

**FINAL YEAR REPORT ON ‘STIBGK’: A COMPARATIVE STUDY BETWEEN  
BANGLA-ENGLISH LANGUAGES ON QUESTION-ANSWER USING LSTM  
BASED ATTENTION MECHANISM**

**BY**

**Nusrat Nabi  
181-15-10524**

**Sakib Ahmad Siddiquee  
181-15-10776**

**Sumiya Islam  
181-15-10661**

**Syed Rakib Al Hossain  
181-15-11027**

This Report Presented in Partial Fulfillment of the Requirements for the  
Degree of Bachelor of Science in Computer Science and Engineering

Supervised by

**Md. Sazzadur Ahamed**

Senior Lecturer

Department of CSE

Daffodil International University

Co-Supervised by

**Md Zahid Hasan**

Assistant Professor

Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY  
DHAKA, BANGLADESH  
JANUARY 2022**

## **APPROVAL**

This Project titled “STIBGK”: A Comparative Study between Bangla-English Languages on Question-Answer Using LSTM Based Attention Mechanism”, submitted by Nusrat Nabi, ID No: 181-15-10524, Sumiya Islam, ID No: 181-15-10661, Sakib Ahmad Siddiquee, ID: 181-15-10776 and Syed Rakib Al Hossain, ID: 181-15-11027 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 5<sup>th</sup> December 2021.

## **BOARD OF EXAMINERS**



**Chairman**

---

**Dr. S.M Aminul Haque**  
**Associate Professor and Associate Head**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



**Internal Examiner**

---

**Naznin Sultana**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



**Internal Examiner**

---

**Raja Tariqul Hasan Tusher**

**Senior Lecturer**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



**External Examiner**

---

**Dr. Dewan Md. Farid**

**Professor**

Department of Computer Science and Engineering  
United International University

## DECLARATION

We hereby declare that, this project has been done by us under the supervision **Md. Sazzadur Ahamed, Senior Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree.

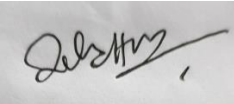
### Supervised by:



---

**Md. Sazzadur Ahamed**  
Senior Lecturer  
Department of CSE  
Daffodil International University

### Co-Supervised by:



---

**Md. Zahid Hasan**  
Assistant Professor  
Department of CSE  
Daffodil International University

### Submitted by:



---

**Nusrat Nabi**  
ID: 181-15-10524  
Department of CSE  
Daffodil International University

*Sumiya Islam*

---

**Sumiya Islam**

ID: 181-15-10661

Department of CSE

Daffodil International University

*Ahmad*

---

**Sakib Ahmad Siddiquee**

ID: 181-15-10776

Department of CSE

Daffodil International University

*Rakib*

---

**Syed Rakib Al Hossain**

ID: 181-15-11027

Department of CSE

Daffodil International University

## ACKNOWLEDGEMENT

First of all, we express our gratitude to the Most Merciful God for enabling us to successfully complete the project phase of the year.

We are very grateful and thanks our honorable Supervisor **Md. Sazzadur Ahamed** for getting this work done properly as he has helped us to make our work a success with his proper advice. His guidance and support gave us with the confidence level increase to complete this research project accurately and correctly. He was the first to inspire us to do this with the Bengali language. He is doing extensive work on Bengali language and served us all work related resources and topical knowledge to complete this research for the Bengali language. We thank our honorable Co-Supervisor **Md Zahid Hasan** for supporting us in completing our work. We are grateful to our honorable department head **Professor Dr. Touhid Bhuiyan** for helping us to do research on this Bengali language. Moreover, I would like to thank the members of other faculties and the staff of our department for their support.

Lastly, we would like to express our gratitude and appreciation to our family and friends who have given their full support to make this research a success.

## **ABSTRACT**

Recently, breakthroughs of NLP research have improved a range of activities, most notably the Question Answering System for many languages. Since the last few years, question answering (QA) systems have grown at a breakneck pace. With the continuous development of the network, the question- and-answer method has become a way for people to get information quickly & precisely that the user will ask and with the increase in web sourcing, any information has become available to the people as the relevant data is stored in that source. LSTM has been introduced, a focus-based deep learning model for the Q&A method in this study. It matches one of the sentences in the question and answer and solves the problem of unexpected features. Using the attention mechanism in the system provides accurate answers by focusing on the specific questions of the candidate. Furthermore, we have proposed an adequate knowledge addition-based framework for the Q&A method. This memory contains a nested word or character level encoder that handles problems outside the words in the dataset or some rare words. We compare both Bangla and English-based question- answer for the dataset domain based on International GK, Bangladesh GK, and Science & Technology. A Sequence to Sequence LSTM based question-and-answer system with a total number of 10,000 data has been proposed through an attention mechanism with (99.91 and 99.48) % accuracy for Bangla and English data, respectively. Overall, LSTM works perfectly for both Bengali and English and is the best Q&A model.

## TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	i-ii
Declaration	iii-iv
Acknowledgements	v
Abstract	vi
List of Figures	x
List of Tables	xi
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-4</b>
1.1 Introduction	1
1.2 Motivation	1-2
1.3 Objective	2
1.4 Research Questions	2
1.5 Expected Output	3
1.6 Report Overview	3-4
<b>CHAPTER 2: BACKGROUND STUDIES</b>	<b>5-9</b>
2.1 Introduction	5
2.2 Related Work	5-6



2.2.1 Related work for Bangla QA	6-7
2.2.2 Related work for English QA	7-8
2.3 Research Summary	8
2.4 Scope of the problem	8-9
2.5 Challenges	9
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>10-25</b>
3.1 Introduction	10-11
3.2 Research Title and Apparatus	11-12
3.3 Data collection	12-15
3.4 Data preprocessing	15-16
3.4.1 Split text	16
3.4.2 Tokenization	16-17
3.4.3 Pad Sequences	17
3.5 Statistical Analysis	17-18
3.6 Model descriptions	18
3.6.1 LSTM	18-20
3.6.2 Sequence to Sequence	20-23
3.6.3 Attention Mechanism	23-24
3.7 Representation of Taxonomy for our model	24-25

<b>CHAPTER 4: RESULTS AND DISCUSSION</b>	<b>26-33</b>
4.1 Introduction	26-27
4.2 A comparison of QA models for Bangla and English	27-31
4.3 Prediction Accuracy In real life	32-33
<b>CHAPTER 5: IMPACT ON SOCIETY, ETHICAL ASPECTS AND SUSTAINABILITY</b>	<b>34-35</b>
5.1 Impact on Society	34
5.2 Ethical Aspects	35
5.3 Sustainability Plan	35
<b>CHAPTER 6: CONCLUSION AND FUTURE WORK</b>	<b>36-39</b>
6.1 Summary of the Study	36-37
6.2 Conclusion	37
6.3 Recommendations	38
6.4 Implication for Further Study	38-39
<b>REFERENCES</b>	<b>40</b>

## **LIST OF FIGURES**

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 3.1.1: Work process for Question Answering System	11
Figure 3.4.1: Dataset preprocessing	16
Figure 3.6.1.1: Architecture of LSTM model	20
Figure 3.6.2.1: Architecture of Seq2seq mode	23
Figure 3.6.3.1 Architecture of Attention Mechanism	24
Figure 3.7.1: Architecture of Taxonomy	25
Figure 4.2.1: Graphical representation of loss and accuracy for English	28
Figure 4.2.2: Graphical representation of Validation loss and validation accuracy for English	29
Figure 4.2.3: Graphical representation of loss and accuracy for Bangla	30
Figure 4.2.4: Graphical representation of Validation loss and Validation accuracy for English	31

## **LIST OF TABLES**

<b>TABLES</b>	<b>PAGE NO</b>
Table 3.3.1: Sample data of our Bangla dataset	13-14
Table 3.3.2: Sample data of our English dataset	14-15
Table 4.3.1: Sample prediction 1: Bangla Question answering system	32
Table 4.3.2: Sample prediction 2: Bangla Question answering system	33
Table 4.3.3: Sample prediction 1: English Question answering system	33
Table 4.3.4: Sample prediction 2: English Question answering system	33



# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

One of the core challenges in the field of Natural Language Processing work is question answering systems depending on context. The proposed framework is the finest alternative for reducing time consumption and obtaining the proper answer to any given inquiry. As a result, the system's purpose of this research is to get the entire document of the supplied query answer. We named our research work as "STIBGK": Science, Technology, International and Bangladeshi based General Knowledge comparative study between Bangla-English languages on Question-Answer Using LSTM based Attention Mechanism. Our title explains our domains and methods also to help others to understand our work hints from the title. During this whole study, we looked at both English and Bangla context-based question-answering systems. This project is a closed domain system for both languages, with domains based on International, National (Bangladesh), and Science-based general knowledge questions and answers. Using sequence to sequence LSTM based on attention processes, our system will compare the two languages. The LSTM (long short-term memory) is incredibly beneficial for digesting vast amounts of text and remembering past knowledge in problems of succession.

### 1.2 Motivation

There is much research regarding context-based question answering systems. But in Bangla Language, there are few works where we can compare our language-related work with English. Because if we look at the amount of quality work, the English language has made very good progress in this field. Also, we see that the Bangla language is spoken by 241 million native speakers but the amount of this comparison work is low. In today's society, data is one of the most valuable commodities. Every day, a great variety of question-related

records with answers derived from exclusive sources are added to the database. We find it quite hard to recall the questions and their replies.

So, in this day and age, we require a mechanism for automatically responding to questions. So our work will help to find a proper answer based on the context of the given question and also it will compare with the English language. Shortly, our work will help to improve the current systems that are based on question answering based on context data.

### **1.3 Objective**

The basic goal of a question answering system is to obtain an exact, rapid, and intelligible answer. But with the help of context data, it will do this work more accurately. Context data helps machines to understand what the question is about and which answer is more accurate. The purpose of a question-and-answer system is to deliver an answer to a query. These are used to respond to the user's inquiries. They're also employed in education, such as in classes and lectures. Question answering systems are also being studied for their ability to be transferred over national and international borders. That is why our objective was to include the English language with Bangla to compare its result.

### **1.4 Research Questions**

In your whole research, we have focused on these questions below to find answers. Our report contains the answers to these questions with explanations and we also described the theory behind our answers.

- What are Question Answering techniques?
- How to preprocess Bangla and English data with context?
- Which techniques will be best for preprocessing data?
- What are the differences between Bangla and English QA systems?
- What are the benefits of Bangla and English QA?
- How to get better accuracy using Bangla and English QA datasets?
- How can we implement a QA system using Bangla and English language?

## **1.5 Expected Output**

We have gathered a huge amount of data manually for Bangla and English languages. Then we preprocessed it and made it ready for our research. We are expecting better accuracy for both languages from which we can make sure our used techniques and model is on the right way. Also, we are expecting to develop a user-friendly system. Using two important languages for comparison and using these to develop a QA system was very challenging. Though we have focused as much as we can to get better results for both languages.

Bangla and English context-based question answering systems need to be compared to know the future goal and scope of work needed to be done. We are hoping our proposed framework will show better results as we have used a large dataset and better techniques. Automatic question answering systems help many systems to enrich their usability and user experience. We are hoping that our systems can be implemented where Bangla and English languages are used for communication. Lastly, our developed system can use both Bangla and English language users which will double its usage for further work.

## **1.6 Report Overview**

In our report, there are a total of six chapters where we have discussed and explained all our work in as organized a way as possible. Below, we are giving a very short overview for understanding it better.

### **Chapter 1**

We have covered all our introduction, motivation of work, our objectives, and research related questions in this chapter. We have discussed the theory and work-related information before and after our research.



## **Chapter 2**

In this section, we have discussed related work for Bangla and English QA systems. Also, we covered the research summary and scope of our research problem. At the end of this chapter, we covered challenges we have faced during this research.

## **Chapter 3**

In this chapter, we explained the methodology of our work. Also covered the required technologies and equipment we have used. We have shown some sample data and the source for English and Bangla QA both. Lastly, we covered data pre-processing, statistical analysis, model description, and representation of the Taxonomy of our model.

## **Chapter 4**

We have discussed results and shown the comparison between Bangla and English data graphs. We have shown real-life prediction accuracy using tables for both of the languages we have worked on.

## **Chapter 5**

We covered chapter five by the impact on society, ethical aspects, and sustainability plan to understand the importance of our research work.

## **Chapter 6**

Finally, we covered and discussed future work. We have explained the summary of the study, recommendation, conclusion, and implication for further study.

## **CHAPTER 2**

### **BACKGROUND STUDIES**

#### **2.1 Introduction**

Question answering technique is software that can extract data and respond to questions in the same way that a person can. In recent years, much work has focused on computer vision to automatically answer questions presented in plain language on any domain or subject. It is divided into two categories: open domain-based and close to the primary. Closed domains interact with specialized topics, while open domains engage with inquiries concerning everything based on world knowledge. Due to the numerous varied ways human language expresses the same information demand, question answering is difficult. Question answering is just a process in which the respondent asks a question & receives an automatic response. As a consequence, minor differences in semantically comparable questions can lead to different outcomes. Deep learning is one of the finest opportunities to expand an automatic system for automatic question answering. Humans can save time and energy by using computerized question responses, which saves money. Nearly every day, humans post a number of queries on various websites in the hopes of receiving responses. In a short amount of time, a large number of people can locate or search for an answer to a query. However, preserving and finding the answers to those questions is indeed a time-consuming process. Our model will strive to provide an automated response to your questions in an efficient manner. Our model contains context-based queries while we took a closed-domain approach. We analyze question-answers in Bangla and English for the system, which is based on International GK, Bangladesh GK, and Science & Technology.

#### **2.2 Related Work**

This section is about the description of QA system-related tasks where we highlighted the models, outcomes, how imported datasets have been used to build QA systems, and

datasets, etc. Question classifiers, automatic question responses, question features, open domain question answers, question taxonomies, factoid question answers, and so on are all related works. Nowadays question answering systems-related tasks are leading research fields.

### **2.2.1 Related work for Bangla QA**

The study of Mumenuunnessa et al. [1], has established a seq2seq model of deep learning which is a context-based Question Answering system of the Bangla Language. The authors took 2000 Bangla data for their research. The model of this research has a 99 percent accuracy for this dataset. In terms of answer prediction, the trained model performs well. In another study, [2] researchers compared the resemblance of the questions and context for their QA system and also used cosine distance for embedding and calculating the text resemblance. The model computed the coefficient of similarity and it was 0.41. In this paper [3], the writers have created state-of-the-art work in order to develop a quality assurance system. Used a dataset of smaller sentient QA from Bangladeshi Wiki to test their models by them. Lastly, a survey was done to establish their outstanding point by comparing the proposed models. The study of Sourav et al. [4], has described their model of closed domain QA system of Bangla language. With and without mentioning the object name, their suggested technique for retrieving responses from several resources derive the answer. Over five coarse-grained categories, provide 75.3 percent and 90.6 % accuracy for the question and document. In another study [5], completed some of the preliminary work on the Bengali language QA system. The initial step of the QA system, question detection, was completed using Logistic Regressionism, K-Neighbor classifier, LSTM Multilayer Perceptron. With the linear kernel trick, they got the best performance of SVM. In this paper [6], authors have created a factoid QA system for the Bangla language. They talked about the difficulties in developing a Bengali question-answering relevant sentence extraction as well as ranking in the paper. A technique for extracting ranked answers from relevant sentences is also proposed. The study of Md. et al [7], used the sequence to sequence architecture to create a QA system that is automated. An attention mechanism is

used for decoder-level encoders (a bi-directional LSTM) is also used. The model successfully answers the question and reduces 0.003 of training loss.

### **2.2.2 Related work for English QA**

The study of Chengfei Li et al. [8], provides a hybrid approach for the intelligent question-answering challenge involving contextual queries. The model uses a bidirectional LSTM network and a convolutional neural network for the QA system and performs well according to the experiments on Babi data. Many models have been used, among them, BiLSTM gives 0.99% accuracy. In another study [9], compared four distinct architectural designs under the same conditions, and make a deeper understanding of the influence of architectural design choices. The subject inference architecture is built around a typical LSTM model that has been trained to recognize the span of the topic in the inquiry. In this paper [10], the author presents a detailed overview of many models for the QA task. New deep neural networks (DNN) and traditional information retrieval perspectives are encompassed. Also, provide well-informed datasets for the task as well as a present outcome from the review so that other researchers can compare different alternative strategies. The study of Kai Lei et al. [11], suggested a deep learning model with symbol level KBQA process for single fact QA. To achieve good accuracy in item presence identification, the BiLSTM-CRF method was utilized to reduce unnecessary information. The study of Linlong Xiao et al. [12], introduces Attender which is a reading comprehension program of QA systems. The BiLSTM system is provided, which is built on the attention-based & allows for the normal linear conversion of self-attention value computation. The model gets 71.31% accuracy and the F1 score is 79.7. In this paper [13], the authors present a QA model of non-factoid which is an unsupervised matching strategy that works with numerous context-specific embedding representations that method the content at multiple levels of analysis. Though it's a simple model, it performs well. The accuracy of the Challenge partition of ARC is 34.01%, AI2 Reasoning Challenge dataset is 64.6% top-three tasks on the WikiQA dataset accuracy is 74.0. In another study [20], the authors began by conducting a literature assessment of the recent state-of-the-art and

essential methods in question answering systems. A simple Question Answering interface style with passage attention visualization is built. As part of their continuous effort, they also suggest a strategy for progressing the existing state of the art.

### **2.3 Research Summary**

In our thesis, we constructed an automatic system for both Bangla and English question answering systems, as well as compared the accuracy of both. 10,000 datasets are utilized to train and test our Question answering automated system for this purpose. For each language, we collected 5000 datasets. This study is a closed domain system for both languages where the domains are Bangladesh GK (general knowledge), International Gk, Science, and Technology based. We used many forms of Wikipedia, books, and websites to get data. In the process of collecting data we created three columns where the first column contains questions, the second one contains context, and the third one is answered. Before applying the deep learning model the preprocessing of datasets is very essential. We encode the data, and the encoder changes it from one ordination to the next. First, the encoder splits the input into text, then tokenizes it before removing stop words. Following preprocessing, we employed pad sequences to verify that each of the orders or series has an equal list range. After that, we used our model on the datasets. As attention is used with a fixed-length internal representation encoder-decoder architecture, in this study we employed an attention-based LSTM method. The model delivered the best result on the created dataset after training and testing.

### **2.4 Scope of the problem**

Our system will compare the two languages using sequence to sequence LSTM based on attention mechanisms. The LSTM (The long short-term memory) is most relevant for processing large text to remember historical information in sequence problems. The sequence-to-sequence (seq2seq) paradigm was used for our research. The seq2seq system is based on an encoder-decoder structure, in which the encoder analyzes the input data and compresses the data into fixed-length context words. During translation, the seq2seq

paradigm performs well for short sequences of context. A modern development for Bangla NLP studies is a question answering system. The seq2seq paradigm was utilized in this study to create a Bangla and English Question Answering Machine based entirely on International GK, Bangladesh GK, and Science & Technology questions. In this project, brief questions are good outcomes for question responding. So, this collection of rules also provides a response to a short-term inquiry. The seq2seq model is often unable to correctly handle long entrance sequences, therefore the context vector for the decoder is ideally based on the encoder's closing hidden state RNN. As a consequence, this model is unable to provide correct results when addressing long-term questions.

## **2.5 Challenges**

Data collection is the most difficult task in any research-based project. In this project collecting 10,000 data was the most challenging part. We manually collected data from different websites, books, and Wikipedia which took three months. After that, we generated context for both Bangla and English. Generating context for the Bangla language is also a difficult part. We used regular expressions for creating automatic context. But compared with English, the Bangla language is more challenging to work with. Because we do not have many resources for Bangla, it is quite challenging. It includes the NLTK library, which has made data processing in Bengali a bit simpler. As a result, during the data preparation stage, we'll need new coding to put the facts together as a version's input. We want the Unicode of each punctuation and receive it using fresh coding if we have swapped punctuation from the data. Another shortcoming is that stop words are removed from the data. Other languages, such as English, offer an integrated library for removing stop phrases and words from data. For Bengali, collect stop phrases or words from the internet, then place data in a text report and remove stop phrases or words from the data using that document. So, working with the Bangla language is quite challenging in this project.

## **CHAPTER 3**

### **RESEARCH METHODOLOGY**

#### **3.1 Introduction**

Every project has a distinct technique, and ours is no exception. Now we'll go over the entire process of our work in depth. The research project's purpose is to uncover something new by employing a novel problem-solving approach. The methodology component of all of this is included in the application. We'll go through each component of the model that we used in our work. To develop an automatic question answering system in Bangla and English, we need to perform some obligatory steps. The model applied in this research is a sequence to sequence LSTM based on the attention mechanism. To begin with, comparing the accuracy score of the automatic system of QA in Bangla and English is a type of work not done previously. This work consumes 10,000 datasets for comparison. We have to know what we're working on and why before we run the deep learning model. To get an accurate result, we'll need a complete dataset. Before performing any algorithm, it is critical to create and preprocess datasets thoroughly. This section explains how the many components of our methodology work together to produce excellent results. The capacity to labor and provide for the aristocracy is enhanced by good storytelling and rationalization of the procedure. The method's graphical sketch and a mathematical equation, as well as their representation, aid in comprehending the complete endeavor. Working in the future necessitates a full grasp of technique, which is a critical component.

Furthermore, the graphical representation of the proposed model is essential for a quality research study. Anyone can easily understand a clear process of that work and learn about the models in a short amount of time. Anyone looking at a transparent process graph will have no trouble learning about the models in a short amount of time. The method's footpases are discussed in detail in this section. The figure 3.1.1 represents the full process of the proposed model of a question answering system.

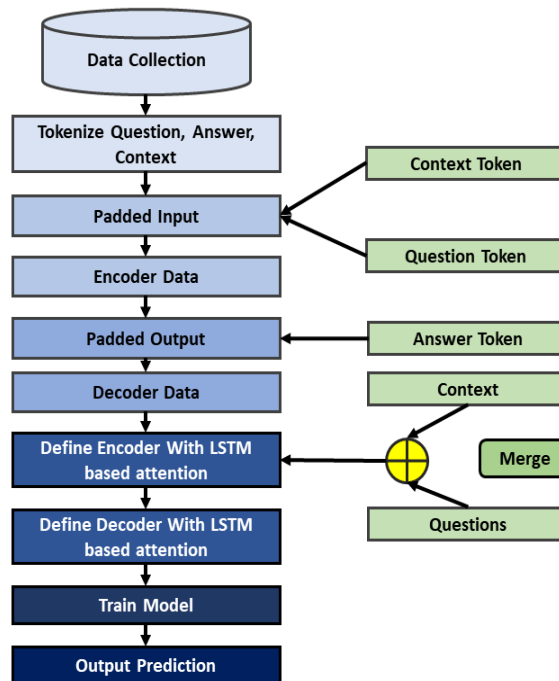


Figure 3.1.1: Work process for Question Answering System

### 3.2 Research Title and Apparatus

Our project-based research title is “STIBGK”: A comparative study between Bangla-English languages on Question-Answer Using LSTM based Attention Mechanism. STIBGK stands for Science, Technology, International and Bangladeshi based General Knowledge which is the domain of the dataset. QA is the trendiest issue in NLP these days. Running a deep learning model with 10,000-dataset demands a high-end PC, GPU, and other equipment, and it will be quite tough for us to work. The following is a summary of the necessary equipment and technologies to get our model up and running.

Development Tools:

- Windows 10
- Python 3.9
- 2.6 version of Tensorflow Backend
- Keras



- NLTK
- Pandas
- Numpy

#### Software and Hardware

- Google Colab with provided GPU
- Intel Core i3 8th generation with 4GB RAM
- 1TB HDD

### **3.3 Data collection**

The deep learning research field needs a huge amount of datasets for getting the better performance of the models and that's why we collect an equal number of data for both Bangla and English language which is respectively 5000. In total 10,000 datasets were used to run our models and also for comparison. To begin with, collect data for Bangla then translate all Bangla data into English. Collecting datasets is the most challenging task. By reading books, questions are made by us manually. Using science books in particular while collecting data for science-based questions. In addition, after all the data has been gathered checking data by one whether to take the correct one or not. Furthermore, the context of our dataset has been done by the coding system cause doing manually context was time-consuming. It takes 5 months with the data collection process. Creating similar datasets for both languages which helps us to compare appropriately. With the help of Wikipedia, books, and many websites our datasets have been collected. Our data was separated into three columns, questions, context, and replies. A tabular representation of our own gathered dataset is shown below.

Table 3.3.1: Sample data of our Bangla dataset

Question	Context	Answer	Domain
বাংলাদেশের সর্ববৃহৎ জেলা কোনটি?	বাংলাদেশের সর্ববৃহৎ জেলা রাঙ্গামাটি	রাঙ্গামাটি	Bangladesh
সেন্টমার্টিন দ্বীপে কত প্রজাতির উভচর রয়েছে?	সেন্টমার্টিন দ্বীপে ৪ প্রজাতির উভচর রয়েছে	৪	
নজরুলের প্রথম প্রকাশিত কবিতার নাম কি?	নজরুলের প্রথম প্রকাশিত কবিতার নাম মুক্তি	মুক্তি	
সূর্যোদয়ের দেশ কোনটি ?	সূর্যোদয়ের দেশ জাপান	জাপান	International
বিশ্বের সবচেয়ে বড় শহর কোনটি ?	বিশ্বের সবচেয়ে বড় শহর লন্ডন	লন্ডন	
বৃটিশ ভারতের শেষ ভাইসরয় কে ছিলেন ?	বৃটিশ ভারতের শেষ ভাইসরয় লর্ড মাউন্টব্যাটেন ছিলেন	লর্ড মাউন্টব্যাটেন	
জিহ্বায় কোন পেশী থাকে ?	জিহ্বায় সরেখ পেশী পেশী থাকে	সরেখ পেশী	Science
আমরা যে চক দিয়ে লিখি তার উপাদান কী?	আমরা যে চক দিয়ে লিখি তার উপাদান ক্যালসিয়াম সালফেট	ক্যালসিয়াম সালফেট	
জীবজগতের জন্যে সবচেয়ে ক্ষতিকারক রশ্মি কোনটি?	জীবজগতের জন্যে সবচেয়ে ক্ষতিকারক রশ্মি গামা রশ্মি	গামা রশ্মি	
পয়েন্ট টু পয়েন্ট প্রোটোকল অথেনটিকেশন কি কি ?	পয়েন্ট টু পয়েন্ট প্রোটোকল অথেনটিকেশন পিএপি এবং সিএইচএপি	পিএপি এবং সিএইচএপি	

সোর্স থেকে দূরবর্তী জায়গায় কিভাবে নিয়ন্ত্রণ করা যায়?	সোর্স থেকে দূরবর্তী জায়গায় ডিএইচসিপি রিলে নিয়ন্ত্রণ করা যায়	ডিএইচসিপি রিলে	Technology
কম্পিউটার না থাকলে "বুট" করতে পারে না কোনটি ?	কম্পিউটার না থাকলে "বুট" করতে পারে না অপারেটিং সিস্টেম	অপারেটিং সিস্টেম	

Table 3.3.2: Sample data of our English dataset

Question	Context	Answer	Domain
Which is the largest district in Bangladesh?	Rangamati is the largest district in Bangladesh	Rangamati	Bangladesh
How many species of amphibians are there on St. Martin's Island?	There are 4 species of amphibians on St. Martin's Island	4	
What is the name of Nazrul's first published poem?	The name of Nazrul's first published poem is Mukti	Mukti	
Which is the country of sunrise?	The country of sunrise is Japan	Japan	International
Which is the largest city in the world?	London is the largest city in the world	London	
Who was the last Viceroy of British India?	The last Viceroy of British India was Lord Mountbatten	Lord Mountbatten	
What muscles are in the tongue?	The tongue has straight muscles	straight muscles	

What are the elements of the chalk we write with?	The chalk we write with contains calcium sulfate	calcium sulfate	Science
Which is the most harmful ray for the living world?	The most harmful ray to the living world is gamma-ray	Gamma-ray	
What is point-to-point protocol authentication?	Point-to-Point Protocol Authentication PAP and CHAP	PAP and CHAP	Technology
How to control remote places from the source?	DHCP relays can be controlled remotely from the source	DHCP relays	
Which can't "boot" without a computer?	Without a computer, the operating system cannot "boot"	operating system	

### 3.4 Data preprocessing

We have followed a few steps for data pre-processing. After collecting data started preparing the datasets. Take three columns as questions, context, and answers. Where questions and answers are collected manually and contexts are created for the collecting questions. To create context, we use regular expressions in our excel datasets. We replace the regular expressions for instance “কোন | কি | কোথায় | কোন দেশ | কে | কত | কী কী | কোনটি | কী| কে| কয়টি” from Bangla questions with the corresponding answers along with the stop words such as “? | ” with “ ”. For English, translate Bangla datasets. For data processing, we must take certain steps. Initially, we employed encoder methods to convert our data format. Then the encoder layer is applied to the datasets and the encoder splits the datasets. There is a built-in Tokenizer in the encoder so it also tokenizes the data into individual word segments. The data is then checked for split text. After that, we separated the data and tokenized the datasets. We have applied pad sequence for data preparation.

Finally, in order to manage our datasets, we combine context and question data. The preprocessing flow is given in figure 3.4.1:

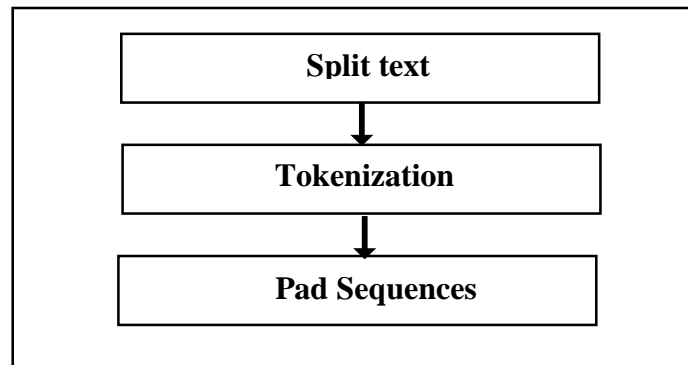


Figure 3.4.1 Dataset preprocessing

### 3.4.1 Split text

Researchers can comprehend how complicated texts in sentences operate. For this aim, various NLP tools have already generated English models. If you're given a long paragraph of material to examine, the simplest method to do so is to break it up into numerous sentences. We quantify data in real talks by separating sentences into different levels and evaluating the consonant words. In the case of text paragraphs, nevertheless, splitting sentences accurately in raw code can be problematic, which can be done rather easily with the help of NLT.

### 3.4.2 Tokenization

Tokenization is a well-developed strategy that uses a kind of data security and randomly generated synthetic values to replace a clean standard. Which usually requires more relevant testing, such as altering downstream applications, enhancing data exchange, and securing data. In the modern data environment, the very first step in comprehending tokenization is to employ more complex and inventive approaches to acquire access and control of data to outsiders. The second phase is to use fences and walls as a boundary.

However, when the entity's borders become more dispersed due to the use of the Cloud, SaaS, and third-party data management, the solution is to manually protect the data using fine-grained data security. Replace a clear-text value, such as 'Robert,' with 'ABC.' Following that, rather than the delicate plaintext 'Robert,' use 'ABC' for retention, transport, and evaluation. Tokenization necessitates a random mapping between the actual value and the resulting encrypted value. This unpredictability exists because encoding is a more secure method of securing data than storing versions.

### 3.4.3 Pad Sequences

Pad sequence is mainly used to make sure the sequential equivalent list of the length. In the Encoding layer, the padding 0 is performed for starting all sequences to the identical length as a prolonged sequence of all sequences. For doing this task easier we used pad sequence and Keras function. For Bangla pad sequence can represent as ([[‘ম্যানগ্রোভ’, ‘কি’], [‘সংশপ্তক’, ‘কার’, ‘রচনা’]]). The pad sequence is a 2D array and considers the first row and second row as [1, 2] and [3, 4, 5]. From the example, we can see that [3,4,5] is a prolonged sequence that’s why putting 0 in the first sequence matches the identical length as [0,1,2]. For English, the method is similar. To illustrate ([[‘Which’, ‘acid’, ‘in’, ‘vinegar’], [‘what’, ‘is’, ‘light’]]) is also 2D array can be examined in the first row as [1, 2, 3, 4] and the second one is as [5, 6, 7]. Hence [1, 2, 3, 4] is a prolonged sequence for that putting 0 into the second one for matching the same length as [0, 5, 6, 7].

### 3.5 Statistical Analysis

In our project work, we collected 10,000 datasets in both Bangla and English languages. Where the datasets are equal for both languages. Our datasets contain four closed domains as Bangladesh Gk, International Gk, Science Gk, and Technology Gk. In this domain the amount of data is equal and the amount is 1250 for each. There are two encoder part questions and context, the decoder part is the answer. For Bangla the questions input maximum length is 23, the context input maximum length is 25 and the answers input

maximum length is 11. For English, the question input maximum length is 36, the context input maximum length is 34 and the answers input maximum length is 13. On the other hand, for Bangla data, the number of Input tokens for questions is 7414, the number of Input tokens for context is 10303 and the number of Input tokens for answers is 4579. For English data, the number of Input tokens for questions is 6246, the number of Input tokens for context is 8081 and the number of Input tokens for answers is 3932. Here for Training, we use 3500 samples and for validation, we use 1500 samples for both Bangla and English languages. We used a Microsoft Excel file to save the dataset. Now represent the length & input token of context, question, and answer for Bangla and English of our datasets.

### **3.6 Model Description**

For developing an automatic question answering system in Bangla and English, need to perform some obligatory steps. This section elaborates the whole process of the extraction of the question answering system. Starting from data collection is a very challenging task in any research. After that, pre-process the dataset and model implementation. The model applied in this research is the sequence to sequence LSTM based on the attention mechanism. To begin with, comparing the accuracy score of the automatic system of QA in Bangla and English type work not done previously. In this section we are going to describe about our three layer model of our work. Which are LSTM, Sequence to Sequence and Attention Mechanism.

#### **3.6.1 LSTM**

The LSTM model is an RNN based structure used in NLP and time-series data forecasting. The LSTM resolves a significant problem with recurrent neural networks by consuming short memory methods. The LSTM knows how to keep, ignore or forget data points based on a prediction model by using a sequence of "doors". LSTM is also used to handle issues with explosive and disappearing gradients. These difficulties arise as a result of repetitive customization while a neural network trains. After making a forecast,

the model is used to predict the following data in the series. Each forecast introduces some imperfection into the system. Data are 'squashed' using sigmoid & tanh activation laws prior to door entrance & output to minimize rising differences. This is the way LSTM works in this model. LSTM is a type of recurrent neural network to use sequential datasets. RNN cannot work for long-term memory. LSTM works to overcome this boundary by adding memory structure. There are few equation for LSTM is below,

$$f_t = \sigma ([h_{t-1}, x_t] + b_f) \dots\dots\dots (1)$$

$$C_t = f_t * C_{t-1} + i_t * \sigma ([h_{t-1}, x_c] + b_c) \dots\dots\dots (2)$$

$$O_t = \sigma ([h_{t-1}, x_t] + b_o) \dots\dots\dots (3)$$

$$i_t = \sigma ([h_{t-1}, x_t] + b_i) \dots\dots\dots (4)$$

$$h_t = O_t * \sigma (C_t) \dots\dots\dots (5)$$

LSTM is explain as protect a memory cell =  $C_t$  which is reset, write and read from following to the forget gate =, the input gate =  $i$ , the output gate =  $O_t$ , hidden state =  $h_t$  and here  $\sigma$  is activation function. The architecture of LSTM model is given in figure 3.6.1.1.



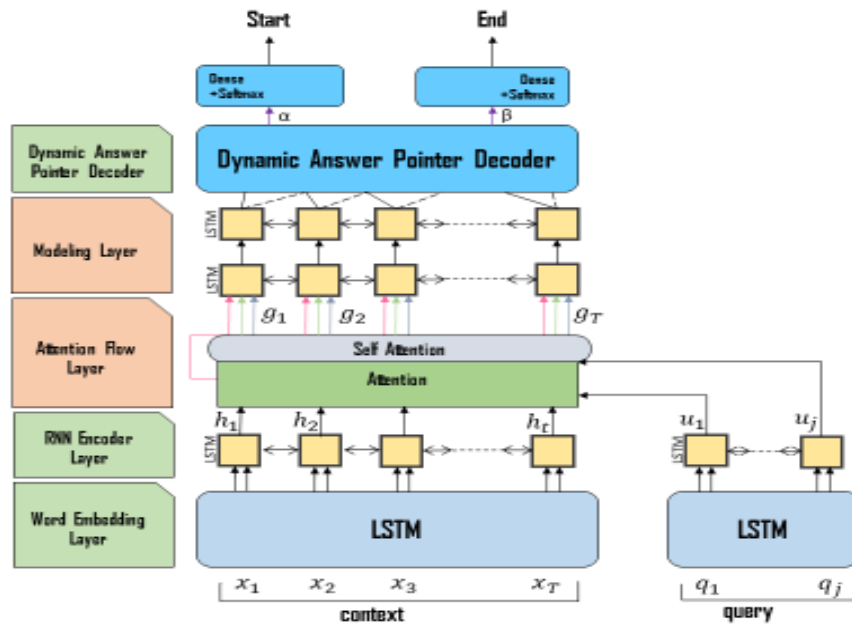


Figure: 3.6.1.1 Architecture of LSTM model

### 3.6.2 Sequence to Sequence

To complete the Question Answering challenge in the Bangla-English Datasets, a sequence-to-sequence attention reading comprehension framework was constructed. Many of the technologies are based on sequence to sequence concepts. For instance, the seq2seq model powers Machine Translation, online Chabot's, voice-activated gadgets and QA systems. This model is a particular part of LSTM architectures. In a seq2seq paradigm, the encoder and decoder are made up of two parts. In both of them, LSTM blocks may be discovered. The goal of using the encoder of an LSTM is to transform the representation of a significant vector of input series, then use an LSTM in the decoder to retrieve the target series. Word embedding was then used to complete the vocabulary for the files that had to be included as input. The motive for employing a LSTM in encoder is to create the input series to massive illustration of vector and a LSTM with decoder to induce the target series. The Sequence to Sequence method alter an input order into output order or series. The orders or series should be defined by X and Y. Then the input sequence i-th element and output sequence the j-th element can be narrated as  $x_i$  and  $y_j$ .

The element of each onehot vector tokens is also define by  $x_i$  and  $y_j$ . The vocabulary of the inputs and outputs are define by  $^{(s)}$  and  $^{(t)}$ . All the important component of  $x_i$  and  $y_j$  part  $x_i \in R^{|\mathcal{V}^{(s)}|}$  and

$y_j \in R^{|\mathcal{V}^{(t)}|}$ . Now the equation for X and Y is,

$$X = (x_1 \dots x_I) = (x_i)_i^I = 1 \dots \dots \dots (1)$$

$$\text{And, } Y = (y_1 \dots y_J) = (y_j)_j^J = 1 \dots \dots \dots (2)$$

For equation (1) and (2),  $I$  = length of input sequence and  $J$  = length of output sequence. Again, For Natural Language Processing,  $y_0$  is the vector of BOS, it is virtual word declaim the initializing of the sentence.  $y_{i+1}$  is EOS, it works adds token for terminate the end.

Now, we discuss the conditional feasibility equation. The feasibility of  $(j | Y_{<j}, X)$ ,

$$P_\theta (Y | X) = \prod_{j=1}^{J+1} P_\theta (y_j | Y_{<j}, X) \dots \dots \dots (3)$$

Here, conditional probability =  $\square (\square | \square)$ , Seq2seq modeling the probability =  $\square (\square | \square)$ , The feasibility of j-th components of  $y_j$  given  $Y_{<j}$  and  $\square = P (y_j | Y_{<j}, X)$ .

The next discussion part is sequence model processing. The method produce standing vector  $z$  and the input  $X$ . Now we could mention,

$$z = \Lambda (\square) \dots \dots \dots$$

..... (4) Here,  $\Lambda$  = Function of the recurrent neural network of LSTM methods.

Now,

$$P_\theta (y_j | Y_{<j}, X) = \gamma (h_j^{(t)}, y_j) \dots \dots \dots (5)$$

$$\text{And, } h_j^{(t)} = \Psi (h_{j-1}^{(t)}, y_{j-1}) \dots \dots \dots (6)$$

Here,  $\square$  = hidden vector of  $h^{(t)}$ ,  $\gamma$  = calculated the probability of vector  $y_j$ .

For encoder embedding layer transfers the embed vector of every words. So the equation of embedding vector is,

$$\bar{x}_i = E^{(s)} x_i \dots \dots \dots (7)$$

In encoder the embedding matrix is  $^{(s)}D \times |\mathcal{V}^{(s)}|$ . The encoder recurrent layer generate hidden

vectors.  $\Psi^{(s)}$  is the uni-directional Recurrent Neural Network (RNN) function. So equation is,

$$h_j^{(s)} = \Psi^{(s)}(\bar{x}_j, h_{j-1}^{(s)})$$

$$= \dots \dots \dots \tanh \left( W^{(s)} \begin{bmatrix} h_{i-1}^{(s)} \\ \bar{x}_i \end{bmatrix} + b^{(s)} \right) \dots \dots \dots (8)$$

Here, activation function is tan h.

For decoder embedding layer recurrent layer function is  $\Psi^{(t)}$ . The equation of decoder is below,

$$\bar{y}_j = {}^{(t)}y_{j-1} \dots \dots \dots (9)$$

$$h_i^{(t)} = \Psi^{(t)}(\bar{x}_i, h_{i-1}^{(t)})$$

$$\dots (10) \quad \quad \quad = \tanh \left( W^{(t)} \begin{bmatrix} h_{i-1}^{(t)} \\ \bar{x}_i \end{bmatrix} + b^{(t)} \right) \dots \dots \dots$$

$$\text{So, } h_0^{(t)} = z = h_i^{(s)}$$

The decoder output layer equation is below,

$$P_j = P_{\theta} (| Y_{<j}) = \text{softmax} (O_j). y_j$$

$$= \text{softmax} (W^{(o)}. h_i^{(t)} + b^{(o)}). y_j \dots \dots \dots (11)$$

A graphical view and architecture are important when understanding for every model. What has happened or research methodology inside a model can be easily understood by looking at an architecture in figure 3.6.2.1.

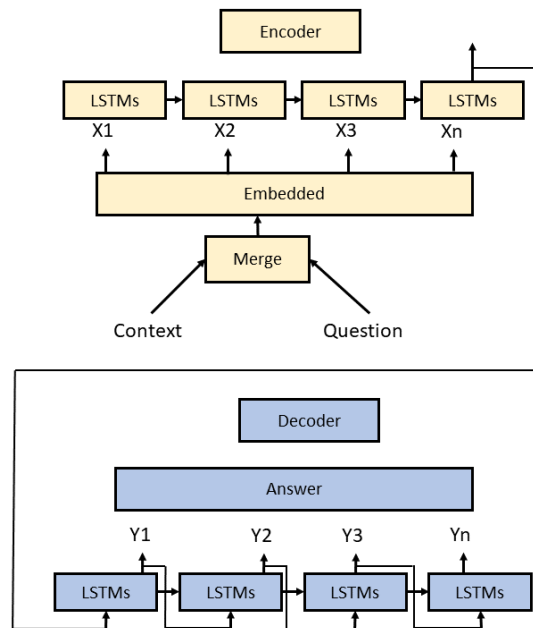


Figure 3.6.2.1: Architecture of Seq2seq model

### 3.6.3 Attention Mechanism

Contexts and queries are used as inputs in the encoding section before being sent into a layer for embedding. Both embedding vectors are then sent via LSTM. Following that, we output both contexts and queries and then transmit the unified output to the attention layer. The next decoder layer generates the queries in the encoder and decoder attention layers, while the output generates the storage keys and data. This allows each decoder site to focus on each individual place inside the input pattern. The encoder includes self-attention layers. A self-attention layer's context, responses, and questions all originate from the same source, in this case, the encoder's previous layer's response. Every encoder site can respond to all previous encoder layers. Another LSTM is utilized in the decoding process to evaluate information. After calculating all of the necessarily encoded vectors, we will use a softmax function to normalize them. The weights of the encoded vector are then multiplied to form a "time-dependent" that is given into the decoder. Every encoder site can respond to all previous encoder layer locations. The model is given in figure 3.6.3.1.

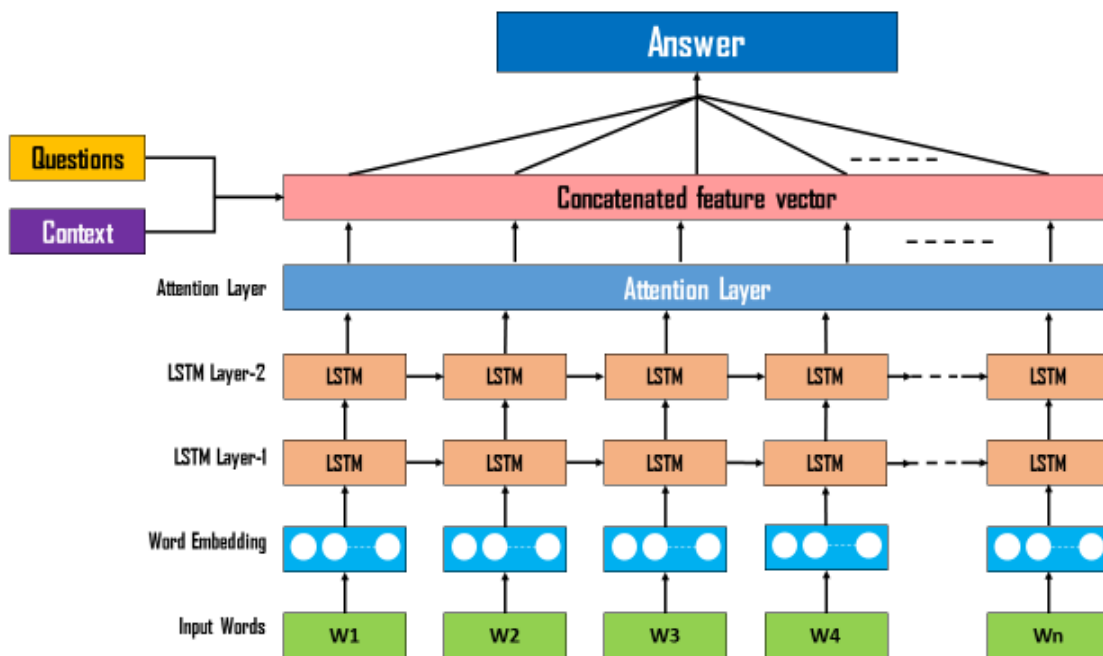


Figure 3.6.3.1: Architecture of Attention Mechanism

### 3.7 Representation of Taxonomy for Our Model

Taxonomy seeks to give concepts and the words used to express them a chronological context. In order to make information useful for both humans and technology, taxonomists are interested in how individuals use language to classify and identify concepts. NLP in taxonomy design is essentially a bottom-up process in which Named Entity Recognition captures the content's lowest level concepts. The taxonomist can then classify these words into bigger groups. This is supplemented by top-down analysis when working with SMEs to identify and fine-tune the categories, completing EK's hybrid taxonomy design technique.

In our project work, the taxonomy is divided into five sections: task description, evaluation methods, corpus, and questions type and answer type of our proposed methods. Where the

task description classifies as Context, visual QA, and Closed domain. The Evaluation classifies as accuracy, validation, tokenization, padding, context question answering. Again context question answering is classified as length, merge, and embedding. The corpus classifies as context type text, size of train test, data collection, data source, and language. The question type classifies as text and close. Last of all the answer type classifies as span and free form text. The taxonomy diagram is given in figure 3.7.1.

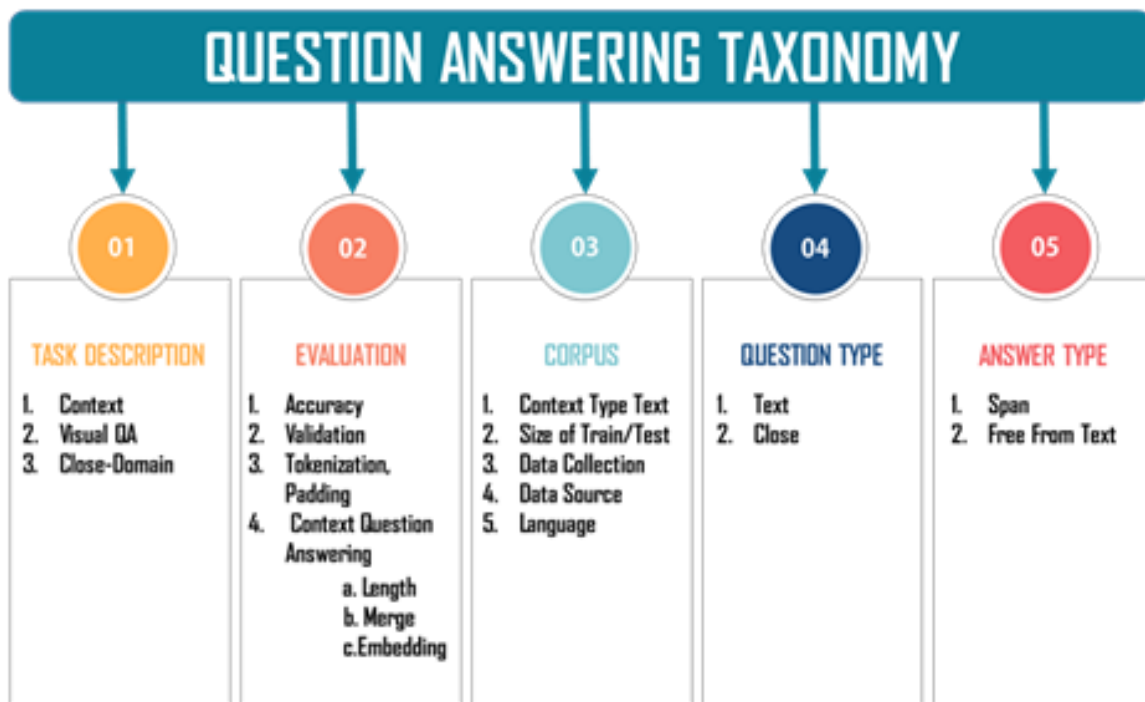


Figure 3.7.1: Architecture of Taxonomy

## **CHAPTER 4**

### **EXPERIMENTAL RESULTS AND DISCUSSION**

#### **4.1 Introduction**

In Natural Language Processing, Automated Question Answering is a challenging and critical topic. Having the correct responses from a machine is a really complex process. Whenever someone gives the machine a question, it will respond, but that is not the goal. The machine's primary function is to provide accurate results instantly in response to any question. Probability plays a crucial role in this question-answering method. Since a machine's correct output is based on the highest plausibility. Because after the machine has learned each term, it will calculate the probability and provide us with the response to the corresponding question. A machine is in charge of every trained model. In the tensor flow 2.6.0 version, the seq2seq LSTM model was employed. Sequence to sequence LSTM has been deployed in this system, which is based on attention mechanisms. Hyper-parameters are predetermined parameters that are used in the training procedure. After training, the model is able to develop a customized responding method for the machine. It will get feedback from the datasets for the answer, and the length will be generated at random. The training process of every variable was calculated using the Adam optimizer with epochs, batch size, and validation split, verbose. Our system has values for batch size, epochs, verbose, and validation split for two languages because it consumes two languages.

For the English language & Bangla language the value of batch epochs=35, size=32, validation split=0.15, verbose=1. Now, the batch size decides the amount of input sequence sent for the network. Batch size is just another element that influences categorization efficiency. The larger the batch size, the more time it takes to train the dataset, but as a condition, the model's reliability and storage requirements deteriorate. However, we should tread cautiously while choosing the batch size. Another parameter is epochs = 35, which represents the number of repetitions in epochs. The number of iterations is a major parameter that determines how several repetitions the learning algorithm can run over the

whole testing set. Every instance in the testing set got the opportunity to modify the inner hyper parameters one every epoch. Using sparse gradients, Adam optimization enhances the speed of problems. Its purpose is to create a deep neural network.

Adam optimizer is particularly significant for the retraining phase since it analyzes all parameters. When a deep learning algorithm model is trained, a well-configured computer or laptop is essential. A GPU is required for database training. Initially, we used a direct computer to train our model. As a reason, running the model takes a while, and the obtained outcomes are unsuitable for addressing questions. So, for this project, we used Google Colab to train our model. It allows consumers a free GPU function.

#### **4.2 A comparison of QA models for Bangla and English**

Almost all know that no system can produce 100% precise findings. The methodology we utilized for our project provided satisfactory accuracy, however, it was not usually 100 percent accurate. It occasionally produced incorrect output in a brief span of time. When dealing with Bengali Language processing, the amount of incorrect output is minimal.

Now the graphical representation of the Loss, Accuracy for English is given in figure 4.2.1.



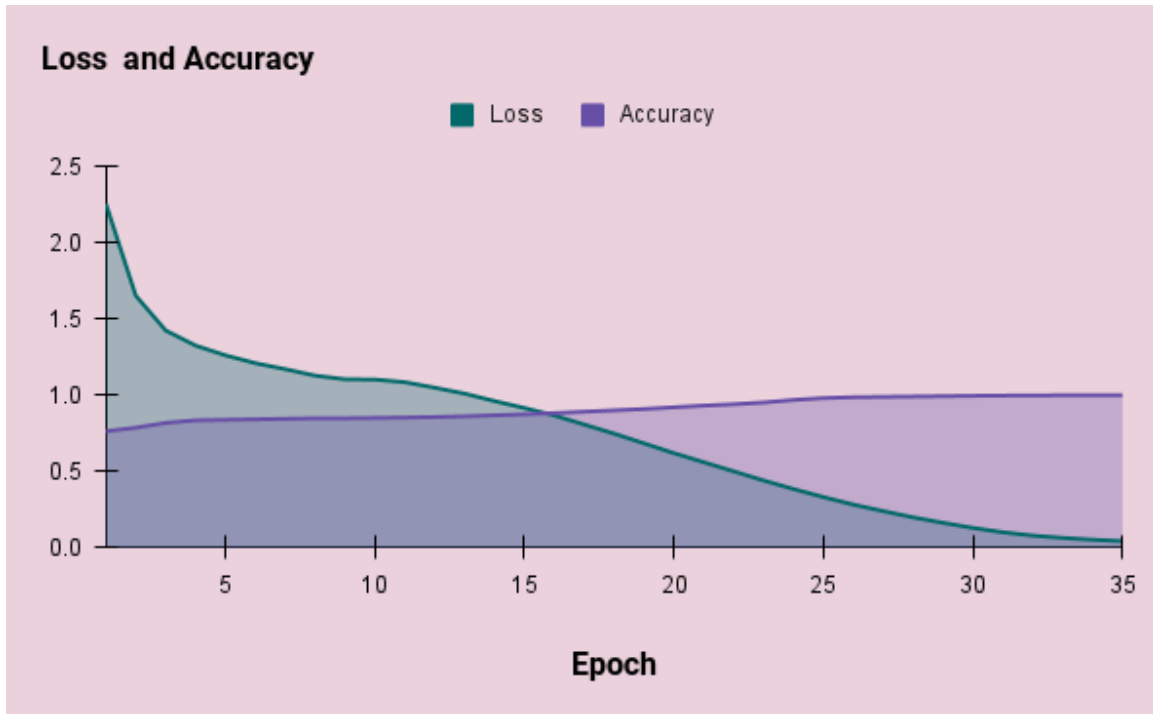


Figure 4.2.1: Graphical representation of loss and accuracy for English

The loss and accuracy corresponding to an epoch are depicted in, with epoch on the x-axis and loss and accuracy on the y-axis. The experiment ran 35 epochs and the loss reduces from 2.0494 to 0.1397. On either side, the accuracy expanded from 0.7829 to 0.9948. As a result of the statistics, it can be concluded that when the loss decreases the accuracy increases.

Now the graphical representation of the VAL\_loss, Val\_accuracy for the English is given in Figure 4.2.2.

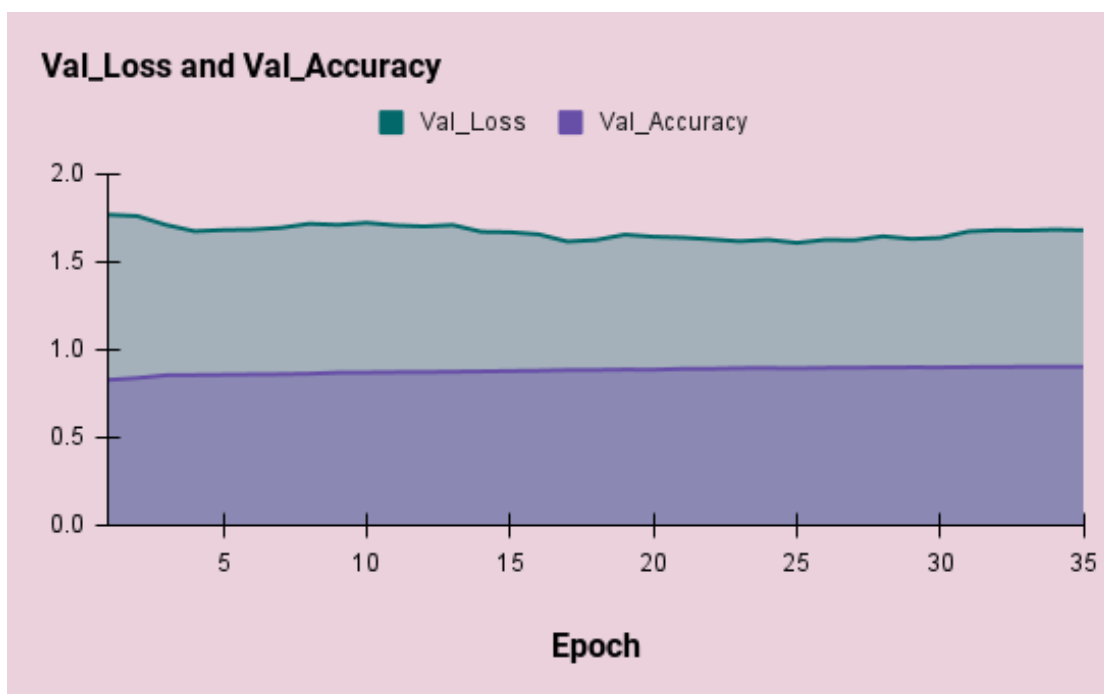


Figure 4.2.2: Graphical representation of validation loss and validation accuracy for English

The diagram depicts Val\_loss and Val\_accuracy for an epoch, with epoch on the x-axis and Val\_loss and Val\_accuracy on the y-axis. Epoch is less than training loss when compared to Val\_loss, with a value of 1.5109 which is 0.53.85 less than loss. However, at epoch35, Val\_loss is more than loss, which is 1.2838 which is more than 1.1441 that is why the Val\_accuracy is increased by train accuracy. The validation accuracy is 0.9016. The accuracy declines from training accuracy as the validation loss grows in epoch 35.

We can observe that the efficiency of our training model and validation accuracy are suitable to be used on the English Question Answering system. Overall, the accuracy of the English QA dataset is excellent. And we can notice that the reliability loss and validity loss for our training model are both quite little.

Now graphical representation of the loss, Accuracy for Bangla is given in Figure 4.2.3.

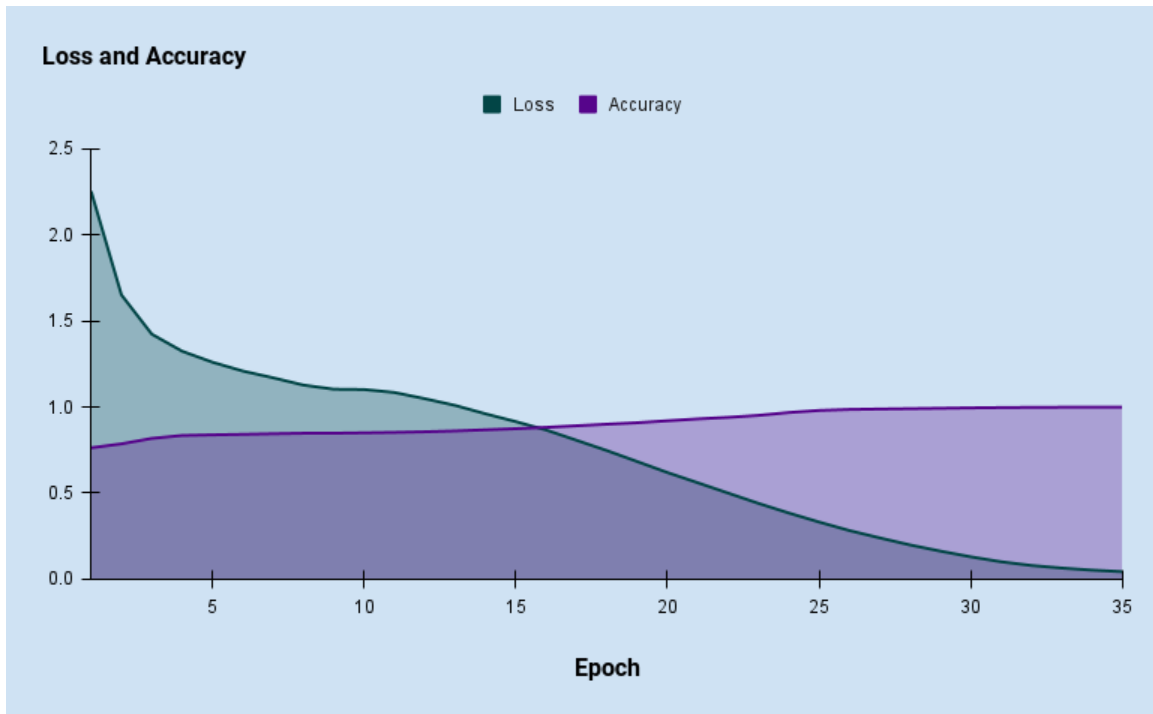


Figure 4.2.3: Graphical representation of loss and accuracy for Bangla

The above loss and accuracy corresponding to an epoch where we put epoch to x-axis and loss and accuracy in the y-axis. The experiment ran 35 epochs and the loss reduces from 2.2542 to 0.0423. On either side, the accuracy expanded from 0.7626 to 0.9991. As a result of the statistics, it can be concluded that when the loss decreases the accuracy increases.

Now graphical representation of the Val\_loss, Val\_accuracy for Bangla is given in Figure 4.2.4.

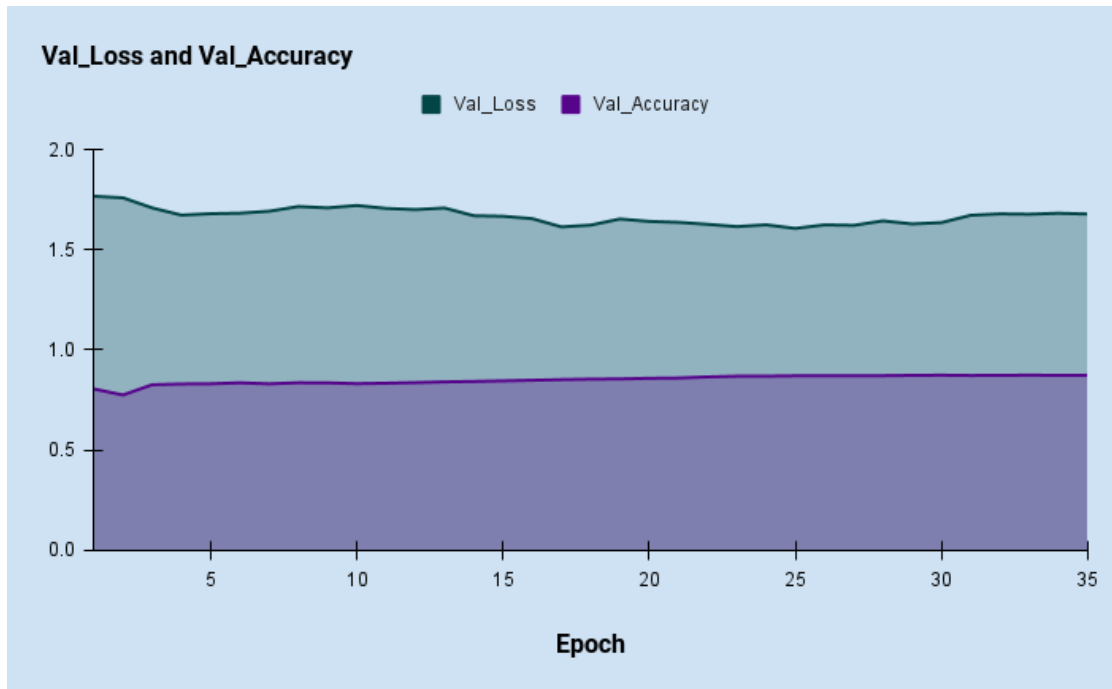


Figure 4.2.4: Graphical Representation of Validation loss and Validation accuracy for Bangla

The above graph shows the loss and accuracy associated with an epoch, with the epoch on the x-axis and the loss and accuracy on the y axis. The epoch is on the x-axis, and Val\_loss and Val\_accuracy are on the y-axis in the above 8 (b) fig concerning Val\_loss and Val\_accuracy relating to an epoch. Epoch is less than training loss when compared to Val\_loss, with a value of 1.7673. This is 0.4869 points less than the loss. However, Val\_loss is increasing faster than training loss in epoch 35, and Val\_loss is now 1.6787. Because train accuracy is more than 1.6787, the Val\_accuracy is enhanced. The accuracy of validation is 0.8736.

We can observe that the efficiency of our training model and validation accuracy are suitable to be used on the Bangla Question Answering system. Overall, the accuracy of the Bangla QA dataset is excellent. Many of the previous studies we've studied at haven't been able to achieve as high accuracy using Bangla datasets. And we can notice that the reliability loss and validity loss for our training model are both quite little.

### 4.3 Prediction Accuracy in Real Life

We come to the conclusion that the English language predicts well and the prediction accuracy is 90.16% wherein in Bangla the prediction accuracy is 87.36%. In real life, if the user asks a slightly different question, the algorithm properly anticipates the answer. After training & testing then we gave individual user input for testing our model in real life. Real-life testing is very important whether it predicts correctly or not. So, we gave slightly different questions to the system and it predicted the answer correctly. After that, we can say our model predicts well for Both Bangla and English languages.

Now the response of the model after training, there are two sample prediction outcomes for Both Bangla and English language of our research is given under in table 3, table 4 and table 5, table 6. In this table, we showed user input which is different from learned questions. The machine's input on the model received after training is shown below:

Table 4.3.1: Sample prediction 1: Bangla Question answering system

<b>Context</b>	বিশ্বকাপ ক্রিকেটে বাংলাদেশ দলের অভিষেক হয় ৭ম বিশ্বকাপে
<b>Question</b>	বিশ্বকাপ ক্রিকেটে বাংলাদেশ দলের অভিষেক হয় কবে?
<b>Answer</b>	৭ম বিশ্বকাপে
<b>User Input</b>	বাংলাদেশ দলের অভিষেক হয় কোন বিশ্বকাপ ক্রিকেটে?
<b>Prediction</b>	৭ম বিশ্বকাপে

Table 4.3.2: Sample prediction 2: Bangla Question answering system

<b>Context</b>	আমরা যে চক দিয়ে লিখি তার উপাদান ক্যালসিয়াম সালফেট
<b>Question</b>	আমরা যে চক দিয়ে লিখি তার উপাদান কী?
<b>Answer</b>	ক্যালসিয়াম সালফেট
<b>User Input</b>	লিখিত চকের উপাদান কী ?
<b>Prediction</b>	ক্যালসিয়াম সালফেট

Table 4.3.3: Sample prediction 1: English Question answering system

<b>Context</b>	In the VTP trunking protocol, server mode villain is added
<b>Question</b>	In VTP trunking protocol, where is the villain added?
<b>Answer</b>	Server mode
<b>User Input</b>	Where is the villain added in the VTP trunking protocol?
<b>Prediction</b>	Server mode

Table 4.3.4: Sample prediction 2: English Question answering system

<b>Context</b>	Rangamati is the largest district in Bangladesh
<b>Question</b>	Which is the largest district in Bangladesh?
<b>Answer</b>	Rangamati
<b>User Input</b>	Which district is the largest in Bangladesh?
<b>Prediction</b>	Rangamati

## **CHAPTER 5**

### **IMPACT ON SOCIETY, ETHICAL ASPECTS AND SUSTAINABILITY**

#### **5.1 Impact on Society**

Our research will have a great impact on our society. In this study we have worked with Automatic Bangla and English question answering. Now the present time is the age of information technology, now in our society information technology has touched all fields. Chabot or Automatic question answering System is currently saw and used in many languages, especially for English language. We have run a model to do this kind of work in Bangla and English language. The model has given very good results for both languages. The main goal of this exploration is to expand the use and innovation of machines in Bangla and English language. Such research will play a very important role in our society as well as worldwide for the both Bangla and English speaking people. Using this kind of research, we can also do Bangla chatbot type work in the future, which will be very useful for the Bengali speaking people of our society to use.

This work will enrich our Bengali and English NLP world as well as pave the way for such work in the future. A lot of work has been done for English language but no such work has been done for Bengali language. We have got interest to work in Bengali language keeping in view this aspect. This is our small effort keeping in mind the innumerable Bengali speaking people living all over the world. Bengali is our mother tongue. So there are many people in our society who have difficulty reading and understanding English. If we can get an answer to a question automatically as soon as we ask a question, it will be very useful for us and it will save us time.

## **5.2 Ethical Aspects**

From an ethical point of view, our work model or type does not violate any human rights and privacy. We have collected “STIBGK” based data to give automatic question answer. We did not collect anyone's name, address or other personal information when we collected the data. Therefore, the data we have cannot be used to identify or harm anyone. We have not done any work or collected data by harming or intimidating people while doing our job. Since our work is data dependent, we have used utmost care while collecting and storing data. We did not take the work of any other organization or person as our own while completing our work. We are 100% hopeful that this work of ours will never harm anyone, this work will help our Bengali along with English NLP a lot in future research. We used our own PCs while doing our job. We have not used any equipment used by any other person and we have not stolen any information or data from any other person. We have done our research maintaining honesty, obedience to the law, integrity, legality and transparency.

## **5.3 Sustainability Plan**

Our main goal is to automatically generate answers from Bangla and English questions. In future, by using our program, many changes can be made in the business organization. Our model will only work for specific datasets. So in order to make this work more advanced and sustainable in the future, we need a lot of datasets related to Bangla and English questions. Applications on the business organizations and various types of organizations like online chatbot are being used in many parts now. In the future, enriches and improves the required our dataset this model can be used for the purpose of educational sectors, military sectors, industrial sectors, business sectors etc with automatically question answering generate system. Therefore, if we can bring the chatbot system of online organizations in both languages for the benefit of the Bengali nation in such a promising work, then many benefits will come to our consumers in the future



## **CHAPTER 6**

### **CONCLUSION AND FUTURE WORK**

#### **6.1 Summary of the Study**

Our entire work with this research is concentrated on Bangla Linguistics. We used a deep learning algorithm to complete our research for Bangla & English Question Answering. This project is really effective in giving an automated Bengali and English question answering system. There has never been any previous work on comparing Bangla and English language both in queries answering systems using a deep learning method. We received excellent efficiency for Bengal and English after fixing the issues of question answering using a deep learning model. Our approach will be useful in the field of Bangla and English NLP research, as well as in upcoming comparative study with any language in question answering initiatives researchers can get help from our work. From the time we started collecting data until the time we finished the project, it took us five months. To complete this, we must go through a series of processes individually. The entire work process has now been described this way:

Process 1: Gather data from Google and Facebook to answer questions.

Process 2: Gathered the answers to the questions.

Process 3: Save all of your data as a.csv file in Excel.

Process 4: Data pre-processing from the dataset.

Process 5: Estimate the vocab.

Process 6: To determine the length of the data sequence, use the Pad sequence.

Process 7: Combine the data from the question and the context.

Process 8: With LSTM based Attention mechanism, use the encoder and decoder.

Process 9: Establish a sequence to sequence structure in

Process 10: Build a model train and put it through its tests.

Process 11: Review the output or results.

We finished the work by executing these procedures in order.

## **6.2 Conclusion**

The main goal of this initiative is to enhance and improve comparative study in any language under the QA system. We took data from Bengali and English questions as input to the model, then after processing the data, the model interacts with the question to produce our customized response as an output. We employed the LSTM based attention mechanism method for encoding and decoding for this automated responding system. We saw a little research in the English and Bangla language individually before we undertook any question-answering related work in both languages. Using LSTM based on attention, comparing the language of both Bangla and English question answering systems in this proposed study. Creating our own datasets has been done on four closed domains. We collect an equal number of data for both the Bangla and English languages which is respectively 5000. In total 10,000 datasets were used to run our models and also for comparison. Both languages have a high level of accuracy and efficiency. The closed domain and our datasets are limitations in this work. If we want answers to any inquiries, our model will not function. This is our program's primary limitation as well. In Bengali, data processing is a little more complicated than in English. There are many processing libraries in English for cleaning data so for English it is quite easier for us than in Bangla. As a result, a preprocessing library for Bengali took more time. Despite the difficulties we had dealing with Bengali, our model produced a pretty decent result in both linguistics. The English data gives 99.48% accuracy where the Bangla data gives 99.81% accuracy. But the validation accuracy of English languages is quite better than in Bangla which is 90.16%. On the other hand, the validation accuracy of Bangla is 87.36%. The result can be concluded that English fared better than Bangla in this model. The accuracy scores of Bangla and English are compared using Seq2Seq LSTM in this study, which is novel work.

### **6.3 Recommendations**

We must put in a lot of effort to deal with a comparative study with another language like Bangla with Hindi or English with France QA system because this kind of study is a complicated approach. We'll need a substantial and well-organized dataset for these studies. We require a good PC configuration and a complete GPU system to operate with it. Before working with Bangla, keep in mind that gaining enough datasets to work with Bengali is really tough. However English is easier to evaluate than another language. We can expand the number of datasets in order to improve the accuracy of the models in the next, and so gain greater outcomes from this research. Humans can witness the influence and role of automatic question answering on numerous online platforms currently, and we can anticipate that this study will have a large future impact by focusing on the Bengali as well as the English language. Our study can help those who cannot understand English but understand Bangla. Our model can give answers to both languages. Repeatedly answer this question which is both a nuisance and a waste of time. If we can properly train the machine while keeping the particular questions that must be answered, the machine will provide us with a good response using the suitable approach, saving us both effort and time.

### **6.4 Implication for Further Study**

Every work has some limitations. Our project also has some limitations. We discovered several difficulties in doing so, such as the fact that we had to deal with closed domains and that our dataset was insufficient. Any form of the model is built for incoming improvement, as we already know. Because any form of experimental effort is a never-ending process that improves day by day. We compared our model in both Bangla - English. This is a new work in this field. But in the future, it will develop more. We've just commenced our research on questions and answers comparing Bangla and English, and we'll need to go much further with it. Next time, we'll try to build a more complex and narrative QA system using open domain and expand our own datasets. Following the completion of this investigation, we will need to expand the model. In this research, we

only employed one model. We'd like to expand the dataset and test some additional models in the future. This will help us figure out which model to utilize for this project. More data will be added to these domains in the future, as well as the ability to extract certain domains from given datasets. Extracting and anticipating the domain of a system of provided questions is a new job for scholars, and that's something I will explore in the future. After the study is completed, we want to focus on how to implement it in our daily life. Because if a piece of study isn't valuable to people, it might be dismissed as insignificant. This study will become a reality through developing web-based and mobile apps based on the future of artificial intelligence.

## REFERENCES

- [1] M. Keya, A. K. M. Masum, B. Majumdar, S. A. Hossain, and S. Abujar, “Bengali question answering system using seq2seq learning based on general knowledge dataset,” ICCCNT,2020.
- [2] M. Keya, A. K. M. Masum, S. Abujar, S. Akter, and S. A. Hossain, “Bengali context–question similarity using universal sentence encoder,”in *Emerging Technologies in Data Mining and Information Security* Springer, 2021.
- [3] T. Tahsin Mayeessa, A. Md Sarwar, and R. M. Rahman, “Deep learning based question answering system in Bengali,” *Journal of Information and Telecommunication*, vol. 5, no. 2.
- [4] S. Sarker, S. T. A. Monisha, and M. M. H. Nahid, “Bengali question answering system for factoid questions: A statistical approach,” 2019, ICBSLP.
- [5] S. A. A. Sagor, N. A. Rizvi, and M. M. H. Nahid, “Machine learning approaches for Bengali automated question detection system,” *IJCA*, vol. 975, p. 8887.
- [6] S. Banerjee, S. K. Naskar, and S. Bandyopadhyay, “Bfqa: A Bengali factoid question answering system,” in *International Conference on Text, Speech, and Dialogue* Springer, 2014, pp. 217–224.
- [7] M. R. Bhuiyan, A. K. M. Masum, M. Abdullahi-Oaphy, S. A. Hossain, and S. Abujar, “An approach for Bengali automatic question answering system using attention mechanism,” in *2020 11th International Conference on Computing, Communication and Networking Technologies*.
- [8] C. Li, L. Liu, and F. Jiang, “Intelligent question answering model based on cn-bilstm,” in *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence*, 2018, pp. 447–450.
- [9] S. Hakimov, S. Jebbara, and P. Cimiano, “Deep learning approaches for question answering on knowledge bases: an evaluation of architectural design choices,” 12 2018.
- [10] Z. Abbasi-taeb and S. Momtazi, “Text-based question answering from information retrieval and deep neural network perspectives: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1412, 2021.
- [11] K. Lei, Y. Deng, B. Zhang, and Y. Shen, “Open-domain question answering with character-level deep learning models,” in *2017 10th International Symposium on Computational Intelligence and Design*.
- [12] L. Xiao, N. Wang, and G. Yang, “A reading comprehension style question answering model based on attention mechanism,” in *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP) IEEE*, 2018, pp. 1–4.
- [13] V. Yadav, S. Bethard, and M. Surdeanu, “Alignment over heterogeneous embeddings for question answering,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2019, pp. 2681–2691.
- [14] Purnendu Mukherjee, “Question Answering Systems with Attention Mechanism”, May 2018.

## STIBGK

### ORIGINALITY REPORT

<b>13%</b>	<b>10%</b>	<b>6%</b>	<b>7%</b>
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

### PRIMARY SOURCES

<b>1</b>	<b>dspace.daffodilvarsity.edu.bd:8080</b> Internet Source	<b>4%</b>
<b>2</b>	<b>Submitted to Daffodil International University</b> Student Paper	<b>3%</b>
<b>3</b>	<b>"Emerging Technologies in Data Mining and Information Security", Springer Science and Business Media LLC, 2021</b> Publication	<b>&lt;1%</b>
<b>4</b>	<b>Mumenunnessa Keya, Abu Kaisar Mohammad Masum, Bhaskar Majumdar, Syed Akhter Hossain, Sheikh Abujar. "Bengali Question Answering System Using Seq2Seq Learning Based on General Knowledge Dataset", 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020</b> Publication	<b>&lt;1%</b>
<b>5</b>	<b>Sayli Uttarwar, Simran Gambani, Tej Thakkar, Nikahat Mulla. "Machine learning based review on Development and Classification of Question-Answering Systems", 2019 3rd</b>	<b>&lt;1%</b>

International Conference on Computing Methodologies and Communication (ICCMC), 2019

Publication

---

6	deepai.org Internet Source	<1 %
7	Gulraiz Khan, Muhammad Ali Farooq, Zeeshan Tariq, Muhammad Usman Ghani Khan. "Deep-Learning Based Vehicle Count and Free Parking Slot Detection System", 2019 22nd International Multitopic Conference (INMIC), 2019 Publication	<1 %
8	Md. Rafiuzzaman Bhuiyan, Abu Kaisar Mohammad Masum, Md. Abdullahil-Oaphy, Syed Akhter Hossain, Sheikh Abujar. "An Approach for Bengali Automatic Question Answering System using Attention Mechanism", 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020 Publication	<1 %
9	Saranlita Chotirat, Phayung Meesad. "Part-of-Speech tagging enhancement to natural language processing for Thai wh-question classification with deep learning", Heliyon, 2021 Publication	<1 %

---

10	Masoumeh Rajabi, Mohammad Ehsan Basiri, Shahla Nemati. "Identifying High-Quality User Replies Using Deep Neural Networks", 2021 7th International Conference on Web Research (ICWR), 2021 Publication	<1 %
11	<a href="http://www.ijcaonline.org">www.ijcaonline.org</a> Internet Source	<1 %
12	<a href="http://www.springerprofessional.de">www.springerprofessional.de</a> Internet Source	<1 %
13	Arpita Mashyal, Basavaraj Chougula, Sukanya Kobal, Harshala Gopal Bajantri, Veeresh. "Facial Mask Detection and Alert System", 2021 International Conference on Intelligent Technologies (CONIT), 2021 Publication	<1 %
14	<a href="http://aut.ac.ir">aut.ac.ir</a> Internet Source	<1 %
15	"Cyber Security Intelligence and Analytics", Springer Science and Business Media LLC, 2020 Publication	<1 %
16	<a href="http://researchr.org">researchr.org</a> Internet Source	<1 %
17	Submitted to University of Westminster Student Paper	<1 %



18	Yaobang Chen, Jie Shi, Xingong Cheng, Xiaoyi Ma. "Hybrid Models Based on LSTM and CNN Architecture with Bayesian Optimization for ShortTerm Photovoltaic Power Forecasting", 2021 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia), 2021 Publication	<1 %
19	Submitted to Midlands State University Student Paper	<1 %
20	Priyanka Ravva, Ashok Urlana, Manish Shrivastava. "AVADHAN", Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, 2020 Publication	<1 %
21	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2020 Publication	<1 %
22	Arnab Saha, Mirza Ifat Noor, Shahriar Fahim, Subrata Sarker, Faisal Badal, Sajal Das. "An Approach to Extractive Bangla Question Answering Based On BERT-Bangla And BQuAD", 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021 Publication	<1 %
23	Submitted to Far Eastern University Student Paper	<1 %

24	<a href="https://sites.google.com">sites.google.com</a> Internet Source	<1 %
25	Submitted to Queen's University of Belfast Student Paper	<1 %
26	Submitted to School of Business & Computer Science Limited Student Paper	<1 %
27	Submitted to King's College Student Paper	<1 %
28	<a href="https://dspace.library.daffodilvarsity.edu.bd:8080">dspace.library.daffodilvarsity.edu.bd:8080</a> Internet Source	<1 %
29	<a href="http://www.mdpi.com">www.mdpi.com</a> Internet Source	<1 %
30	<a href="http://www.sust.edu">www.sust.edu</a> Internet Source	<1 %
31	<a href="http://mrlaptop.pk">mrlaptop.pk</a> Internet Source	<1 %
32	"Proceedings of International Joint Conference on Computational Intelligence", Springer Science and Business Media LLC, 2020 Publication	<1 %
33	Liu, Song, and Fuji Ren. "Paragraph act based pragmatic information extraction in question answering", 2011 IEEE International	<1 %

## Conference on Cloud Computing and Intelligence Systems, 2011.

Publication

34

thesai.org  
Internet Source

<1%

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off