, and answer span positions. The resulting tokenized inputs were organized into a structured format, which included input IDs, attention masks, start positions, and end positions for answer spans.

The tokenized inputs were then transformed into a list of dictionaries, where each dictionary contained input IDs, attention masks, start positions, and end positions. These dictionaries were structured in a way that facilitated feeding the data into the model during training.

To facilitate question answering in Bangla, the AutoModelForQuestionAnswering model was chosen. This model was loaded from the "saiful9379/Bangla-Roberta-Question-and-Answer" [8] pretrained checkpoint. The tokenizer used for tokenization earlier was also loaded with the same configuration for consistency.

Training hyperparameters, including the learning rate, batch sizes, number of training epochs, and evaluation frequency, were configured. The training strategy was set to perform evaluation steps at regular intervals during training. Additionally, weight decay was applied to control overfitting, and the total number of saved model checkpoints was limited to maintain storage efficiency.

Leveraging the Hugging Face Transformers library, we initialized the Trainer class to encapsulate the training process. In each iteration of the training loop, batches of tokenized question-context pairs were processed, generating predicted answer spans. The loss was computed by comparing predictions with actual answer positions. Gradients were backpropagated to fine-tune the model's parameters, with the aim of minimizing the loss. At intervals, the system evaluated its performance on validation data, tracking metrics like accuracy. Throughout training, we logged vital details including progress, losses, and metrics, offering insights into the learning process. Upon completion, the trained model and results were saved, establishing the foundation for a proficient question-answering system in the Bangla language.

V. Evaluation

We follow a comprehensive methodology to assess our model's performance using the F1 score metric. Our focus is on understanding how key parameters impact its performance. Initially, we investigate the effects of certain parameter values: weight decay is set at 0.05, evaluation steps are configured to 1000, the learning rate is set to 2e-5, and the epoch is set to 10. Under these conditions, the model achieves an F1 score of 0.54.

Continuing our analysis, we explore different parameter settings. We vary the weight decay to 0.03, increase evaluation steps to 2000, and set the learning rate to 3e-5. This results in an improved F1 score of 0.65.

In our quest for the optimal configuration, we further experiment with parameters. We test weight decay at 0.01, set evaluation steps to 5000, choose a learning rate of 4e-5, and extend the epoch to 100. This yields the highest F1 score of 0.7.

Through iterative adjustments of these influential parameters, we identify the best configuration that maximizes the F1 score. This process provides valuable insights into the model's adaptability and accuracy when dealing with questions in the context of the Bangla language.

VI. Conclusion

In "Unleashing the Power of Hybrid NLP Models: Reading Comprehension-Based Question Answering through Text-based QA Systems," we delve into the vast potential of hybrid NLP models for reading comprehension-based question answering. With a concentrated effort on the Bengali language, the study introduced BanglaRQA, a robust reading comprehension-based dataset containing an expansive range of question-answer pairs. The paper also innovated in question generation using a BART Transformer-based generative model with semantic constraints, showing marked effectiveness especially on the ARCHIVAL dataset. Leveraging tools such as the Hugging Face Transformers library facilitated the intricate process of tokenization and model training. Through iterative evaluation, the optimal model configuration achieved an F1 score of 0.7, shedding light on the model's adaptability to the Bangla language context. Additionally, a novel avenue for future work involves the task of categorizing the question-answer dataset. By classifying the questions and answers into distinct categories based on their content, question type, or contextual focus, the system could offer more targeted and refined responses. This categorization could provide valuable insights for educational purposes, content organization, and further model enhancement.

References

[1] T. T. Aurpa, R. K. Rifat, M. S. Ahmed, M. M. Anwar, and A. B. M. S. Ali, "Reading comprehension based question answering system in bangla language with transformer-based learning," *Heliyon*, vol. 8, no. 10, p. e11052, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844022023404

[2] M. Zhou, M. Huang, and X. Zhu, "Robust reading comprehension with linguistic constraints via posterior regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2500–2510, 2020.

[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*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 4561–4568. [Online]. Available: https://aclanthology.org/2022.lrec-1.486

[4] 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*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2518–2532. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.186

[5] G. Berger, T. Rischewski, L. Chiruzzo, and A. Rosá, "Generation of english question answer exercises from texts using transformers based models." IEEE, 2022. [Online]. Available: https://hdl.handle.net/20.500.12008/37155

[6] T. T. Mayeesha, A. M. Sarwar, and R. M. Rahman, "Deep learning based question answering system in bengali," *Journal of Information and Telecommunication*, vol. 5, no. 2, pp. 145–178, 2021. [Online]. Available: https://doi.org/10.1080/24751839.2020.1833136

[7] M. Rathod, T. Tu, and K. Stasaski, "Educational multi-question generation for reading comprehension," in *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. Seattle, Washington: Association for Computational Linguistics, Jul. 2022, pp. 216–223. [Online]. Available: https://aclanthology.org/2022.bea-1.26

[8] M. S. Islam, "Transformer based bangla$_r obert_q a$,$''$ 2023.$[Online].Available : https : //github.com/saiful9379/Bangla_Roberta_Question_and_Answer$