A Pattern Recognition Approach for Bangla Question Answering System

1st Rubayet Mahjabin
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
rubayet.mahjabin@g.bracu.ac.bd

2nd Mubtasim Fuad Mozumder
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
mubtasim.fuad.mozumder@g.bracu.ac.bd

3rd Humayra Musarrat
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
humayra.musarrat@g.bracu.ac.bd

4thJamilatun Subarna
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
jamilatun.subarna@g.bracu.ac.bd

5th Mehnaz Ara Fazal
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
mehnaz.ara.fazal@g.bracu.ac.bd

6th Md Sabbir Hossain
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
md.sabbir.hossain1@g.bracu.ac.bd

7th Annajiat Alim Rasel
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
annajiat@gmail.com

***Abstract*—This study intends to improve Bangla QA answering approaches by the help of pattern recognition techniques which include modern Transformer models. The research evaluates models such as BERT and LSTM in identifying question patterns recognition, accuracy measurement while foreseeing the computational efficiency of the research. This research intends to address linguistic challenges, typo and polysemy and grammatical mistakes by leveraging a Bangla QA dataset. We intend to highlight the importance of Transformer Architecture like BERT and GPT-3 in Bangla QA systems. While the methodology involves data collection from a Bangla QA dataset,also preprocess the dataset and analyze it. Our Evaluation metrics include token length histograms and performance metrics specific to QA tasks. Our result aims to contribute to the use of Transformer models in Bangla QA while offering architectural improvements for enhanced understanding of the language. Our study expects some improvements in pattern recognition and language processing that are specific to the language of Bangla.**

I. Introduction

A. Background and Motivation

In the field of Bangla Question answering, we can see a massive growth and progress when we talk about Natural Language Processing (NLP). Creating autonomous devices which can understand human speech and make a response to the understanding which will be similar to humans, this is the goal of NLP. Processing and working with Bangla is more difficult as it has more complexity than other languages we see. We can see complicated morphology in this language and there are no capitalizations needed when we are using Bangla. We need robust solutions when it comes to using Bangla in digital communication as it is growing rapidly.

The underlying intentions and subtle context of Bangla queries can be better understood with the aid of pattern recognition techniques. When we utilize pattern recognition, the machine/ system depends on the ability to identify patterns in the Questions and passages, after that it has to determine relevant contextual information which are needed to give a proper answer to that specific question. Nevertheless, pattern recognition in this language is really difficult due to the complex morphology inflection. Finding patterns in Bangla Questions and then generating answers require sophisticated techniques to adapt and translate to the features we intend to

use. Thus, we need to merge pattern recognition and NLP together to improve Bangla QA systems and get over the linguistic challenges that we are facing with speaking Bangla.

B. Statement of the Problem

When to talk about generating Bangla answers, we find ourselves in a difficult task of finding patterns and intents in natural language questions. Some significant obstacles are typos, polysemy, complex nested named items, and insufficiently big labeled datasets for Bangla QA. On the other hand, novel transformer-based neural models, like as BERT and GPT-3, have proven to be very adept in multilingual pattern recognition and contextual understanding. Regardless of distance, their self-attention architecture is capable of modeling dependencies well. Therefore, improving such models on Bangla might lead to significant advancements.

C. Objectives of the Research

The purpose of the research is to determine how well-suited modern Transformer models are for identifying question patterns in Bangla. More specifically, a combination of descriptive questions will be used to assess contextual language models such as, multilingual BERT (mBERT), and others. Pattern accuracy, intent recognition F1 scores, answerability prediction, computational efficiency, and model uncertainty are some of the metrics that will be measured. In order to progress the State-of-the-Art, these evaluations will highlight the strengths and shortcomings. We will investigate further Bangla corpus synthesis, and self-supervised techniques to achieve optimal fine-tuning. Ultimately, these results will direct architectural improvements to enhance the responding of Bangla questions. Improved pattern recognition will lead to more insightful context modeling, showing possible paths toward natural language understanding that is comparable to that of humans.

II. Literature Review

The goal of question answering (QA) systems is to provide accurate answers to questions made by people in natural language. Earlier attempts to parse and extract replies depended on manually created rules and semantic grammars. However, the flexibility and scalability of these conventional methods to new domains was severely constrained. Additionally, they were unable to grasp the nuances and unpredictability of difficult questions in language. Neural network-based, data-driven statistical methods gained momentum as a result. Early breakthroughs that showed promise for natural language interpretation were made using convolutional neural networks (CNNs) and recurrent neural networks (RNNs), such as LSTMs and GRUs [1]. Nevertheless, in translation tasks, Vaswani et al. (2017)'s revolutionary Transformer architecture outperformed CNNs and RNNs [2]. This model completely discarded recurrence and replaced it with a multi-headed self-attention mechanism to find global links between words in sentences. The Transformer was able to pick up new knowledge far faster and maintain more context thanks to its parallelization ability. Presented as the state-of-the-art approach, transformer networks are widely employed in natural language processing.

Finding the right response structure for textual questions requires QA systems to recognize semantic patterns in the questions, which is a basic difficulty [3]. Syntactic patterns are frequently extracted using methods like dependency parsing and part-of-speech (POS) tagging [4]. But the inherent complexity of human language, such as lexical variations, ambiguous word meanings, sophisticated reasoning, common sense integration, etc., is difficult to overcome by merely applying templated principles [5]. Therefore, there is great potential for this pattern recognition problem using contemporary data-driven Transformer models.

Significant progress in language modeling for downstream NLP tasks has been prompted by the work of Vaswani et al. (2017). To pretrain deep bidirectional representations by simultaneously conditioning on both left and right context in all layers, BERT (Bidirectional Encoder Representations from Transformers) was explicitly suggested in 2018 [6]. This allowed for far deeper contextual mastery, in contrast to earlier unidirectional approaches. With very little pretrained parameter adjustments on target datasets, BERT reached state-of-the-art performance on question answering and other tasks.

Recently, autoregressive models for generation tasks—like GPT-3, or Generative Pre-training Transformer 3—have also been developed [7]. Through the use of almost 175 billion parameters, GPT-3 showed unmatched few-shot generalization capabilities. Additionally, encoder-decoder architectures like T5 are gaining popularity more lately by recasting all NLP issues into a text-to-text format [8]. More natural language competence is now possible as a result of constant Transformer advancements.

III. Methodology

A. Data Collection

We are using the dataset by Haque et al. (2020) for our research on Bangla question answering. This dataset is available in the github under the name "Bangla-Dataset-for-Question-Answer-System". The information in this dataset was gathered by picking quotes from widely read novels, poetry, and public papers on a variety of topics that might be found online. Each row in the Excel file "alldata.xlsx," which contains the dataset, comprises a question and answer pair. We have converted the file in CSV. In the CSV file, the questions are listed in the first column, the matching answers are listed in the second, the name of the content file from where the questions and answers are derived is listed in the third, and the fourth, fifth, sixth, seventh column indicates the sentence segment containing the right response. We have named the first column QUESTION, second column ANSWER, third column content_file, and fourth, fifth, sixth, seventh column with sentence segment , sentence segment 2, sentence segment 3 and sentence segment 4 for easier retrieval of data while coding. The paragraph folder contains all the text files of different paragraphs from where the questions and answers are derived.

B. Preprocessing and Exploratory Data Analysis

In order to facilitate data manipulation and analysis, the Bangla QA dataset is first imported into a Pandas DataFrame from the CSV file. The textual questions and responses are tokenized into word tokens using the NLTK library in the subsequent natural language processing phases. Token length histograms are produced and basic statistics are performed in order to obtain an understanding of the linguistic properties of the dataset. This quantitative summary of the dataset's structure is shown by the histograms, which show that most questions have between five and twelve words (Fig. 1) whereas answers typically peak at one to three words (Fig 2).

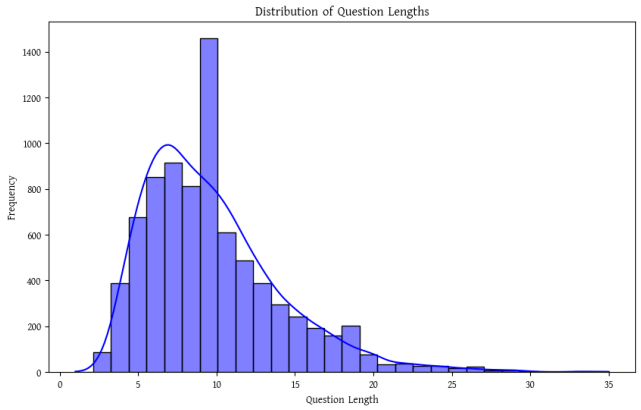

Fig. 1. Distribution of Question Lengths

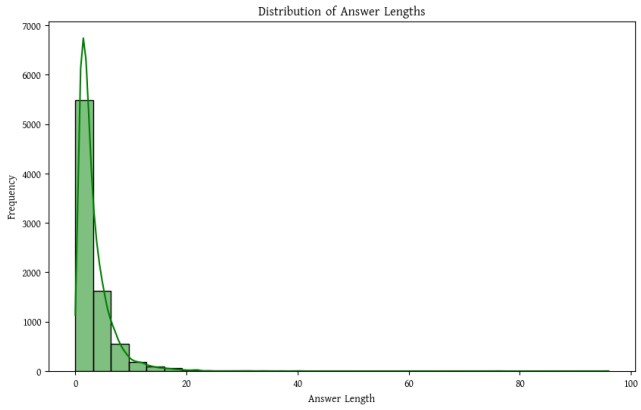

Fig. 2. Distribution of Answer Lengths

Understanding the content relevance between the question and answer lexicons depends heavily on visual analysis (Fig. 3). The full corpus of question and response pairings is used to extract an aggregate collection of all unique terms. The Top 20 most often occurring words throughout questions and replies are revealed through the computation of frequency distributions. The resultant long-tailed bar chart shows the frequency of terms suggesting that a considerable amount of fact-finding queries were interrogative and concerned events, dates, and locations. With regard to the language styles, variety, and topic patterns found in the Bangla dataset, this thorough analysis works as a diagnostic tool and offers insightful information.

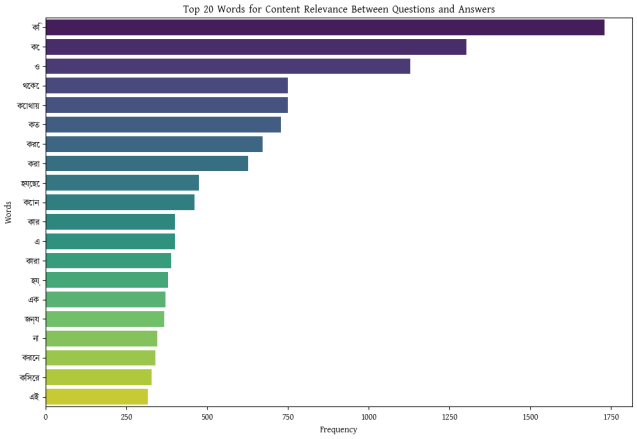

Fig. 3. Top 20 Words for Content Relevance Between Questions and Answers

C. Model Description

1. Bangla BERT: Bidirectional Encoder Representations from Transformers

The Bangla BERT deep bidirectional transformer model has shown state-of-the-art performance in a number of natural language processing applications. It uses the transformer architecture's self-attention mechanism to simulate the left and right contexts for each input character. Bangla BERT is able to fully understand the syntactic and semantic patterns seen in Bangla text because of this. Adding task-specific output layers to the pretrained Bangla BERT can yield accurate results in question pattern classification and extract replies. Because Bangla BERT can represent each query word while conditioning on both left and right context, it provides a very helpful bidirectional feature for answering questions. To represent the diversity and complexity of the Bangla language in real world, we might get help from the bidirectional conditioning.As Bangla BERT has bidirectional representation and contextual modelling skills which we can use in the question answering system, so that it may improve our system more.

2. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTMs) can help solve NLP sequential modeling problems more precisely. Using a gated memory cell structure, we can effectively extract long-term associations from textual input. In particular, we know that LSTM models can sequentially perform tasks such as reviewing questions and paragraph sequences, as well as gathering

contextual and structural inputs for Bangla queries. Through the creation of a connection between current inputs and long-term contexts, LSTM allows us to match questions with relevant texts. Encoder-decoder models can be trained on long short-term memory (LSTM) to help with end-to-end mapping tasks, such as spam responses. Their excellent sequence modeling skills are the reason for this. Their parameter economy may also make them easier to train than intricate pretrained models. All things considered, LSTM-based techniques offer a distinct sequence modeling viewpoint and may be used in conjunction with other pattern recognition techniques.

3. BERT + LSTM - ENSEMBLE

We can solve Bangla questions by combining characteristics of BERT and LSTM models. In this case, LSTM layers that come after can be given the contextual representations that the pretrained BERT levels have learnt. This enables the LSTM decoder to do targeted sequence modeling and dependency learning on the BERT representations, utilizing the rich bidirectional semantic patterns that the BERT encoder provides. While the LSTM may link these representations locally throughout the sequence, the BERT-recognized patterns can supply the global context. By jointly optimizing the stacked BERT and LSTM models on Bangla question-answer pairs, the hybrid model can identify patterns in language from both more general and more specific settings. In comparison to separate BERT or LSTM models, this combined modeling is probably going to perform better on the Bangla question answering task. Additionally, the hybrid technique may overcome the poor sequence modeling capabilities of BERT and the absence of bidirectional context in LSTM. All things considered, the BERT + LSTM design maximizes contextual knowledge by utilizing the dual pattern recognition capabilities.

VI. CHALLENGES AND FUTURE DIRECTIONS

A. CHALLENGES ENCOUNTERED IN THE STUDY

Challenges such as data limitations can be seen through insufficient labeled Bangla QA datasets. The inadequacy of labeled datasets in the Bangla language hinders the training and evaluation of question-answering models. Unlike some widely spoken languages, Bangla lacks comprehensive datasets that capture the breadth and depth of its linguistic diversity. The scarcity of labeled data restricts the ability to train models effectively, impacting their understanding of the nuanced patterns and structures inherent in Bangla questions. Moreover, Challenges in diverse question types and linguistic nuances were also seen. Bangla, as a language, presents a rich tapestry of linguistic nuances and varied question types. From formal inquiries to colloquial expressions, the spectrum of question types is vast. Additionally, the language exhibits intricate linguistic nuances, including complex morphological structures and contextual dependencies. Adapting models to recognize and respond accurately to this diversity demands not only a large dataset but also sophisticated algorithms capable of discerning the subtleties within the language. Another challenge is the generalization issue which is addressing the gap in model performance across varied domains. The challenge extends beyond the confines of data scarcity to encompass the ability of models to generalize effectively across diverse domains. Bangla QA systems must not only comprehend general knowledge questions but also exhibit domain-specific expertise, spanning topics from literature to science. Bridging this gap in performance requires innovative approaches that go beyond traditional training methods, ensuring that models can adapt to the nuances of different subject matters. Real-world scenarios also introduce unpredictability, requiring models to handle questions that may deviate from standard linguistic structures. Achieving adaptability involves exposing models to a variety of contexts, including informal language use, regional dialects, and evolving linguistic trends. Furthermore, the models should be robust enough to tackle questions stemming from dynamic, real-world situations, such as current events or evolving cultural phenomena.

B. RECOMMENDATIONS FOR FUTURE RESEARCH

We can overcome data scarcity in pattern recognition by collaborating for the creation of larger and more diverse Bangla QA datasets. The foundation of any effective machine learning model lies in the quality and quantity of the data it is trained on. Collaborative efforts are crucial to address the scarcity of labeled datasets in Bangla Question Answering. By fostering partnerships between research institutions, industry experts, and linguistic communities, we can pool resources and expertise to create larger, more diverse datasets. These datasets should encompass a wide array of domains, question types, and linguistic styles, ensuring that the models gain exposure to the richness and variability of the Bangla language. At the same time, we can also implement active learning techniques to iteratively improve model understanding. Traditional machine learning models rely heavily on pre-existing labeled data, often struggling when faced with new or complex scenarios. Active learning provides a solution by allowing the model to interactively query humans or other information sources to obtain new, informative data points. By implementing active learning in the Bangla QA context, we empower our models to actively seek clarification on ambiguous or challenging instances. This iterative learning process refines the model's understanding, making it more adept at handling diverse linguistic patterns and nuances.

The exploration of novel transformer architectures through the investigation of custom architectures tailored for Bangla language intricacies can be a path for future research. The uniqueness of the Bangla language calls for tailored solutions in the form of custom Transformer architectures. Traditional Transformer models, while powerful, may not fully capture the intricacies of Bangla grammar, syntax, and semantics. By conducting in-depth investigations into the linguistic characteristics of Bangla, researchers can design architectures specifically crafted to handle its nuances. This involves incorporating features such as enhanced attention mechanisms, domain-specific embeddings, and fine-tuned positional encoding to

ensure that the model effectively captures the essence of the Bangla language.

REFERENCES

[1] Young, T., Hazarika, D., Poria, S., Cambria, E. (2018). Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 13(3), 55-75.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[3] Li, X., Roth, D. (2002). Learning question classifiers. In COLING-02: The 19th International Conference on Computational Linguistics.

[4] Hacioglu, K., Ward, W. (2002, May). Question classification with support vector machines and error correcting codes. In Proceedings of the Human Language Technology Conference (pp. 28-30).

[5] Kumar, V., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., ... Socher, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. In International conference on machine learning (pp. 1378-1387). PMLR.

[6] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (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 (Vol. 1, pp. 4171-4186).

[7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[8] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1-67.