Bangla Question Answering Based Model Using Sequence to Sequence Learning

BY

Bhaskar Majumdar
ID: 172-15-9988

Mumenunnessa Khan
ID: 172-15-10100

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

Supervised By

Sheikh Abujar
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Md. Tarik Habib
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JUNE 2021

APPROVAL

This Project titled "**Bangla Question Answering Based Model Using Sequence to Sequence Learning**", submitted by Bhaskar Majumdar, ID No: 172-15-9988 and Mumenunnessa Khan, ID No: 172-15-10100 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 1th June, 2021.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan — **Chairman**
Professor and Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Fizar Ahmed — **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

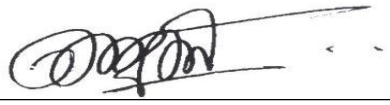
Md. Azizul Hakim — **Internal Examiner**
Senior Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision **Sheikh Abujar, 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:



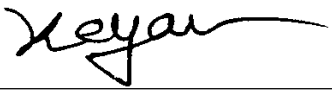
Sheikh Abujar
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Tarik Habib
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Bhaskar Majumdar
ID: 172-15-9988
Department of CSE
Daffodil International University



Mumenunnessa Keya
ID: 172-15-10100
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 **Sheikh Abujar** 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. Tarik Habib** 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.

We sincerely thank our DIU NLP and Machine Learning Research Lab for providing us with the highest facilities. And we are extremely grateful for the convenience of providing the machinery and completing the research work.

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

People can use our technology to make their lives easier. In order to keep pace with the current world, we need to automate everything. Sometime it is very difficult to get an immediate answer after asking any question. In this age of technological revolution, everything has become machine dependent. So we can't imagine doing anything without machines now. At present, the models of Deep Learning are working well for Question answering system. This is why we have done our work using deep based Sequence to Sequence model for Bengali context based QA system. Where context and question have been used as part of encoder and answers have been used as a part of decoders. The Question Answering System for NLP is becoming a potentially huge field of research. The main focus of our work is to get answer from question automatically. Question Answering system produce question to answer automatically. Like other languages, we need to expand our research on Bengali language as there is a dearth of materials needed to research Bengali language. We tried to work an automatic question answering system for the Bengali language using general knowledge dataset. For this project, running with the Bengali language turned into very hard task. However end of the day, we've create a model for computerized Question Answering System for the Bangla language.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

Question answering system can be categorized between two classes. One is closed domain and another is open domain QA system. Closed domain question answering offers with questions in specific dominion and there restrained classes expertise are popular. Open-domain question answering offers with any question and might confide on philosophy and world knowledge. Usually a question can kinds Factoid and complicated. In Factoid question answering - it is works approximately offering concise realism. It is glad with short text and announcement. It deal on questions like-who, when, where and what. Instead of, complicated question answering offer full description of the question in answer multiple line. Questionnaire prescribes computer technological know-how inside the situation of records rescue and NLP, which is set up structures that mechanically solution human generated questions in our own language.

A question respondent widget realize the proper response wherever an inquiry engines like google and yahoo area to urge higher complete documents. Answer the question manner could be a technique that relates a tool that may reply the question automatically. Respondent queries could be a manner of process flavoring Language that manage queries answer and solutions to queries are robotically generated among the method. Question respondent could be a heat topic among the gift era of NLP studies. Net assets are increasing a day that's why studies on question respondent is turning into exceptionally crucial. An automatic question respondent device congratulate on the thanks to reply to user's queries with correct answer.

For that reason an automated QA machine is obligatory. There are many people speak in Bangla in the world. In Bangladesh and the states of West Bengal, Tripura and Assam in India communicate Bangla language particularly. There are numerous research paintings has completed previously in unique languages on question Answering machine and there also exists a number of operating QAS. On this time, we've got tried to build an automated Bangla query Answering system and its miles closed domain factoid query answering system for Bengali language applying deep learning algorithms. If we examine with

Bengali and English question answering studies it'll be discovered that Bengali works is very few to English. For that reason examination on Bengali automatic query answering device is needed to be prolonged.

1.2 Motivation

Question answering system is a system where the answer to a question is automatically found. Many times we ask questions and do not get an easily answer from a question. We have to do a lot of searching to get an answer to a question. It is very time consuming and painful. Automatically answering a question has made life easier for us and made our life more dynamic and time-saving. In the case of closed domains, we can easily get answers some certain questions through this question answering system. Nowadays we see this question answering system automatically on different social media platforms and online shopping etc. But many time we fill depressed because of unsequential data and garbage meaning. So far several question answering system work has been seen in English but very few and poor accuracy rate question answering system work has been done so far for Bengali language.

Information is one of the most treasured things in this modern world. Each day a huge wide variety of question related records with answer comprised of exclusive sources. It is actually very difficult for us to remember the questions and of these answers. So in this modern age we need the method of answering the automatic question.

Bengali is our mother language. There are many people in our country who would like to use the Bengali language in their daily work as well as writing and speaking. Most of the time we have seen English language being used for question and answering systems, where both answers and questions are in English. Using these makes it very difficult for our Bengali speaking people. We want to work with the Bengali language to alleviate the suffering of the Bengali nation and the Bengali speaking people. But it matter of sorrow that, Natural Language Processing useful resource for this language isn't enough for the person. So we need to construct NLP resources and equipment and technology. That's why the primary recognition of this studies is constructing an automatic query answering device for Bengal to reach the Bengal Natural Language Processing opulence.

1.3 Rational of the study

Millions of people around the world have been proudly using Bengal Language as their native language. History of the Bengali language is very exuberant. However in this modern world, the instruments and technology of Bangla language aren't affluent like other common languages in the world. We want to enrich technology to improve this Bangla language. Numbers of the problems question answering related may be solved by way of the Natural Language Processing equipment and techniques. Earlier there are few works on Bangla QA methods were finished but and that's why we tried to make an automated query answering so as to produce the correct solution of consumer's query. For NLP Many techniques and methods have been used or builds in other languages for example English, Urdu, Chines, French, Hindi etc. However for Bengali QA system some model had been construct which isn't sufficient. Consequently, the research region of Bangla Natural Language Processing desires to be improved for query answering. The principle impediment for Bengali automatic question answering system is understanding question and gives correct answer. The system can't recognize a number of the Bengali characters and logos. Some Unicode characters and symbols are used to solve this problem. NLTK library isn't always available for Bengali textual content. For all motives, Bengali language tools are not as powerful as other languages. So research is the best way to offer an option to those styles of troubles. That why in this research activity, we effort to give question properly and send this question answer automatically for the Bengali language. That helps to saves us valuable time and easily answers questions.

1.4 Research Questions

- What is Question Answering?
- How question answering works?
- What are the benefits of Bengali question answering?
- What are the differences between Bengali and English question answering?
- How to preprocess Bengali question in NLP?
- What are the future works of Bengali question answering?
- How Bengali question answering Model works?

1.5 Expected Output

Our project is related about research. So, our main goal is to publish research papers on related projects. A lot of analysis can be done very easily through a research paper and answers too many questions are easily found. Then the developer develops the tools and makes user-friendly by using the research paper. The maximum number of question answering related studies work and equipment are advanced using closed area within the Bengali language. For everyone huge numbers of researcher and developer aren't inquisitive to offer their dataset and documents. As a consequence, a few research work is not been used. For Bangla language question answering is a brand new research topic. A few research tasks are carried out in preceding for Bangla question answering. However the end result changed into not enough first-rate for making an automated Bengali query answering. For any types of automated system we need training a machine and consequently the device needs to learn. There after the mastering version is working in the backend of a machine together with a web or cellular application. On this research, we introduce a system mastering technique for near domain using fashionable knowledge based totally Bengali question answering and display to important steps on a way to build a model for computerized Bengali query answering.

1.6 Report Layout

There are total six chapters. Different perspectives of each chapter are discussed and every chapter has several part explaining in details. This report paper contains the following contents as given:

Chapter 1

In this chapter there are some parts such like- 1.1 Discuss about Introductions, 1.2 Discuss about Motivation, 1.3 Project Rational Study, 1.4 Discuss about Research Questions, 1.5 Expected Output.

Chapter 2

In this chapter we discuss about Background Studies of the work. There are some sections like- 2.1 Introductions part, 2.2 Related works part, 2.3 Describe about Research Summary, 2.4 Highlighted Scope of the Problem, 2.5 Discuss about Challenges.

Chapter 3

In this chapter we explained about full working flow of our work and with some sections like- 3.1 Introduction, 3.2 Research Title and Apparatus, 3.3 Data collection procedure, 3.4 Data preprocessing part, 3.5 Discuss about Statistical Analysis, 3.6 Discuss about Implementation Necessity.

Chapter 4

In this chapter covers Experiment and Results of the research and some relevant discussions as- 4.1 Described about Introduction part, 4.2 Described about Experimental Results, 4.3 Described about Descriptive Analysis part.

Chapter 5

This chapter we have discussed about Social Impact on our Society, Ethical Aspects and Sustainability plan in our research work like- 5.1 Impact on Society, 5.2 Ethical Aspects, 5.3 Sustainability plan.

Chapter 6

In this chapter it consists of the belief and destiny works of the studies with some sections like- 6.1 Summary of the Study, 6.2 Conclusion, 6.3 Recommendations, 6.4 Implication for Further Study.

CHAPTER 2

Background Studies

2.1 Introduction

Question answering is a method where the questioner will ask a question and will gets that question answer automatically. Finding accurate, quickly and understandable answer is the main purpose of question answering system. Repeatedly giving and receiving some fixed question answers is annoying and difficult for a human. Therefor automatic Question answering system is a great way to help a human. For automatic question answering, machine learning approach is one of the best way to expand an automatic device. Automatic question answering facilitates us to save time and save man power which also reduces the cost.

Every day we ask a lot of questions on different sites on the internet to get answers from them. A big number of users can find or search answer shape a query in time. But saving and locating the answer from question is a complicated system and additionally need a big time to get solution. Our system will work to get automatic question answer fastly and effective way. There are two types of Question answering: Closed domain and Open domain. Closed domain question answering deals with questions mainly region and there limited classes question. Open domain query answering can offers with something.

2.2 Related Work

Question Answering is a warm topic in the present era of natural language processing. There are lots of work in research area about question answering system. Again there many ongoing working on different languages for QA method. There are a huge numbers of related task as question classification, automatically question answer, question characteristic, question taxonomies, QAS on factoid, QAS on open domain etc.

Author Somnath Banerjee et al. [1] worked on classification in question by using Bangla language. They have used syntactic, phonetic, semantic properties to classify in Bengali QA system. Again they mention single-layer classification process. For this work they have used two types of classification algorithm one is Naive Bayes and another is Decision Tree

Classifier. For Naive Bayes accuracy is 80.65% and Decision tree model accuracy is 87.63%. The Author Young Gang Cao et al. [2] worked on clinical studies interrogatory respondent system and output destination on the response of extract epitomes with complicated queries. The technique of this work method is query inquiry, summarization and answer outgiving technique. There are 5 number of categories resource they used in the system –MEDLINE abstracts system, PubMed principal complete-text articles process, medicine archives part, clinical recommendations and Wikipedia articles part.

Sourav Sarker et al. [9] complete worked for Bangla language on factoid Question Answering system by using closed domain method. They used four algorithm for their question classification: Stochastic Gradient Descent, Decision Tree, Support Vector Machine and Naive Bayes. In future, their intentions to expand the method of answering questions primarily based on complicated and narrative queries. Author G. Rohit et al. [10] worked on Question answering of NLP area by using construction of Longest Short Term Memory network to perform closed-domain question answering. Again they have compared and agreement both of the pattern and perform on selected datasets.

Mohammad Nuruzzaman et al. [5] worked with Chatbot's Question Answering system by using attention-based totally structure for sequence labelling on deep continual neural networks (DRNN). Here used to a wide range of sequence labelling tasks in different languages and domains. The attention based DRNN provides consistent enhancements and crush then ancient approaches or totally different RNN variations. Author Fenglong Ma et al. [6] used three types of datasets for Question Answering system: the real-world Stanford question answering dataset (SQuAD), the synthetic single word answer bAbI dataset and the synthetic multi word answer bAbI dataset for Question Answering system. They used two algorithms LSTM, LTMN and compared with each other. Here LTMN perform better than LSTM for this respective dataset.

The author Tasmiah Tahsin Mayeesha et al. [13] successfully applied the transformer model to Bengal Language first time. They used BERT model for Bangla Question Answering system and the use of survey experiments they compare their models with human kids to installation a benchmark score.

Discussed about concerning the subsistent power of Natural Language Processing and wonder feasible appeals of Artificial Intelligence and information salvation and Question

Answering using AI for NLP worked by author Mukul Aggarwal et al. [3]. He bring up the IQAS that support students become higher reader. For making communicate agent proxy via query answering responsibilities changed into cited and the machine perform response enquiries with the help of chaining statistics, simple fetching, taken conclusion and greater and purposed to categories the device between ability kits and stated that researchers be able discover about the cause of weakness their systems. Through using this memory Networks version Young Gang Cao et al. [4] worked by clear up the hassles.

Author Mohamed Adanyet al. [11] done an incredible job about the Holy Quran. He used Holy Quran text on their project with unparalleled domain Arabic language and used on obstructed domain queries. Again he used on Natural Language processing Question Answering system and removing stop words, diacritical and particular symbols. Author Suzan Verberne al. [12] worked on a mechanically respondent why related questions and she use 395 dataset developing approach for why-questions. Each and every queries she supply documents and one or a lot of user formulated answers square measure obtainable among it. She manufactured query analysis in the sake of the question methodology supported syntactical classification and response kind of realization. She targeted upon the important problems with wherever, once and what. She focused of this analysis paper is to search out whether or not the employment of factoid question isn't appropriate for respondent the question.

Chandra Obula Reddy et al. [7] worked for supported survey of the kinds of query answering. For this survey, they described several types of question answering and architecture of proposed like- Open dominion question answering process, Closed dominion question answering process, Factoid type, Catalog type, Verification type Questions, Causal type problem, Suppositional exploration type etc. A large number of works successfully done by factoid QAS. Radu Soricu et al. [8] worked above factoid Question Answering system. For this work they collect 1 million data from the web page and create a Frequently Asked Questions (noisy-channel) based architecture.

For our work we have used Bangla General Knowledge related dataset for question and response. We have inside our datasets mixed by International and Bangladesh related question with answers. We worked closed domain dataset and use sequence to sequence learning for making Question Answering system.

2.3 Research Summary

In this project, we founded a method generated in the sake of Bangla Question Answering. Using deep learning we create a model. Our own Bangla dataset we used to make this architectural model. We assembled 2000 Question with answer from Social media, Internet. The full dataset create based on Bangla General Knowledge by using International and Bangladesh related questions. We had to create three column for the dataset, First column Question part, Second column context part and third column answer part. We pre-processed our dataset because it is important to process the dataset before running any research work model. We encode the data and encoder alter the data from one ordination to another. At first encoder split text the data and also tokenization, then remove stop words from the data. Tokenization is a method that replaces a clean standard with a form of data protection like- strings in words, symbols, phrases, keywords and randomly generated synthetic values. After preprocessing we have used pad sequences accustomed make ensure that each one the orders or series have equivalent range of the list. With the aid of default it is done via padding zero to the distance of every order to a comparable quantity because the longest order of each series. Next, we use LSTM model for Encode and Decode data with neural network schematized by sight sequence to sequence methods. After using all process and library we tend to train the model for over 4 hours. Then the machine gave us a good result.

2.4 Scope of the problem

For our work we used sequence-to-sequence (seq2seq) model. The seq2seq model is generally composed of an encoder-decoder architecture, wherever the encoder approaches the input sequence and encodes or compresses the facts into a context vector of fixed length. For short length sequence of context in machine translation seq2seq model provide a good result. Question answering system is modern invention for studies in Bangla NLP. This research paintings used seq2seq model for Bangla Question answering system based totally on Bengali general knowledge question. In our dataset question context is not huge but right sufficient for question to answering however the context question is the small range. In this project intention we used sequence-to-sequence version with LSTM methods. So short question are reaction a proper outcome for question answering. As a result our set

of rules also reply for a short term questions answer. The seq2seq model is generally not able to appropriately technique for long enter sequences, considering best the closing hidden state of the encoder RNN is used as the context vector for the decoder. So, long term question answering this model can't give accurate output.

2.5 Challenges

Constructed Bangla data for Question answering is not obtainable easily. So, for this work data collection was very challenging step. Some data we have to collect from Facebook but it was a very lengthy process. We collect one question then we collect that's question answer and we check is it right or wrong. In addition to Facebook, we have created a new data set from Google one by one with general knowledge based questions, this task was very difficult. Any work with Bangla is very difficult because we do not have much resources for Bangla. It has NLTK library which has made the data processing work related to Bengali a little easier. Consequently, for data preprocessing step we necessity to fresh coding to put together the facts as enter of a version. If we have switch punctuation from the data, we want Unicode of each punctuation and receive away it by using fresh coding. Another drawback is stop word take away from the data. Others languages like English have a build-in library to get rid of stop phrases or words from data. For the Bengali language accumulate stop phrase or words from online then place data in text report and take away stop phrases or words from data using that document. For Bangla language if we want good accuracy we need large number of dataset. If we provide a large number of dataset we can get a good accuracy for our work.

CHAPTER 3
Research Methodology

3.1 Introduction

Every research work has a specific methodology, our work also has a specific methodology. Now we will describe in details the whole process of our work. The goal of the research work is to discover something new using some unique solving technique. The application of all these includes methodology part. We will describe each parts of the model applied to our work. For this research work, we have used deep learning based Long Short-Term Memory model for question answering. Research has become easier with the inclusion of the use of deep learning algorithms for the Bengali language. Long Short-Term Memory (LSTM) is exploited for the question answer type data processing hassle in deep learning knowledge. Before we run the deep learning model we need to know what we are working on and why. Again, we needs a proper dataset to find an accurate result. It is very important to make and preprocess datasets well before running any algorithm. In this section describes how the different parts of our approach work very well. A proper narration and rationalization of method enhancement the ability for working and supply the aristocracy. Graphical sketch and mathematical equation of the method with their representation is supporting to apprehend the entire pursuit. To work in the future requires a thorough understanding of the methodology is very important part. If mathematical term and graphical waves are fully functional then anyone looking at work process can get a lot of ideas easily.

Again, for good research paper needs a good working flow chart. Anyone can looking at a transparent workflow chart is very easy and will be able to know about the models in a short time. Anyone looking at a transparent workflow chart is very easy and will be able to learn about the models in a very short time. Every footpace of the method are described on this section. Here Sub phase of a few middle sections are facilitates to recognize the gist of the model with it cause of the use of that's will very helpful to understanding the methodology part. The full work process is given below which will serve as a summary of the entire research work-

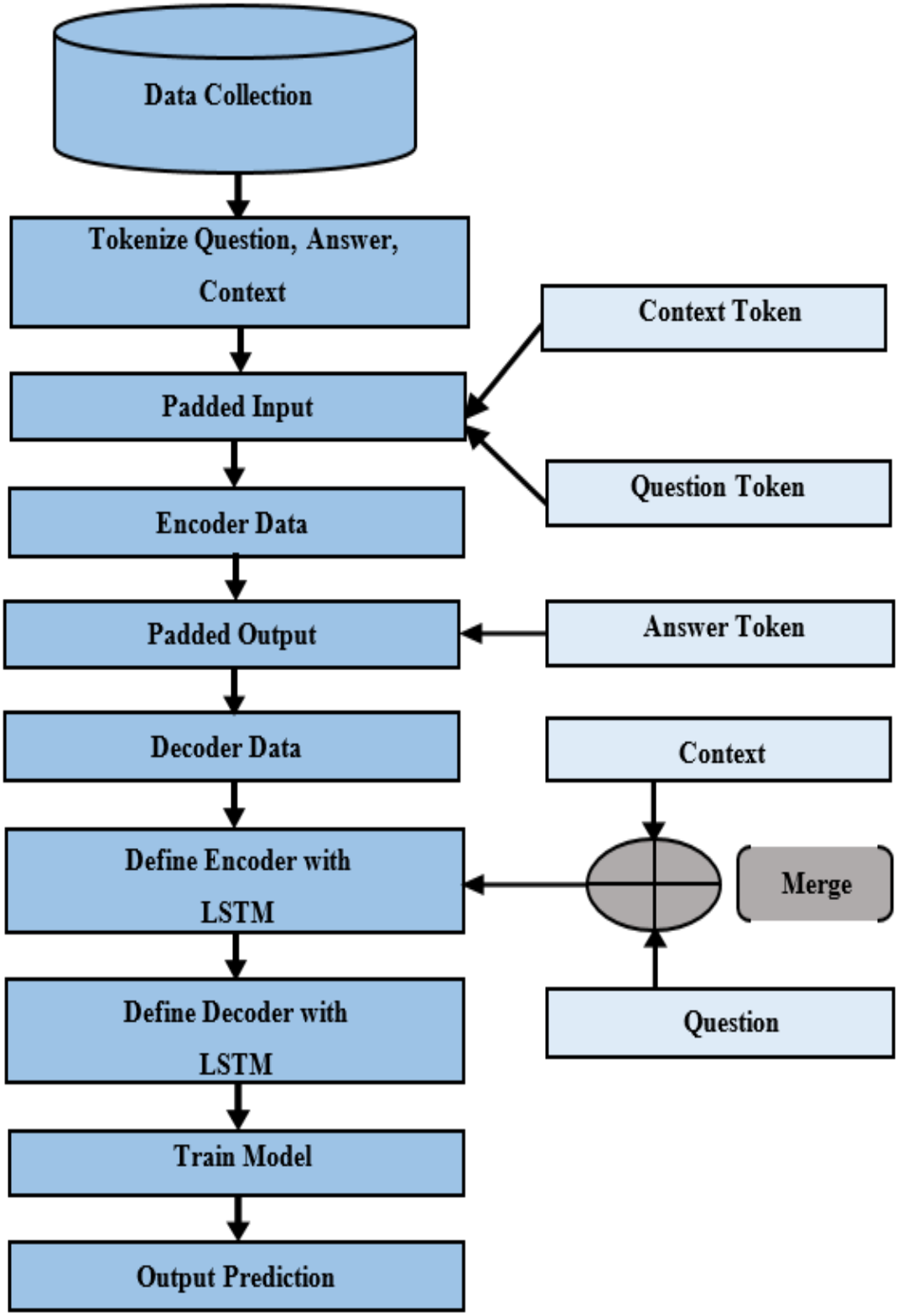


Figure 3.1.1 Work process for Question Answering System.

3.2 Research Title and Apparatus

We propose our thesis title is “QAGK: Question Answering Based on Bangla General Knowledge Data”. Now a days QA is the hottest topic in Bangla NLP. We described on the answer to the question of closed domain in Bangla general knowledge founded with the mathematical and theoretical way in first. For running a deep learning model requires a high configuration PC and GPU and others apparatus moreover it will be very difficult for us to working. A list is given below of the important equipment and technology to get our model running properly-

- Software and Hardware:
  - Intel Core i5 8th generation with 4GB RAM
  - 1TB HDD
  - Google Colab with GPU
- Development Tools:
  - Windows 10
  - Python 3.7
  - 1.15.2 version of Tensorflow Backend
  - Numpy
  - NLTK
  - Pandas.

3.3 Data collection

We took our own dataset collected for research work. The biggest problem when it comes to working for Bengal language is collecting data. Data has been collected from Google and social media from Facebook. A lot of data is needed for a study to get good accuracy. The more data we can collect, the better accuracy the algorithm models will give us. We collected 2000 datasets on a general knowledge based on national and international basis mixed. Since there is no such thing as easy data collection anywhere in Bengal, we have faced those problems in data collection. One by one we have taken the initiative of general knowledge based questions for Bengali and have taken the answers along with the questions. After collecting the questions and answers and checking whether they are correct

then we replaced them with CSV format. We divided our dataset into three columns- questions part, context part and answers part. Below is a short picture of our own collected dataset in a tubular form-

Table 1: Sample of our dataset

Question	Context	Answer
বাংলাদেশের সাংবিধানিক নাম কী?	বাংলাদেশের সাংবিধানিক নাম গণপ্রজাতন্ত্রী বাংলাদেশ	গণপ্রজাতন্ত্রী বাংলাদেশ
বাংলা ভাষাকে দেশের দ্বিতীয় ভাষার মর্যাদা দিয়েছে কোন দেশ?	বাংলা ভাষাকে দেশের দ্বিতীয় ভাষার মর্যাদা দিয়েছে সিয়েরা লিয়ন	সিয়েরা লিয়ন
প্রথম বাংলাদেশি কোন ক্রীড়াবিদ অলিম্পিকে সরাসরি অংশ নেওয়ার যোগ্যাতা অর্জন করে?	প্রথম বাংলাদেশি গলফার সিদ্দিকুর রহমান অলিম্পিকে সরাসরি অংশ নেওয়ার যোগ্যাতা অর্জন করেন	গলফার সিদ্দিকুর রহমান
পূর্ব পাকিস্তানের নাম "বাংলাদেশ" করা হয় কবে?	পূর্ব পাকিস্তানের নাম "বাংলাদেশ" করা হয় ৫ ডিসেম্বর ১৯৬৯ সালে	৫ ডিসেম্বর ১৯৬৯ সালে
পারমানবিক চুল্লীতে পরিবাহক হিসেবে কোন ধাতু ব্যবহৃত হয় ?	পারমানবিক চুল্লীতে পরিবাহক হিসেবে সোডিয়াম ধাতু ব্যবহৃত হয়	সোডিয়াম ধাতু
‘অসমাপ্ত আত্মজীবনী’ বইটির লেখকের নাম কী?	‘অসমাপ্ত আত্মজীবনী’ বইটির লেখকের নাম বঙ্গবন্ধু শেখ মুজিবুর রহমান	বঙ্গবন্ধু শেখ মুজিবুর রহমান
দেশের প্রথম স্যাটেলাইট বঙ্গবন্ধু ১ মহাকাশে পাঠানো হয়েছে কবে?	দেশের প্রথম স্যাটেলাইট বঙ্গবন্ধু ১ মহাকাশে পাঠানো হয়েছে ১২ মে ২০১৮	১২ মে ২০১৮
করোনা ভাইরাস যখন ধাতব কোনো পৃষ্ঠের উপর পড়ে তখন এটি কত সময় পর্যন্ত বাঁচতে পারে?	করোনা ভাইরাস যখন ধাতব কোনো পৃষ্ঠের উপর পড়ে তখন এটি ১২ ঘন্টা পর্যন্ত বাঁচতে পারে	১২ ঘন্টা পর্যন্ত
কম্পিউটারের প্রথম প্রোগ্রামিং ভাষার নাম কি?	কম্পিউটারের প্রথম প্রোগ্রামিং ভাষার নাম ফোরট্রান	ফোরট্রান
বাংলায় ইউরোপীয় বণিকদের মধ্যে বাণিজ্যের জন্য প্রথম এসেছিলেন কারা?	বাংলায় ইউরোপীয় বণিকদের মধ্যে বাণিজ্যের জন্য প্রথম এসেছিলেন পর্তুগিজরা	পর্তুগিজরা

3.4 Data preprocessing

After collection datasets we need preprocess our all data. We have to use few steps for data processing. First we converted our data format using an encoder methods. Then we check split text the data. Then we split the data as well as tokenized the datasets. For data preprocessing we have to clean our data and used pad sequence. Then we merge context and question data for processing our datasets.

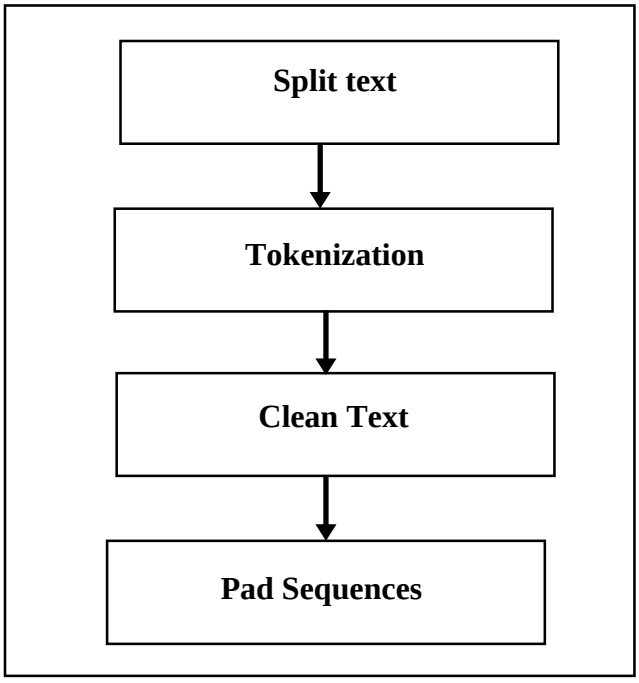


Figure 3.4.1 Dataset preprocessing

3.4.a Split text

Very few people fully understand how texts in sentences can work in a complex way. Already most NLP frameworks have developed English models for this purpose. If ever a large paragraph of text is given, the best way to analyze it is to divide the text into several sentences. In real-life conversations, we calculate information by dividing the sentences into different levels by analyzing the words in the consonant. However, in the case of text paragraphs, trying to split sentences correctly can be difficult in raw code, which fortunately can be done quite easily with the help of NLT.

3.4.b Tokenization

Tokenization is a finely crafted method that replaces a clean standard with a form of data protection and randomly generated synthetic values. Which usually stands for ‘token’ which means more meaningful testing including changing downstream applications, enhanced data sharing and secure data. The first step in understanding tokenization is to use more sophisticated and innovative methods to gain access and control of data to outsiders in the modern data landscape. The second step is that childhood and walls will act as a good barrier, but when the boundaries of the enterprise become more fragmented with the adoption of SaaS, Cloud and third party data processing, the solution is to secure the data manually using fine-grained data protection. Change a clear text value like 'Bob' to 'XYZ'. Subsequently using ‘XYZ’ storage, transmission and processing instead of sensitive clear text ‘Bob’. Tokenization requires some random mapping between the original value and the resulting protected value. The reason for this randomness is that it saves encryption a more secure way of protecting data than saving formats. Regulations like DSS in PCI do not require strong sensitive credit card data to be recreated by tokenizing, but the data needs to be encrypted if this rotation is required.

3.4.c Clean Text

We cannot apply machine or deep learning models directly from any raw text. We must first clear the text, which means we have to split the word through this process and handle the punctuation and case very well. Clean text often means a list of words or tokens that we work on in machine learning models. This means the strategy of converting the raw text into a list of words. The easiest way to do this is to split new lines, tabs, and many more documents into white space.

3.4.d Pad Sequences

Pad sequences serve to assign equal lengths to the list of all sequences. For each sequence it is performed by 0 padding by default, here the same length as the longest sequence of each sequence is considered. We have used the keras pad sequences function for this long sequence related work. Pad sequences can be represented as like- ([[‘জসীমউদদীনের’,’ছদ্মনাম’,’কি’], [‘ম্যানগ্রোভ’, ‘কি’]]). From here we can see that our

sequential form has worked for 2D matrix and like the 1st row of the array as [1, 2, 3] and 2nd one as [4, 5]. Here we can see that [1,2,3] is the longest sequence for this two matrix, so 0 will be padded to the 2nd sequence and for the 2nd matrix the matching length will be [0,4,5]. This pad sequence is used to bring each data to an equal length.

3.5 Statistical Analysis

In this work we collected 2000 datasets on a general knowledge based questions. There are two encoder part question and context, decoder part is answer. For questions input maximum length is 15 and number of Input tokens is 3738. Again for context part input maximum length is 16 and number of Input tokens for context is 5318. Answer output maximum length is 10 and number of output tokens is 2355. Here for Training we use on 1600 samples and for validation we use on 400 samples. We used Microsoft excel file is used for save the dataset.

3.6 Implementation Necessity

There are many models in deep learning and different types of model used for various types of motive. While we are operating with question and answer, longest short term memory (LSTM) are going to be very useful for question text modeling. A machine translation is necessary for learning machine concerning about question sequence. For our Question Answering System we have used Sequence to Sequence model. A Sequence to Sequence method is a method that use for change an input order or series into an output order.

3.6.a. Logical Equation

For our QA system we perform analysis puzzle of mathematically by using feasibility equation. Let, the question set as feasibility of P(Q) and the response feasibility of P(A) . So, we consider the user's questioners as feasibility of P(B).

The equation will be –

$$P(Q \text{ and } B) = P(Q).\ P(B|Q) \quad \text{.....(1)}$$

And the response will be,

$$P(A) = P(Q).\ P(B|Q) \quad \text{..... (2)}$$

Cause, $P(Q \text{ and } B) = P(A)$

So, from the feasibility equalization we can analyse the mathematical or logical problem of our work.

3.6.b. Sequence to Sequence Learning

The Sequence to Sequence model works for machine translation, text summaries, question answering, language systems, responsive queries, catbots etc. Sequence to Sequence models is a particular form of Recurrent Neural Network architectures. Sequence-to-sequence learning (Seq2Seq) models convert sequences from one domain to sequences another domain. It could be used for machine translation or free-from question respondent generally. There are a unit multiple ways in which to handle this task, either using RNNs or 1D convents. Here we are going to target RNNs. Sequence to sequence model have two parts: one is encoder part and another is decoder part. Encoder-decoders each contain LSTM blocks. 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 one-hot vector tokens is also define by x_i and y_j. The vocabulary of the inputs and outputs are define by $V^{(s)}$ and $V^{(t)}$. All the important component of x_i and y_j part $x_i \in R^{|V^{(s)}|}$ and $y_j \in R^{|V^{(t)}|}$. Now the equation for X and Y is,

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

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

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 $P_\theta\ (\ y_j \mid Y_{<j}, X\ )$,

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

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

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

$$z = \Lambda\ (X) \quad \ldots\ldots (4)$$

Here, **Λ** = Function of the recurrent neural network of LSTM methods.

Now,

$$P_\theta\ (\ y_j \mid Y_{<j}, X\ ) = \boldsymbol{\gamma}\ (\ h_j^{(t)}, y_j\ ) \quad \ldots\ldots (5)$$

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

Here, Ψ = hidden vector of $h_j^{(t)}$, $\boldsymbol{\gamma}$ = 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,

$$\bar{x}_I = E^{(S)} x_i \quad \ldots\ldots (7)$$

In encoder the embedding matrix is $E^{(S)} \epsilon R^{D \times |V^{(8)}|}$. The encoder recurrent layer generate hidden vectors. $\Psi^{(S)}$ is the uni-directional Recurrent Neural Network (RNN) function. So equation is,

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

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

Here, activation function is $\tan h$.

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

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

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

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

So, $h_0^{(t)} = z = h_l^{(S)}$

The decoder output layer equation is below,

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

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

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.

Now the architecture of this model is given below,

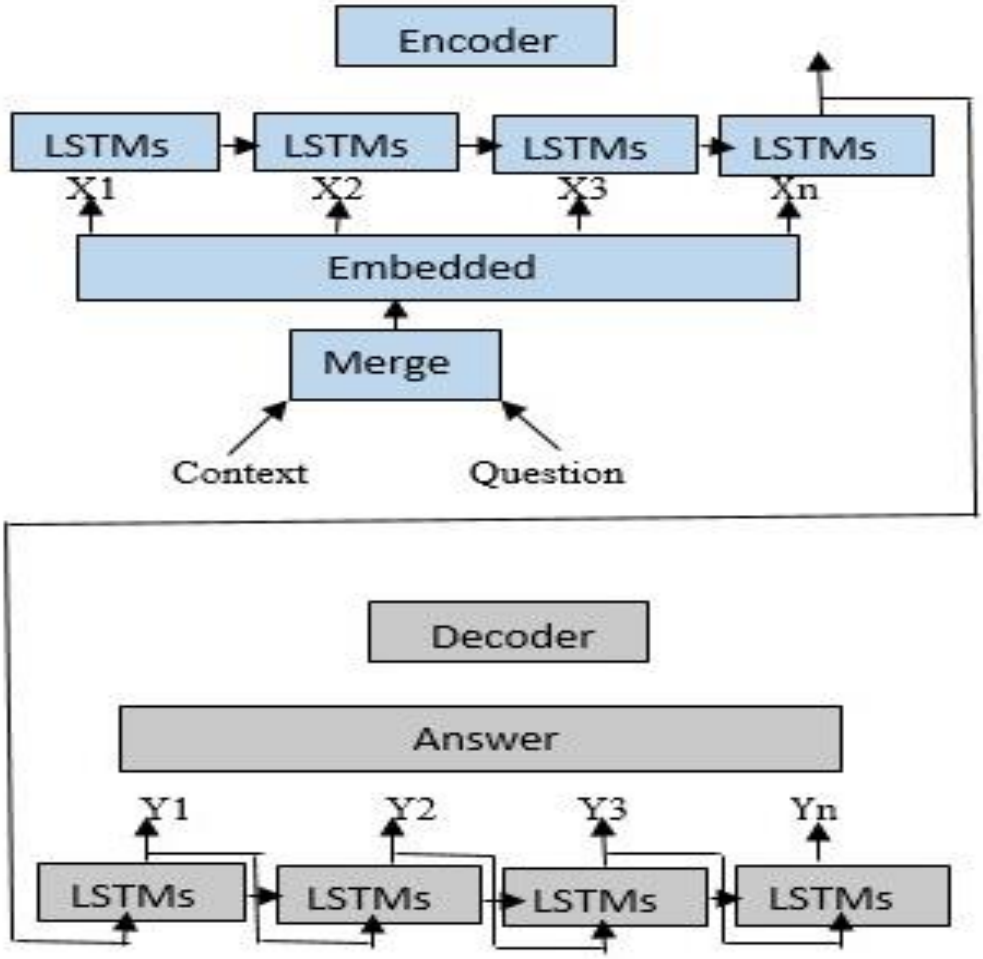


Figure 3.6.1 Architecture of Seq2seq model.

3.6.c. Activation Function

For activation function model we tend to use activation operate of softmax. It mention outcome with reference to neural network and confirm it operate would actuate or not.

The softmax activation function count the primary pursuit or logistical activation. For this operation the outcome range within 0 and 1. So the equation of softmax activation function is,

$$f_{(Z)_j} = \frac{e_j^{(z)}}{\sum_{K=1}^{K} e_k^{(z)}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (12)$$

*H*ere, z = performance output, j = record output.

3.6.d. Longest Short Term Memory 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,

$$f_t = \sigma\ (\ W_f[h_{t-1}, x_t] + b_f) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (13)$$

$$C_t = f_t * C_{t-1} + i_t * \sigma\ (\ W_c[h_{t-1}, x_c] + b_c\ ) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (14)$$

$$O_t = \sigma\ (\ W_o[h_{t-1}, x_t] + b_0) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (15)$$

$$i_t = \sigma\ (\ W_i[h_{t-1}, x_t] + b_i) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (16)$$

$$h_t = O_t * \sigma\ (C_t) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (17)$$

LSTM is explain as protect a memory cell = C_t which is reset, write and read from following to the forget gate = f_t , the input gate = i_t , the output gate = O_t , hidden state = h_t and here σ is activation function.

CHAPTER 4

Experimental Results and Discussion

4.1 Introduction

Automatic Question Answering is very hard and complicated problem in Bangla Natural Language Processing. Finding answers to the right questions from a machine is very difficult task. If you ask the machine, he can give an answer, but that's not the main thing. The main task of the machine is to automatically give the correct answer after receiving any question. For this question answering system probability is very important part. Because a machine provides its correct output depending on the maximum probability. After every word is trained by machine then calculate the probability and give us the related question answer. Each deep learning model is driven by a machine or engine.

For this project we used Seq2Seq model in 1.15.2 version of Tensor flow for backend. Whereas a machine receives some question then it will be model training by a model and when model stop training that machine will provide automatically answer that question. The datasets provides answer randomly and our machine collecting accurate answers for the given questions. A dataset training requires some basic parameters like- epochs, batch size, validation split, verbose etc. For this project we applied attention mechanism technique.

In this work we place batch size = 32. The number of input samples sent to the network is determined by the batch size. Another factor that affects classification accuracy is batch size. The larger the batch size, the longer it takes to train the dataset, and the models precision suffers as a result, as does the memory requirement. As a result, when deciding on batch size, we must proceed with caution. Now another parameter epochs = 50, the number of iterations is measured in epochs. The number of epochs is a hyper parameter that specifies how many times the learning algorithm can process the entire training dataset. Once per epoch, each sample in the training dataset had the chance to update the internal model parameters. The value of another parameters are- define verbose = 1, define validation split = 0.15, Embedded Units = 100, Hidden Units = 256 and we used Adam optimizer for minimize the optimization and loss the model. After a good model train it

may be able send good accuracy. For train period Adam optimizer is very important because it's calculate every parameters. A good configuration pc or laptop is very important for deep learning algorithm model train. GPU is must needed for a dataset training..

At first we train our model in direct pc. As a result it take a long time to run the model and the results obtained are not good for question answering. So, this work we train our model using google colab. It provides cost free GPU service for the user.

4.2 Experimental Results

We know that no machine can give 100% accurate output. The model we used for our work gave good output but could not always give 100% accurate output. In a very short time it occasionally gave some wrong output. This wrong output is negligible when working with Bangla NLP.

Now the graphical view of our model after testing the accuracy-

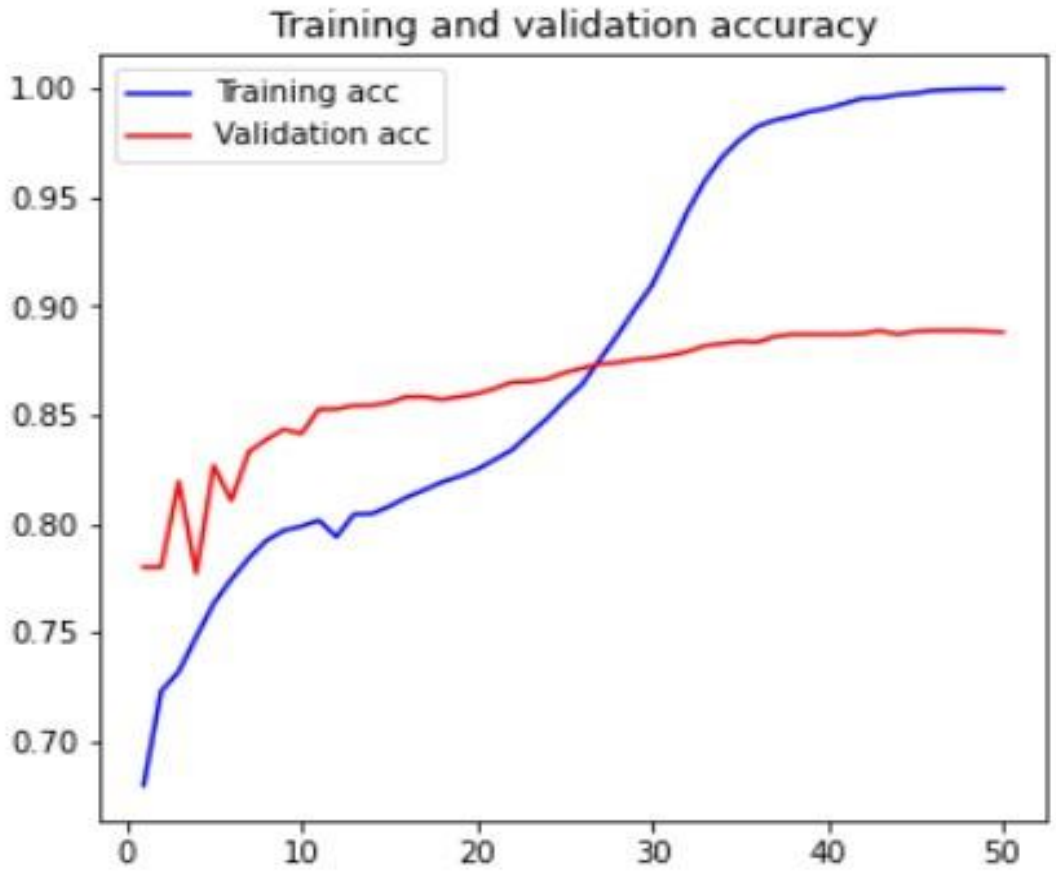


Figure 4.2.1 Training and validation accuracy.

For this graph (Figure 4.2.1), we can see that our training model accuracy and validation accuracy is good for using Bangla Question Answering system. Our train model accuracy rate is 0.99 and validation accuracy rate is 0.89. That's all accuracy is very good by using Bangla QA dataset. Many of the works we have reviewed before have not been able to get such good accuracy using Bangla datasets.

Now the graphical view of our model for the train and validation loss-

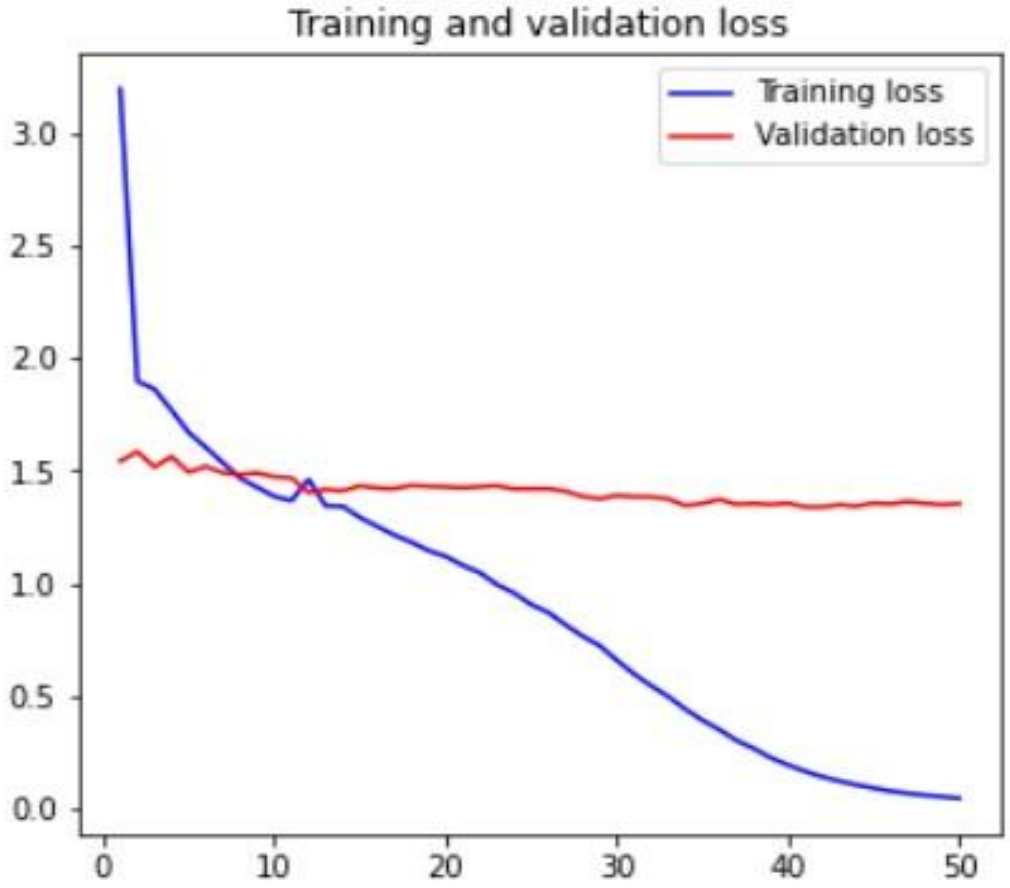


Figure 4.2.2 Training and validation loss.

Now we can see (Figure 4.2.2) that for our model training accuracy loss and the Validation loss is very few. We store our model file and rename the file name is "model.ckpt" file for check our output. Then we create a tensor flow session to reload the graph and saved in the previous step. To reduce the loss we used cross entropy in categorical which is used in any classification work for a single level.

Now the response of the model after training, there are three sample outcome of our research is given under in table 2, table 3 and table 4. Every table context and question part is raw data which was collected from Google and Facebook. The feedback of the model obtained by the machine after training is given below-

Table 2: Sample response one for Bengali Question Answering

Context	ইউনেস্কোর ৩১তম সম্মেলনে একুশে ফেব্রুয়ারিকে আন্তর্জাতিক মাতৃভাষা দিবস হিসেবে ঘোষণা করা হয়
Question	ইউনেস্কোর কততম সম্মেলনে একুশে ফেব্রুয়ারিকে আন্তর্জাতিক মাতৃভাষা দিবস হিসেবে ঘোষণা করা হয়?
Answer	৩১তম
Response	৩১তম

Table 3: Sample response two for Bengali Question Answering

Context	পূর্ব পাকিস্তান আওয়ামী মুসলিম লীগ প্রতিষ্ঠা লাভ করলে শেখ মুজিবুর রহমান সেখানে যুগ্ম সম্পাদক পদ পান
Question	পূর্ব পাকিস্তান আওয়ামী মুসলিম লীগ প্রতিষ্ঠা লাভ করলে শেখ মুজিবুর রহমান সেখানে কী পদ পান?
Answer	যুগ্ম সম্পাদক পদ
Response	যুগ্ম সম্পাদক পদ

Table 4: Sample response three for Bengali Question Answering

Context	বাংলাদেশের স্বাধীনতার স্থপতি তথা জাতির জনকের নাম বঙ্গবন্ধু শেখ মুজিবুর রহমান
Question	বাংলাদেশের স্বাধীনতার স্থপতি তথা জাতির জনকের নাম কী?
Answer	বঙ্গবন্ধু শেখ মুজিবুর রহমান
Response	বঙ্গবন্ধু শেখ মুজিবুর রহমান

4.3 Descriptive Analysis

We created the Bangla Question Answering Model. In previously we were well conscious of how the Question Answering model works for English. However, no model has been used for Question Answering in Bengal before. We've seen some work on the Question Answering System using English before and learned about the accuracy of the model in those works. Accuracy rate was much better for English. So we were worried about whether the accuracy rate would be better using the same model in Bengali. However, using the same model we can see the results of our work that the accuracy rate in Bengali is very good like in English. Question answering system model is construct to minimize the loss operation in the model. During data training we calculate all types of loss functions and we wait until all iteration to calculate the final loss function. Before the model training, we divided the data into test and train.

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 question answering. Now the present time is the age of information technology, now in our society information technology has touched all fields. Chatbot or Automatic question answering System is currently saw and used in many languages, especially for English language. However, such work for the Bengali language has not been read or used in our eyes in that way. We have run a model to do this kind of work in Bengali and the model has given very good results. The main goal of this exploration is to expand the use and innovation of machines in Bengali language. Such research will play a very important role in our society as well as for the Bengali speaking people. Using this kind of research, we can 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 help us to take Bangla language to a higher level.

This work will enrich our Bengali 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 General Knowledge 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 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 Bengali 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 Bengali 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 Bangla 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 Bengali 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

For this project, our whole work is connected to Bangla Natural Language Processing. We have done our work for Bangla Question Answering using Deep learning model. Our work is very helpful for provide an automatic Bengali question answering. No work has been done for Bangla question answering using general knowledge based dataset and deep learning model before. After solving the problem of question answering using deep learning model in English language, we got very good accuracy for Bengali just like we saw the accuracy. Our model will helpful for Bangla NLP research area and it will be helpful in future Bengali question answering related all works. Our work is to enrich the Bengali NLP world. It took us four months from data collection to full work completion. We had to go through many steps one by one to get this done. We have now summarized the whole work step is below-

Step 1: Question answering dataset collected form Google and Facebook.

Step 2: Collected answer for questions.

Step 3: Store all data in Excel .csv file.

Step 4: Preprocessing dataset.

Step 5: Vocabulary calculation.

Step 6: Use Pad sequence for data sequence length.

Step 7: Merge Question and Context data.

Step 8: Use Encoder and Decoder with LSTM.

Step 9: Create sequence to sequence model.

Step 10: Model train and test.

Step 11: Output or result check.

By following these all steps sequentially we completed our task. In this model will support our Bangla Natural Language Processing research area to make a complete automatic question answering system and further experiment of Bengali question answering related any work.

6.2 Conclusion

Our principal objective of this project is to enrich and strengthen NLP research in Bengali language. We have used Bengali general knowledge based question data as the input of our model and taking the input the model works with the question which will provide our specific answer as output. For this automatic answering system we used encoding and decoding with LSTM process. Before we did question answering related work on Bengali, we saw some work in English language. From there, we are inspired to work on Bengali language. Our dataset is not big for Bengali question answering system. We used only 2000 data for this work. But even after using so few data the machine gave good accuracy for our model. We hope that this work will play an important role in all future work related to automatic Bangla question answering system. We did our work with the factory data of the close domain so our project works for fixed dataset. Our model will not work if we want answers to any questions. This is the main limitation of our project. Data processing in Bengali is a bit more difficult than in English. There are many libraries in English for processing but not in Bengali. So we need to create preprocessing library for Bengali language. Our model has given a very good output despite having so many problems working with Bengali language. After working with Bangla Dataset, our model gave 99 percent accuracy for prediction result. No machine can work with 100% success. Our question answering system model also has some limitations. But above all for Bangla we have to admit that this model will play very important role. In this work is a contribution for our Bangla NLP that will be conducive for future project or research work on Bangla.

6.3 Recommendations

Since Bengali is a complex language compared to other languages, we have to work hard to work with Bengali. If we want to work with Bengali language then we need clear and large dataset. In order to work with Bangla language, we need to have a good PC configuration and a complete GPU system. Dataset must be kept in mind before working with Bangla as it is very difficult to get enough datasets to work with Bengali. In order to get better results from this work, we can increase the amount of dataset so that the accuracy of the models will be better in future. Nowadays we can see the importance and impact of Automatic question answering on various online platforms and we can say that this work

will have a significant impact in the future by targeting Bengali language. Repeatedly answering this question is a matter of annoyance and it is a matter of time. If we can train the machine well keeping in mind the specific questions, then the machine will give us a good output using the right model, which will reduce both our time and effort.

6.4 Implication for Further Study

We noticed some limitations in doing this, such as we dealt with certain questions about closed domains and our dataset is not enough. We know any types of model is constructed for incoming improvement. Because any kinds of experimental work is an incessant method and day by day it's goes to better place. As a result, our model for Bengali language will be further developed day by day. We have just started this research on questions and answers about Bengali language and then we have to go much further with this research. Next time we will attempt to create complicated and narrative QA system with open domain and we will increase our own datasets. We need to expand the model after completing this study. We have used only one model in this work. In the future, we want to enrich the dataset and run some more new models. This will allow us to understand which model we would like to use for this task. We want to work on how to apply it in our lives after the research is done. Because if any research work is not useful for human beings, then it can be said that it is not important. Building web-based and mobile applications based on the future of artificial intelligence will make this research a reality. So our work will help to create an application for automatic Bengali question answering system.

REFERENCES

[1] Somnath Banerjee, Sivaji Bandyopadhyay: Bengali Question Classification, "Towards Developing QA System." (Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 25–40, COLING 2012, Mumbai, December 2012.)

[2] YoungGang Cao [a,1] , Feifan Lia [a], Pippa Simpson [b], Lamont Antieau [a], Andrew Bennett [c,d J. James Cimino [e], John Ely [f], Hong Yu [a,g,*]: AskHERMES, "An online question answering system for complex clinical questions." (Journal of Biomedical Informatics 44 (2011),277–288).

[3] Mukul Aggarwal, "Information Retrieval and Question Answering NLP Approach: An Artificial Intelligence Application." (NCAI2011, 13-14 May 2011, Jaipur, India, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-NCAI2011, June 2011)

[4] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merri¨enboer, Armand Joulin & Tomas Mikolov, "TOWARDS AI-COMPLETE QUESTION ANSWERING: A SET OF PREREQUISITE TOY TASKS." (ICLR 2016).

[5] Mohammad Nuruzzaman, Omar Khadeer Hussain, "Identifying Facts for Chatbot's Question Answering via Sequence labelling Using Recurrent Neural Networks." (ACM TURC 2019: Proceedings of the ACM Turing Celebration Conference - China, May 2019, Article No.: 93 Pages 1–7, https://doi.org/10.1145/3321408.3322626).

[6] Fenglong Ma, Radha Chitta, Saurabh Kataria, Jing Zhou, Palghat Ramesh, Tong Sun, Jing Gao**, "**Long-Term Memory Networks for Question Answering." (Proceedings of IJCAI Workshop on Semantic Machine Learning (SML 2017), Aug 19-25 2017, Melbourne, Australia.)

[7] Chandra Obula Reddy [1], Dr. K. Madhavi [2], "A survey on Types of Question Answering system." (IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, ISSN: 2278-8727, Volume 19, Issue 6, Ver. IV (Nov.- Dec. 2017), PP 19-23).

[8] Radu Soricu and Eric Bril, "Automatic Question Answering using the web Beyond the Factoid." (Journal of Information Retrieval - Special Issue on Web Information Retrieval, 2006, Kluwer Academic Publishers).

[9] Sourav Sarkar, Syeda Tamanna Alam Monisha, Md Mahadi Hasan Nahid, "Bengali Question Answering System for Factoid Questions: A statistical approach." (International Conference on Bangla Speech and Language Processing (ICBSLP), 27-28 September, 2019).

[10] G. Rohit, Ekta Gautam Dharamshi and Natarajan Subramanyam, "Approaches to Question Answering Using LSTM and Memory Networks." (Springer Nature Singapore Pte Ltd. 2019).

[11] Mohamed Adany Hamdelsayed Adany "An automatic question answering system for the Arabic Quran." (August 2017)

[12] Suzan Verberne "Developing an approach for why question answering." (EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy).

[13] Tasmiah Tahsin Mayeesha, Abdullah Md Sarwar and Rashedur M. Rahman (2020), "Deep learning based question answering system in Bengali." (Journal of Information and Telecommunication, DOI: 10.1080/24751839.2020.1833136)

[14] Md. Kowsher, M M Mahabubur Rahman, Sk Shohorab Ahmed and Nusrat Jahan Prottasha, "Bangla Intelligence Question Answering System Based on Mathematics and Statistics." (2019 22nd International Conference on Computer and Information Technology (ICCIT), 18-20 December 2019, DOI: 10.1109/ICCIT48885.2019.9038332).

[15] Abu Kaisar Mohammad Masum, Sheikh Abujar, Md Ashraful Islam Talukder, AKM Shahariar Azad Rabby and Syed Akhter Hossain, "Abstractive method of text summarization with sequence to sequence RNNs." (10th ICCCNT - 2019 July 6-8, 2019, IIT - Kanpur Kanpur, India).

[16] Sheikh Abujar, Abu Kaisar Mohammad Masum, S. M. Mazharul HaquenChowdhury, Mahmudul Hasan and Syed Akhter Hossain, "Bengali Text generation Using Bidirectional RNN." (10th ICCCNT - 2019 July 6-8, 2019,IIT - Kanpur Kanpur, India).

[17] Zhang, Han et. al., "A Retrieval-Based Matching Approach to Open Domain Knowledge-Based Question Answering." Conference in Natural Language Processing and Chinese Computing. Springer, pp. 701-711, (2018).

Bangla Question Answering Based Model Using Sequence to Sequence Learning

ORIGINALITY REPORT

9%	6%	5%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	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 Publication	3%
3	Md Ashraful Islam Talukder, Sheikh Abujar, Abu Kaisar Mohammad Masum, Fahad Faisal, Syed Akhter Hossain. "Bengali abstractive text summarization using sequence to sequence RNNs", 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019 Publication	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		