Bangla Question Answering System Using BanglaBERT Language Model and AdaFactor Optimizer

Sanjana Sabah Khan , Mehreen Khan , Md Humaion Kabir Mehedi ,
Md Mustakin Alam , and Annajiat Alim Rasel

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
Email: {sanjana.sabah.khan, mehreen.khan, humaion.kabir.mehedi,
md.mustakin.alam}@g.bracu.ac.bd, and annajiat@gmail.com

***Abstract*—In recent years, there has been uncountable research on Question Answering (QA) systems in different languages. Still, QA is a critical NLP problem and a long-standing artificial intelligence milestone which is yet to reach perfection. Choosing the right language model and training it on a standardized dataset is the key to perfecting a QA system. Bidirectional Encoding Representations for Transformers (BERT) models are known to perform competently on complex information extraction tasks. There are many BERT based models listed in Hugging Face for numerous tasks such as Fill-Mask, Question-Answering, Summarization, Sentiment Analysis, etc. One of these fine models is BanglaBERT which is a based on Natural Language Understanding (NLU) model that is pretrained in Bangla, the fifth most-spoken native language and the seventh most spoken language by total number of speakers in the world. In this paper, we have fine-tuned the BanglaBERT model using AdaFactor optimizer. To achieve this goal, we have collected two datasets from from two different published sources and sliced each of them into 3 splits - training, validation and test sets. We have compared our obtained Exact Match (EM), F1 and accuracy scores with other BERT based models - mBERT, RoBERTa, DistilBERT and IndicBERT - to emphasize on their performance.**

***Index Terms*—Natural Processing Language, Question Answering system, BERT, BanglaBERT, AdaFactor**

I. Introduction

Question Answering has been a very crucial part of any language study. The term Question Answering (QA) comes from the reading comprehension (RC) where the reader is provide with a paragraphs and some related questions to answer. For this task, many Natural Processing Language (NLP) models are trained for RC question answering using machine learning algorithms. These models can interpret the textual context and answer the correlated questions using our natural language. In addition, they also try to interpret the sentiments of questions correctly and search for the distinct text, sentence or words. With the Internet making vast quantities of text available at our fingertips, QA systems have become critical for us to better understand and quickly extract key information from text.

As each section in the dataset does not have a predetermined number of sentences and so the responses might range from one to many words, this makes it quite challenging to create a training set for a supervised learning model. Computers need to construct context in order to grasp the meaning of ambiguous words in text. This problem was solved in 2018 when Google launched a language model - Bi-directional Encoder Representation from Transformer (BERT) - which can be used for Natural Language Processing (NLP) tasks [1]. BERT is a machince learning model based on transformer that is trained for a single task, yet it can be modified and reused for other tasks as well. There are many BERT based models listed in Hugginface for numerous tasks such as Fill-Mask, Question-Answering, Summarization, Sentiment Analysis, etc. One of these excellent models is BanglaBERT [2], which is a BERT-based Natural Language Understanding (NLU) model that was pretrained in Bangla, the fifth most widespread native tongue in the world and the seventh most spoken language by the majority of people.

In Bangladesh, 98% of people speak Bangla as their first language, making it the official and national language of the country. In addition, Bangla is the second official language of Jharkhand and the official language of West Bengal, Tripura, and the Barak Valley part of the state of Assam in India. It is the sixth most widely used Indo-European language in the world. Designing a QA system for Bangla is quite challenging as Bangla is a very different language from English in terms of alphabet, pronunciation, grammar and vocabulary. With this pager, we have contributed the folowing:

- We have fine-tuned the BanglaBERT model using AdaFactor optimizer.
- We have collected two datasets [2], [3] and after thorough analyzation we have spit each of these datasets in training, validation and test sets in the 80:10:10 ratio.
- We have compared our obtained Exact Match, F1 and

978-1-6654-6316-4/22/$31.00 © 2022 IEEE

accuracy scores with other BERT based models - mBERT, RoBERTa, DistilBERT and IndicBERT.

II. Related Works

For the past few years, there has been substantial research on English QA systems, but only a handful for Bangla. In 2021, Arnab Saha, Mirza I. Noor, Shahriar Fahim and three other authors from Rajshahi University of Engineering and Technology (RUET), designed BERT-Bangla which a language model pre-trained on Bangla unlabelled text [4]. They tested BERT-Bangla on different Bangla NLP classification tasks and attained better results than any other existing Bangla language models. They also developed the Bangla Question Answering Dataset (BQuAD) from scratch. Then, they fine-tuned the pre-trained language model on BQuAD for QA tasks and achieved an EM accuracy of 45% and F1 accuracy of 78.5%. Later in 2022, Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad and five other researchers from Bangladesh University of Engineering and Technology (BUET), proposed two pre-trained models in Bangla - BanglaBERT and BanglishBERT - in a proceeding paper [2]. They created Bangla Natural Language Inference (NLI) and QA dataset.They also introduced the Bangla Language Understanding Benchmark (BLUB). BanglaBERT performs better than mBERT and XLM-R (base) in the supervised setting by 6.8 and 4.3 BLUB scores, whereas BanglishBERT performs better than them in zeroshot cross-lingual transfer by 15.8 and 10.8 exactly.

There has been more intensive implementation and benchmarking on several variants of BERT language model. In a published article [5], Tahsin Mayeesha Sarwar, Md Abdullah and M. Rahman described their work in which they have used the multilingual BERT models (mBERT) for zero shot transfer learning. Thy have fine-tuned mBERT model on their synthetic training for Bangla reading comprehension. They have also bench-marked other versions of BERT model like RoBERTa [6] and DistilBERT [7] on both zero shot and fine-tuned model setting. In another paper [5], Tahara Aurpa, Khandakar Rifat, Shoaib Ahmed and 2 other collaborators talks about their predictive algorithm, which can determine the response to any text and reading comprehension question. They used the most recent NLP method, known as transformer-based learning BERT, which makes use of the self-attention mechanism as well as the pre-training language model for prediction. The suggested technique has been used by the authors in a real-world benchmark dataset that is brand-new to the Bangla language. This dataset has the potential to be an excellent addition to Bangla NLP. Their model produced an acceptable outcome with testing accuracy of 87.78% and training accuracy of 99%, while ELECTRA offers training and testing accuracy of 82.5% and 93%, respectively.

In addition to these studies, there are also some notable works for Arabic QA system as well which contributed to the enhancement of QA models. In 2014, Heba Abdelnasser, Maha Ragab, Reham Mohamed and four other researchers from Alexandria University,Egypt, proposed a new Arabic QA system - Al-Bayan, designed for the Holy Qur'an [8]. This system can achieve upto 85% accuracy using the top-3 results. They have built a Semantic Information Retrieval module that converts pieces of natural language text into a weighted vector of Quranic concepts, using machine learning techniques. The Qur'anic Reading Comprehension Dataset (QRCD) and a shared assignment of answering questions about the Holy Qur'an were both assigned at the OSACT 2022 workshop. One solution to this shared task, a proceeding paper was published on 3 June 2022 which suggested certain post-processing actions to improve the accuracy of the responses in line with the official measure [9]. On the official test set, the authors got a Partial Reciprocal Rank (pRR) score of 56.6 percent after fine-tuning a number of BERT models suited for Arabic language.

TABLE I
Number of question-passage pairs in each split.

Dataset	Train	Valid	Test
Dataset-1	3,040	380	380
Dataset-2	5,990	750	750

III. Dataset Details

A. Dataset-1

For our research, we have collected two datasets from open-sources. The first collected dataset is a reading comprehension based question answering dataset in Bangla Language, developed by the research students at Jahangirnagar University on December, 2022 [3] for developing automatic reading comprehension system. They initially assembled a lengthy real-world passage in Bengali that was taken from various articles, books, biographies, etc., before creating questions based on the texts. Their dataset consists of 3,800 strings of passage-question-answer triplets.

B. Dataset-2

The second collected dataset was published by the research students at North South University (NSU) on November, 2020 [10].They have translated a large subset of SQuAD 2.0 [11] from English to Bengla. SQuAD [12] is a large-scale reading comprehension dataset collected in English with annotations from crowd workers. It consists of 100,000 question-passage pairs that are coupled with their extractive answers. SQuAD 2.0 incorporates the SQuAD data with additional 50,000 adversarial questions which are absolutely unanswerable. The adversarial questions are written in such a way that they look exactly like answerable questions. From here, we have taken 7,488 strings of passage-question-answer triplets.

For our project, we have analyzed these two datasets and split them each into training, validation and test sets in the 80:10:10 ratio, shown in Table I.

IV. QA System Framework

After pre-training the BERT architecture on Bangla, transfer learning is used to fine-tune the model for QA tasks in the QA system. The following subsections discuss the QA system's framework.

A. BERT

The design of the BERT model is based on a multilayered transformer encoder with bidirectional self-attention. BERT is trained on a huge corpus of unlabeled text using two unsupervised tasks: next sentence prediction (NSP) and the masked language model (MLM) [1]. By paying attention to both the left and right contexts, the model learns to predict the token that has been randomly selected to be hidden during the MLM method. A model is needed to determine the relationship between comparative sentences for NLP tasks like answering questions and interpreting ordinary language. The binarized next sentence prediction task, one of the unsupervised tasks in BERT, helps a language model understand the link between two phrases since masked language modeling does not account for the relationship between sentences. Using the embeddings or weights from a pre-trained model on labeled data allows for fine-tuning for subsequent NLP tasks including named entity recognition (NER), sentence classification, and quality assurance (QA).

The internal process of BERT are as explained: A special token called CLS is added at the beginning of a sequence, and the token's final state is used for classification tasks. A second special token, SEP, is used to separate two sequences that are part of a pair. Another unique MASK token replaces the masked words. The sum of corresponding input embeddings, segment embeddings, and position embeddings of the same dimension is fed into the first transformer block as the input representation. Pre-training a BERT model requires a significant amount of text data, making it computationally expensive, whereas fine-tuning a model using labeled data is relatively inexpensive. On eleven English NLP tasks, Google-trained BERT models produced cutting-edge results.

B. Supervised Fine-tuning using BanglaBERT model

The BERT classifier is trained with the answers, the attention mask, and the input sequence after they have been created. The BERT tokenizers are also used to tokenize the response. The collected datasets are used to fine-tune the models for a subsequent reading comprehension task using pre-trained BanglaBERT transformers [2]. The attention mask, input ids, and token type ids is entered into the model's three input layers. Finally, the activation softmax function was utilized by the output layer.

The SEP token separates the question and reference text, which are packed together as a single sequence. For the question and the reference text, B segment embeddings are utilized by BERT. Start and end token classifiers receive embeddings from the BERT model's final transformer encoder. Each token's embedding is weighted independently by the start token classifier and the end token classifier. The span of the response in the text is returned for the tokens with the highest likelihood of being the start and end tokens. The probabilities that that token will serve as both the start and end token are as follows:

$$P_{is} = \frac{e^{s.t_i}}{\sum_j e^{s.t_j}} \tag{1}$$

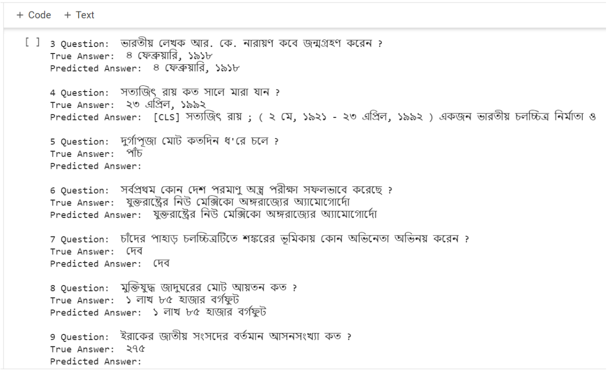
$$P_{ie} = \frac{e^{e.t_i}}{\sum_j e^{e.t_j}} \tag{2}$$

V. Experimental Evaluation

A. Proposed Supervised Fine-tuning

For fine-tuning the BanglaBERT model, we have used the AdaFactor optimizer which internally adjusts the learning rate depending on the `weight_decay`, `scale_parameter`, `relative_step` and `warmup_init` options (see Table II). AdaFactor uses `optim.zero_grad()` which zeros all the gradients of the variable and it updates the learnable weights of the model. We can also say that it will set the gradients of all the optimized torch tensors to zero. This gives the optimizer in the training loop additional flexibility in how the gradient is calculated and applied. When the model or input data is large and just one training batch can fit on the GPU card, this is essential. Our system evaluates the automated learning rates using two different `max sequence length` values of 382 and 512 for Dataset-1 and Dataset-2 respectively with right padding and truncation. The algorithm encodes the texts with stride set to 128. We have run our model on 10 epochs for Dataset-1 and 15 epochs for Dataset-2 while the batch size for both is set to 16 for training and validation sets. Following the training of our classifiers, we used unlabeled test datasets and generated some findings based on projected responses. For evaluation, we have determined the F1 score, Exact Match (EM) and accuracy measures.

B. Results



```
[ ] 3 Question:  ভারতীয় লেখক আর. কে. নারায়ণ কবে জন্মগ্রহণ করেন ?
    True Answer:  ৪ ফেব্রুয়ারি, ১৯১৮
    Predicted Answer:  ৪ ফেব্রুয়ারি, ১৯১৮

    4 Question:  সত্যজিৎ রায় কত সালে মারা যান ?
    True Answer:  ২৩ এপ্রিল, ১৯৯২
    Predicted Answer:  [CLS] সত্যজিৎ রায় ; ( ২ মে, ১৯২১ - ২৩ এপ্রিল, ১৯৯২ ) একজন ভারতীয় চলচ্চিত্র নির্মাতা ও

    5 Question:  দুর্গাপূজা মোট কতদিন ধ'রে চলে ?
    True Answer:  পাঁচ
    Predicted Answer:

    6 Question:  সর্বপ্রথম কোন দেশ পরমাণু অস্ত্র পরীক্ষা সফলভাবে করেছে ?
    True Answer:  যুক্তরাষ্ট্রের নিউ মেক্সিকো অঙ্গরাজ্যের অ্যামোগোর্দো
    Predicted Answer:  যুক্তরাষ্ট্রের নিউ মেক্সিকো অঙ্গরাজ্যের অ্যামোগোর্দো

    7 Question:  চাঁদের পাহাড় চলচ্চিত্রটিতে শঙ্করের ভূমিকায় কোন অভিনেতা অভিনয় করেন ?
    True Answer:  দেব
    Predicted Answer:  দেব

    8 Question:  মুক্তিযুদ্ধ জাদুঘরের মোট আয়তন কত ?
    True Answer:  ১ লাখ ৮৫ হাজার বর্গফুট
    Predicted Answer:  ১ লাখ ৮৫ হাজার বর্গফুট

    9 Question:  ইরাকের জাতীয় সংসদের বর্তমান আসনসংখ্যা কত ?
    True Answer:  ২৭৫
    Predicted Answer:
```

Fig. 1. Sample output generated using Test dataset.

For the predicted answers, our proposed system calculate the True Positives (TP), False Positives (FP), and False Negatives (FN). True Positives are the number of tokens that are shared by the predicted answers and the ground truth. Misleading Up-sides show the anticipated tokens that are not in the test information's result, and Bogus Negatives are the tokens from ground truth's responses that the classifier can't foresee. WIth

TABLE II
PARAMETERS AND THEIR RESPECTIVE VALUES FOR ADAFACTOR OPTIMIZER.

Parameter	Value
`weight_decay`	0.1
`scale_parameter`	True
`relative_step`	True
`warmup_init`	True
`max_sequence_length` (dataset-1)	382
`max_sequence_length` (dataset-2)	512

TP, FP, and FN, we use Equation (3) and (4) to determine the Precision and Recall,

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recal = \frac{TP}{TP + FN} \tag{4}$$

Next, the determined precision and recall are used for calculating the F1 Score using the Equation (5). The F1 score measures the degree of token overlap between the top predicted answer and correct answer. The system provides 58.64% F1 score for the dataset-1 and 58.14% F1 score for dataset-2.

$$F1score = \frac{2 * precision * recall}{precision + recall} \tag{5}$$

The system also determines Exact Match (EM) score using Equation (6) which is a binary measure that will be equal to one when the top predicted answer exactly matches the correct answer. Our fine-tuned model provides 58.60% EM score for dataset-1 and 57.24% EM score for dataset-2.

$$EMscore = \frac{\sum_{i=1}^{N} F(x_i)}{N} \tag{6}$$

$$whereF(x_i) = \begin{cases} 1, & \text{if predicted answer is correct} \\ 0, & \text{otherwise} \end{cases}$$

For calculating accuracy, the predicted output and the test output are compared. BERT, with a batch size of 16, an automated learning rate and a maximum sequence length of 382, produces 63% accuracy for dataset-1 and 62% accuracy for dataset-2. In Table III, we have compared our F1 and EM scores with mBERT, RoBERTa and DistilBERT models [2] and accuracy with IndicBERT [10]. In comparison to these model, we can say that our proposed fine-tuning method for BanglaBERT model has worked well, if not best.

VI. CONCLUSION

The main objective of our work is to develop an effective QA system for Bangla reading comprehension that has the potential to contribute in the education system following Bangla curriculum. We used the recently discovered BERT-based model to achieve our goal. We have fine-tuned the BanglaBERT model, proposed by Abhik Bhattacharjee et al [2]. To prove the method's viability, we used two significant evaluation matrix. For Dataset-1 [3], our proposed fine-tuning provides testing accuracy of 63% and training accuracy of 69%, and for Dataset-2 [10], it provides testing accuracy of 62% and training accuracy of 66%. It is clear that our testing accuracy is lower than our training accuracy and is not as great as other exiting models' accuracy. However, we intend to include more samples in our experiments to improve our testing accuracy.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT PRE-TRAINED MODELS FOR QUESTION ANSWERING SYSTEM.

Models	Language	EM	F1	Accuracy
mBERT	Multiling.	47.18	48.52	-
RoBERTa	Multiling.	54.41	54.41	-
DistilBERT	English	50.05	51.18	-
IndicBERT	Multiling.	-	-	50.84
BanglaBERT (dataset-1)	Bangla	58.60	58.64	63
BanglaBERT (dataset-2)	Bangla	57.24	58.14	62

In future, we plan to take this research further ahead to create an effective real-world Bangla Reading Comprehension. We intend to develop an algorithmic solution that will improve test data accuracy. We plan to employ this approach for tasks based on reading comprehension like , match the following, fill in the blank, true/false and multiple-choice. In addition, we want to work more on our programming skills to improve our testing accuracy. Lastly, we hope that a valuable contribution to Bangla NLP may have been made by this proposed fine-tuning.

ACKNOWLEDGMENTS

We are very thankful to our senior faculty Annajiat Alim Rasel for his support, advice and guidance in every step of our work. We would also like to thank our research assistant Humaion Kabir Mehedi for his constant supervision and advice. Lastly, we are immensely grateful to BRAC University for providing us with this opportunity and its adequate resources. We are privileged to be a part of this institute.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, jun 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[2] A. Bhattacharjee, T. Hasan, W. Ahmad, K. S. Mubasshir, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, jul 2022, pp. 1318–1327. [Online]. Available: https://aclanthology.org/2022.findings-naacl.98

[3] T. T. Aurpa, R. K. Rifat, M. S. Ahmed, M. M. Anwar, and A. B. M. S. Ali, "Uddipok: Reading comprehension based question answering dataset in bangla language," Dec 2022. [Online]. Available: https://data.mendeley.com/datasets/s9pb3h2cjy

[4] A. Saha, M. I. Noor, S. Fahim, S. Sarker, F. Badal, and S. Das, "An approach to extractive bangla question answering based on bert-bangla and bquad," in *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, 2021, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9528178

[5] 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

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[7] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019. [Online]. Available: http://arxiv.org/abs/1910.01108

[8] H. Abdelnasser, M. Ragab, R. Mohamed, A. Mohamed, B. Farouk, N. El-Makky, and M. Torki, "Al-bayan: An Arabic question answering system for the holy quran," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 57–64. [Online]. Available: https://aclanthology.org/W14-3607

[9] M. Elkomy and A. M. Sarhan, "TCE at qur'an QA 2022: Arabic language question answering over holy qur'an using a post-processed ensemble of BERT-based models," in *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*. Marseille, France: European Language Resources Association, jun 2022, pp. 154–161. [Online]. Available: https://aclanthology.org/2022.osact-1.19

[10] 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

[11] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," *CoRR*, vol. abs/1806.03822, 2018. [Online]. Available: http://arxiv.org/abs/1806.03822

[12] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100, 000+ questions for machine comprehension of text," *CoRR*, vol. abs/1606.05250, 2016. [Online]. Available: http://arxiv.org/abs/1606.05250