Bangla Question Anwering System Using BanglaBERT Language Model.

Sanjana S. Khan
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
sanjana.sabah.khan@g.bracu.ac.bd

Mehreen Khan
Computer Science
Brac University
Dhaka, Bangladesh
mehreen.khan@g.bracu.ac.bd

Annajiat Alim Rasel
Senior Lecturer
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
annajiat@bracu.ac.bd

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. For our work, we have collected two dataset and sliced each of them into 3 splits - training, validation and testing sets. Mainly, we have fine-tuned the BanglaBERT model using the collected datasets and compared our obtained accuracy score with other BERT based models - mBERT, RoBERTa, DistilBERT and IndicBERT.

Keywords : Natural Processing Language, Question Answering system, BERT, BanglaBERT

1 Introduction

Question Answering has been a very crucial part of any language study. The term Question Answering (QA) here comes from the reading comprehensive where the reader is given certain paragraphs to read and answer some questions related to those paragraphs. Using this concept, there are many models that are trained to do so using machine learning algorithms. They can understand the text context and answer the question related to that context. They can answer using the natural language that we are used to. It not only answers but also tries to correctly understand the sentiments of questions and look for the particular text or sentence it is looking for.

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. However, creating a training set for a Supervised Learning Model is difficult since each portion does not have a predetermined amount of sentences and answers can range from one word to many words. Computers need to understand the meaning of ambiguous language in text by using surrounding text to establish context. 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 (Devlin et al., 2019). BERT is a transfer learning model which means it is trained for one task but can be reused for another task. IT has a trained Transformer Encoder stack, with twelve in the Base version, and twenty-four in the Large version. There is many BERT provided by Hugginface for multiple tasks like, Summarization, Fill-Mask, Question-Answering etc.

Bangla is the official and national language of Bangladesh, with 98% of Bangladeshis using Bangla as their first language. Within India, Bangla is the official language of the states of West Bengal, Tripura and the Barak Valley region of the state of Assam. It is also a second official language of the Indian state of Jharkhand since September 2011. is the fifth most-spoken native language and the seventh most spoken language by total number of speakers in the world. Bangla is the fifth most spoken Indo-European language. 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 (Aurpa et al., 2022b; Bhattacharjee et al., 2022) and after thorough analyzation we have spit each of these datasets in training, validation and testing sets in the 80:10:10 ratio.
- We have compared our obtained accuracy score with other BERT based models - mBERT, RoBERTa, DistilBERT and IndicBERT.

2 Related Work

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 (Saha et al., 2021). 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 (Bhattacharjee et al., 2022). They created Bangla Natural Language Inference (NLI) and QA dataset.They also introduced the Bangla Language Understanding Benchmark (BLUB). In the supervised setting, BanglaBERT outperforms mBERT and XLM-R (base) by 6.8 and 4.3 BLUB scores, while in zero-shot cross-lingual transfer, BanglishBERT outperforms them by 15.8 and 10.8 precisely.

There has been more intensive implementation and bench-marking on several variants of BERT language model. In a published article (Aurpa et al., 2022a), Tahsin Mayeesha Sarwar, Md Abdullah and M. Rahman described their work in which they have used the multilingual BERT model for zero shot transfer learning and fine-tuned it on their synthetic training dataset on the task of Bengali reading comprehension. They have also benchmarked other variants of BERT model ike RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019) on both zero shot and fine-tuned model setting. In another paper (Aurpa et al., 2022a), Tahara Aurpa, Khandakar Rifat, Shoaib Ahmed and 2 other collaborators talks about their model which predicts the answer for any passage and reading comprehension question. They have utilized the latest NLP technique, the transformer-based learning BERT, which uses the self-attention mechanism and the pre-training language model for prediction. The authors have applied the proposed methodology in a real-world benchmark dataset entirely new to Bangla Language. This dataset can be a noble contribution to Bangla NLP. Their model gave a satisfactory result with 87.78% of testing accuracy and 99% training accuracy, and ELECTRA provides 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 (Abdelnasser et al., 2014). This system can achieve upto 85% accuracy using the top-3 results. They used machine learning techniques to build a Semantic Information Retrieval module that maps fragments of natural language text into a weighted vector of Quranic concepts. In OSACT 2022 workshop, a shared task of Question Answering on the Holy Qur'an was given along with The Qur'anic Reading Comprehension Dataset (QRCD). One solution to this shared task, a proceeding paper was published on 3 June 2022 which proposed some post-processing operations to enhance the quality of answers aligning with the official measure (Elkomy and Sarhan, 2022). The authors fine-tuned a variety of BERT models optimized for the Arabic language and achieved a Partial Reciprocal Rank (pRR) score of 56.6% on the official test set.

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

Table 1: Number of question-passage pairs in each split.

3 Dataset

3.1 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 (Aurpa et al., 2022b) for developing automatic reading comprehension system. They have first comprised a real-time and long passage (context) in the Bengali language which is collected from different Bangla articles, novels, biography, etc, and then generated questions based on the given passages. Their dataset consists of 3,800 strings of passage-question-answer triplets.

3.2 Dataset-2

The second collected dataset was published by the research students at North South University (NSU) on November, 2020 (Mayeesha et al., 2021).They have translated a large subset of SQuAD 2.0 (Rajpurkar et al., 2018) from English to Bengla. SQuAD (Rajpurkar et al., 2016) 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 1.

4 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.

4.1 BERT

A multilayer transformer encoder with bidirectional self-attention serves as the foundation for the architecture of the BERT model. Using two unsupervised tasks, BERT is trained on a large corpus of unlabeled text: next sentence prediction (NSP) and the masked language model (MLM). The MLM procedure involves randomly masking a portion of the input text's tokens, and the model learns to anticipate the masked token by paying attention to both the left and right contexts. NLP errands such as question responding to and regular language surmising require a model to figure out the connection between comparative sentences. One of the unsupervised tasks in BERT, the binarized next sentence prediction task, assists a language model in comprehending the relationship between two sentences because masked language modeling does not capture the relationship between sentences. For subsequent NLP tasks like named entity recognition (NER), sentence classification, and quality assurance (QA), fine-tuning involves using the embeddings or weights from a pre-trained model on labeled data.

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.

4.2 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. Pretraining tasks MLM and NSP help the proposed BERT model learn about the input sequence, and the generator distinguishes between original and replaced tokens. Our dataset is used to fine-tune the models for a

subsequent reading comprehension task using pre-trained transformers. 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:

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

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

5 Experimental Evaluation

5.1 Proposed Supervised Fine-tuning

After our classifiers were trained, we applied unlabeled test dataset and produced some results based on predicted answers. For evaluation, we determine the F1 score, Exact Match (EM), accuracy and loss.

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 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 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.

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

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

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

Table 2: Parameters and their respective values for AdaFactor optimizer

For calculating accuracy, the predicted output and the test output are compared. For the optimizer, 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 2). 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 provides more freedom on how the gradient is accumulated and applied by the optimizer in the training loop. This is crucial when the model or input data is big and one actual training batch do not fit in to the GPU card. Our system tests automated learning rates with the batch size - 16 and two distinct `max_sequence_length` values 382 and 512 . 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.

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

Table 3: : Performance comparison of different pre-trained models for Question Answering system.

In Table 3, we have compared our F1 and EM scores with mBERT, RoBERTa and DistilBERT models (Bhattacharjee et al., 2022) and accuracy with IndicBERT (Mayeesha et al., 2021).

6 Conclusion

The primary objective of our work is to develop an effective QA system for Bangla reading comprehension that has the potential to play an important role in the Bangla education system. We used the most recent NLP technique, transformer-based learning BERT, to put the model into action. We have fine-tuned the BanglaBERT model proposed by Abhik Bhattacharjee (Bhattacharjee et al., 2022). To prove the method's viability, we used two significant evaluation matrix. For Dataset-1 (Aurpa et al., 2022b), our proposed fine-tuning provides testing accuracy of 63% and training accuracy of 69%, and for Dataset-2 (Mayeesha et al., 2021), 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.

In the future, we intend to put this research into action as an embedded system to create a real-world Bangla Reading Comprehension system that is more effective. We intend to develop an algorithmic solution that will improve test data accuracy. We want to use this approach for other questions about reading comprehension like true/false, fill in the blank, and multiple-choice. In addition, we intend to expand our dataset in order to provide a useful data source for Bangla Reading Comprehension. A valuable contribution to Bangla NLP may be 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

Heba Abdelnasser, Maha Ragab, Reham Mohamed, Alaa Mohamed, Bassant Farouk, Nagwa El-Makky, and Marwan Torki. 2014. Al-bayan: An Arabic question answering system for the holy quran. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 57–64, Doha, Qatar. Association for Computational Linguistics.

Tanjim Taharat Aurpa, Richita Khandakar Rifat, Md Shoaib Ahmed, Md Musfique Anwar, and A. B. M. Shawkat Ali. 2022a. Reading comprehension based question answering system in bangla language with transformer-based learning. *Heliyon*, 8(10):e11052.

Tanjim Taharat Aurpa, Richita Khandakar Rifat, Md Shoaib Ahmed, Md Musfique Anwar, and A. B. M. Shawkat Ali. 2022b. Uddipok: Reading comprehension based question answering dataset in bangla language.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohamemd Elkomy and Amany M. Sarhan. 2022. 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, pages 154–161, Marseille, France. European Language Resources Association.

Tasmiah Tahsin Mayeesha, Abdullah Md Sarwar, and Rashedur M. Rahman. 2021. Deep learning based question answering system in bengali. *Journal of Information and Telecommunication*, 5(2):145–178.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Arnab Saha, Mirza Ifat Noor, Shahriar Fahim, Subrata Sarker, Faisal Badal, and Sajal Das. 2021. 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)*, pages 1–6.