Jiggasha: A Bengali Question Answering System using Fine-Tuned BERT

BY

Md. Rafiuzzaman Bhuiyan
ID: 172-15-9655
AND

Md. Abdullahil-Oaphy
ID: 172-15-9868

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. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

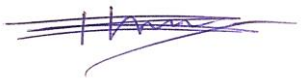
JUNE 2021

APPROVAL

This Project/internship titled **"Jiggasha: A Bengali Question Answering System using Fine-Tuned BERT",** submitted by Md. Rafiuzzaman Bhuiyan, ID No: 172-15-9655 and Md. Abdullahil-Oaphy, ID No: 172-15-9868 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 01.06.2021.

BOARD OF EXAMINERS

Chairman



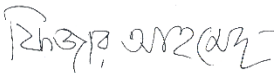
Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

Dr. Fizar Ahmed

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Internal Examiner

Md. Azizul Hakim

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin

Professor

Department of Computer Science and Engineering

Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **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 or diploma.

Supervised by:

Sheikh Abujar
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

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

Submitted by:

Md. Rafiuzzaman Bhuiyan
ID: -172-15-9655
Department of CSE
Daffodil International University

Md. Abdullahil-Oaphy
ID: -172-15-9868
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Sheikh Abujar**, **Senior Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*AI and NLP*" to carry out this project. His endless patience, scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Almighty Allah, Prof. Dr. Touhid Uddin, and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Question Answering System is a system that allows us to query in different type of questionaries' in preferred language and it extracts the exact answer for that question. We tried to work on different important fields and issues of question answering systems. We researched on various established question answering systems and explored their qualities which makes them better in their tasks. Bengali is one of the most popular and commonly used languages in the world. Bengali is the native language of Bangladeshis and 2'nd most popular language in India as it's the native language of the people of West Bengal of India. But it is still in its early stages of research regarding Automated Question Answering system in Bengali language. The type of Question Answering System is conversational Machine Learning and it generates answers which are in natural language to questions raised by users that are humans. A huge progress has been seen in the question answering system and its use in a wide variety of tasks. In recent years, impressive progress has been seen in this sector. Using encoder decoder neural architectures which is trained with big data input helps developing efficient QA system. In this research work, initial steps have been taken by us to bring state-of-the-art Question Answering technology using a BERT model in Bangla Language by designing a Question Answering System which is for basic questions about random subjects.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

An Automated Question answering (QA) system is an Artificial Intelligence system that focuses on generating natural language answers to the questions raised by users and the type of the system is conversational. Initial research works on Automated Question Answering (QA) took place on mid-nineties [1]. Aim of these research works were the same to generate précised answers in response to the questions raised by humans. When users use search engines to get answer of their queries, most of the times search engines can't provide precise answers to the queries. Instead the provide a list of relevant answers. But QA System aims to provide direct answers to the user queries instead of relevant ones and this makes the search engines more relevant. Now-a-days search engines like Google, Bing, Duck Duck Go use automated question answering systems to make the user queries more reliable and user friendly[2]. Machine Reading Comprehension (MRC) and Open-domain QA (OpenQA) are the two main task settings under which textual question answering is being studied on the basis of the availability of contextual information's. Machine Reading Comprehension (MRC) aims to make machines capable of analyzing specified and comprehended context passages to generate answers of queries of user. Machine Reading Comprehension (MRC) was inspired by proficiency exams. On the other hand, Open-domain QA (OpenQA) generates answers of the queries of users without any comprehended or specified context. To generate answers of queries the system search in the local documentations and repositories which are relevant to the queries. Bengali is the eighth most spoken language in the world. In Bangladesh Bengali is spoken as the primary language and is the second most commonly spoken language in India. About 241 million native and 261 million people in total all over the world speak in Bengali. But still compared to English the work in computational linguistic with Bengali language is yet very low. Researches in Bangla language are not as enrich as other languages. In other languages quality research works are done in regular basis to make the use of their language more and more flexible. But in this case, we are far behind them. There is no automated question answering (QA) system in Bengali which will reply to our queries by processing the Bengali Language. But hopefully the situation is changing day by day and researchers are coming forward to enrich the Bangla language researches.

Question Answering Systems can generally be of two domain based. Either of closed-domain based or open-domain based. Closed domain based Question Answering Systems deals with generating answers for questions under specific domains. They don't work out of the specified domains. On the other hands, open domain question answering works on nearly all existing information's and knowledge's of the world.

1.2 Motivation

We in other languages a lot of work has been done in automated question answering system especially in English but not in Bengali language. Bengali is one of the most popular and widely used languages all over the world. A huge number of questions are being raised in Bengali language and the answering process can be simplified by an automated question answering system. Very few work has been done in Bengali question answering system and we don't have much quality research on it and that's why we are working on developing a Bengali question answering system.

1.3 Rational of the study

The key rational are:

1.3.1 Understanding questions in Bengali.

1.3.2 Information retrieval.

1.3.3 General questions with Yes or No output.

1.4 Research Questions

The key questions that this study focuses on are given below:

1.4.1 From where did we get the data?

1.4.2 What is the research objective?

1.4.3 Based on which issue is the study being developed?

1.5 Research Methodology

In the section of methodology of our research paper, we have collected Data, then we analyzed

them, classified the data, selected algorithms, then implemented them, after these steps we evaluated them. At the end of this chapter performance of the proposed model will be described.

1.6 Research Objectives

Our aim is to develop an automated question answering system on Bengali which will be ans to answer question in human manner asked by a user. It will reduce human efforts on question answering procedure.

Some technical issues are pointed below:

1.6.1 Develop an efficient model for question answering system.
1.6.2 To enlighten the software developers to work with ML using the model.

1.6.3 Reduce complexity in Bengali question answering procedure.

1.6.4 Integrate models into mobile apps and websites.

1.6.5 Saving Time.

1.6.6 Reduce complexity in Bengali question answering procedure.

1.6.7 Making queries in Bengali more efficient for answering.

1.7 Research Layout

In chapter 1: all about this project is written here. The reason of choosing this project, how will this project be completed, project motivation, expected outcome and so on is discussed briefly. In a word, chapter 1 is elaboration of introduction of this project.

In chapter 2: related works on this area which were studied are showed. Their findings and limitations are summarized and hence the scope and challenges of the research are also mentioned.

In chapter 3: research methodology will discusses Research Subject and Instrumentation, Data

Collection Procedure, Statistical Analysis and Implementation Requirements

In chapter 4: experimental results and discussion Experimental Results and Descriptive Analysis

In chapter 5: presents a short conclusion. And list of references.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

All over the world a lot of research work is going on in automated question system in many languages. Recently research work in Bengali language has been increased in our country also. But not much quality work has been done in Automated Question Answering system in Bengali.

2.2 Related Works

Research on Automated question answering system one of the most trending topics all over the world. Research work has been done and going on in different languages.

There are machine learning based and rule based approach question answering system [18, 19].

Initially rule based approaches were used in question classifying research works. One of the popular rule based question classifying approach was MULDER which used to classify the pronouns of questions which are interrogated [20].

Verb's objects in questions were used to determine the type of questions. Authors Hermjakob classified 122 classes of questions and wrote 276 rules which were hand-crafted for rule based question classifying approach [21].

People ask a lot of questions on internet about medical cares and treatments and it's very tough to generate accurate answers for this frequent asked questions. Abdelrahman Abdallah and Mahmoud Kasem trained end-to-end model by using encoder decoder and RNN to get useful and valid answers [3].

Although question answering system has developed in many domains but in biomedical still it is a challenge to generate automated answers for frequently asked questions as systems which exists they support few amount of question and answers that’s why more development and research is needed and that’s why Mourad Sarrouti and Said Ouatik El Alaoui intriduced SemBioNLQA which is a semantic biomedical QA system. It has the ability to handle factoid, yes/no, list and summary NLQ (natural language questions) [4].

Question answering helps students to enrich their knowledge and ability. As a huge amount of

documents are generated every day it's impossible for a person to answer all questions. And to solve this problem A. Srivastava, S. Shinde, N. Patel, S. Despande, A. Dalvi and S. Tripathi proposed a state-of-the-art solution that uses NLP(Natural Language Processing) and image captioning techniques which is capable of textual and visual question generation and their answer also[5].

Gathering information from pieces of texts is a tough call. Khot, T., Clark, P., Guerquin, M., Jansen, P., & Sabharwal, A. generated a dataset for multi-hop reasoning that gathers information from a huge corpus and use them to generate multiple-choice answers. In Question Answering via Sentence Composition (QASC) model learns to identify actual information using common scene reasoning [6].

There are a lot of existing method which can generate a structured query for input based questions for syntactic parser. But if the input is parsed incorrectly then a false query is generated and it results answers which are incomplete or flase. When it's about complex questions then the situation becomes worse. Weiguo Zheng, Jeffrey Xu Yu, Lei Zou, and Hong Cheng have proposed a method which is systematic and it uses huge number of binary templates to extract natural language questions rather than semantic parsers [7].

Mio Kobayashi et. al. proposed a hybrid approach which verifies information and statements about historical facts. They collected data from the world history examinations which is in a standardized achievement exams. Their approach predicts the truthfulness of a statement or information word co-occurrence statistic text search, factoid-style question answering [8].

In recent years question answering system has got a lot of attention from researchers. Question Answering system in open domain research has been benefited a lot by Text Retrieval Conference [17]

This conference takes place almost every year. Biomedical Question Answering has been a challenge and very few significant progress has been done on it. J. J. Cimino et. al. [16] developed a question answering system for medical question answering which named MedQA with five components which includes (1) document retrieval, (2) answer extraction, (3) question classification, (4) text summarization (5) query generation.

MedQA helps in automated medical question classification into medical categories of taxonomy developed by J. W. Ely [15] on the basis of a supervised machine learning approach.

Md. Mohsin Uddin et al. presented an end-to-end model which is a heuristic framework used for

paragraphed question answering by using a single supporting line in Bengali Language Question Answering(QA) [9]. A lot of research work has been done using machine learning
In automated and intelligent question answering [10].

Mimir which is an integrated open source framework for semantic text search proposed by Valentin Tablan and Kalina Bontcheva. Mimir is a data driven platform. The matrix of co-occurrence and visualization of information are used for making scene and exploratory search [11].
An algorithm based on graph matching algorithm for r semantic searching for available date in online presented in Universal Networking Language (UNL) was proposed by Mamdouh Farouk, Mitsuru Ishizuka, and Danushka Bollegala. An ontology is not required for a Universal Networking Language (UNL) and the proposed algorithm is supposed to work without any ontology [12].

A complex dataset in Arabic was collected by A. Sayed and Al Muqrishi and they used ontologyin the dataset for semantic searching. Very much complex morphology, huge inflection than other languages made this language one of the most complex language to work with. A. Sayed and Al Muqrishi used Ontological graph and RDF for their research work [13].

"LitVar" is an algorithm which was used by Allot and his team and the dataset on which they worked named SwissProt and ClinVar which were medical datasets containing important information. Allot and his team used text mining which was an advanced text mining technique to extract information which were needed from the medical datasets [14].

2.3 Comparative Analysis and Summary

Generating answers which are in natural language to the human posed questions is the main aim of Question Answering (QA) system. In a Question Answering (QA) system, human and machine are the two participants. There is no other participants in this conversation and this can be defined as a single dialogue. Each participant participants one by one by their turns. In turns of humans they ask questions and the machine answers. In this whole process, different types of topics can be discussed by the Machine agent which is conversational in single turns. When questions use being detected by the Machine agent in the conversation they generates natural language answers to the raised

questions. That's why to develop conversational agents Question Answering Systems are the key components. It's very much popular all over the world and its use is increasing day by day. In this research work, we are designing a Question Answering system for Bangla language and it will help human users to get the answers in Bangla of their frequently asked Bangla questions. This Question Answering system aims to provide precise Bangla answers in response to the user's Bangla questions in natural language.

2.4 Scope of the problem

Since our research work is on Bengali question answering, we can say that it is a very unique work for Bengali. Even though many people worked in Bengal before, their answer predictions were not very good and their dataset was not very big. Our recharge work is much better than that with dataset and new deep learning models that perform very well. This will help us to get a step ahead in Bengali language research domain.

2.5 Challenges

In doing so, as a newbie, we have had to deal with a variety of problems in getting to our actual goal. First of all, it was not a very good dataset in Bengali. So we first collected the data. Then data annotation was a very important step in our problem. Because SQuad like datasets needs a huge human annotation to reach the goal. We faced some issues there as our work is unique. After data annotation and preprocessing we were able to start our actual work.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The methodology covers an absolute of six steps which conclude our research that is displayed in Fig. 3.1. The steps are the following:

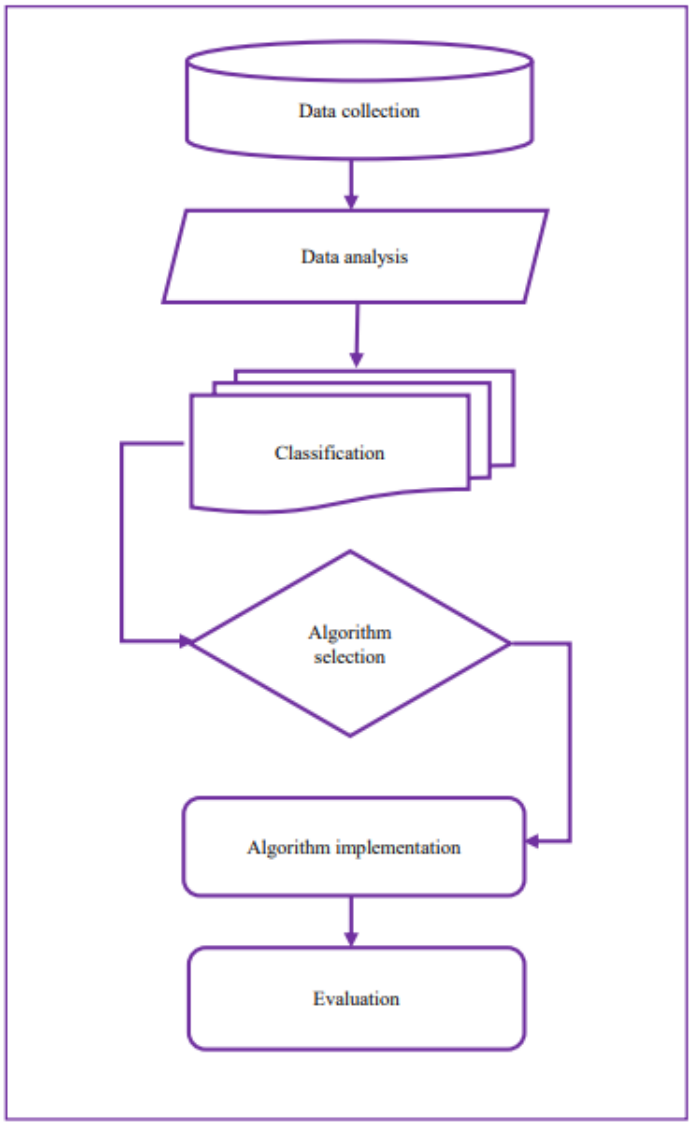


Figure 3.1.1 Methodology diagram

3.2 Data Collection

Data is the most important thing in data-driven techniques. In our, work Bangla is a low resource language and there is no proper human-annotated Reading Comprehension dataset available. So, for making such dataset we first choose Wikipedia as it's a great resource for Bengali content and Wikipedia provides much more valuable insights. That's why we created all the data sets we need from Wikipedia. Our dataset is inspired by the Stanford Question Answering dataset (SQuAD). Dataset contains over 2800 context with over 3000+ Question-answer pairs. Our dataset mimic the format of SQuAD 1.0/2.0 dataset. The list of the things as follows -

• context: Paragraph of which the question has been asked

• qas: questions and answers in a list. Contains as follows:

-id: for each question a unique id

- question: the question

– is_impossible: if the answer is predicted from given questions. Returns a Boolean response

– answers: list contains as follows:

* answer: answer to the given question

* answer_start: starting index of the context that answer start

* answer_end: ending index of the context that answer end

Table 1: Our context data

Context	Question	Answer
ড্যাফোডিল ইন্টারন্যাশনাল ইউনিভার্সিটি বাংলাদেশের একটি বেসরকারী পর্যায়ের উচ্চ শিক্ষা দানকারী প্রতিষ্ঠান। প্রাইভেট বিশ্ববিদ্যালয় অ্যাক্ট ১৯৯২ অনুযায়ী বাংলাদেশের রাজধানী ঢাকার ধানমন্ডি এলাকায় ২০০২ সালের ২৪ জানুয়ারি এ বিশ্ববিদ্যালয়টি প্রতিষ্ঠিত হয়। উচ্চ শিক্ষা প্রতিষ্ঠার জন্য জাতিসংঘের সাথে চুক্তি স্বাক্ষরিত হয় ড্যাফোডিল ইন্টারন্যাশনাল ইউনিভার্সিটির। এ বিশ্ববিদ্যালয়টি বাংলাদেশের আরো ৪টি বিশ্ববিদ্যালয়সহ ইন্টারন্যাশনাল	ড্যাফোডিল ইন্টারন্যাশনাল ইউনিভার্সিটিতে কয়টি অনুষদ রয়েছে ? ড্যাফোডিল ইন্টারন্যাশনাল ইউনিভার্সিটির বিভাগ কয়টি ?	৫ টি ২৩ টি

অ্যাসোসিয়েশন অব ইউনিভার্সিটিসের সদস্য।		
Menth spicata; পরিবার : Labitae; ইংরেজি নাম : Mint পুদিনার সুগন্ধির কারণে বিভিন্ন মুখরোচক কাবাব, সলাদ, বোরহানি ও চাটনি তৈরিতে ব্যবহার হয়। কাঁচা পুদিনা সবচেয়ে বেশি ব্যবহার হয় চাটনি ও সালাদে। ইদানিং বিভিন্ন সামাজিক অনুষ্ঠানে টক দই এবং বোরহানি তৈরির জন্য পুদিনার ব্যবহার বাড়ছে। এছাড়া মাছ, মাংস, সস, স্যুপ, স্টু, চা, তামাক, শরবত তৈরিতে পুদিনা পাতা ব্যবহার হয়। ইউরোপের দেশগুলোতে ভেড়ার মাংসের রোস্ট ও মিন্ট জেলি তৈরিতে পুদিনা পাতা ব্যবহার হয়। বিভিন্ন দেশে পুদিনার বেশি ব্যবহার হচ্ছে তেল তৈরিতে।	ইংরেজিতে পুদিনা পাতার নাম কি ? পুদিনা পাতার সবচেয়ে বেশি ব্যবহার কোথায় ?	Mint চাটনি ও সালাদে

3.3 Data preprocessing

Data processing is a very important state. Data processing after data collection becomes very important for textual & sequence-related problems. Since our dataset contains all textual documents we first add contraction in our dataset, then remove punctuation from them. After replacing Unicode finally we got the full-furnished dataset.

3.4 Data annotation

A supervised machine learning algorithm needs data that are properly labeled. For this labeling process, there comes the data annotation part. Data annotation is a technique that properly labeled the text, images & videos for the machine learning algorithms. For our task, we need to make our dataset like SQuAD [24]. So for that, we use an online tool "Haystack annotation tool" [22] to properly labeled our dataset in SQuAD format. Figure shows the annotation steps we used. At first, we need to upload the context. Next, we choose a question-answer pair to annotate from the given context.

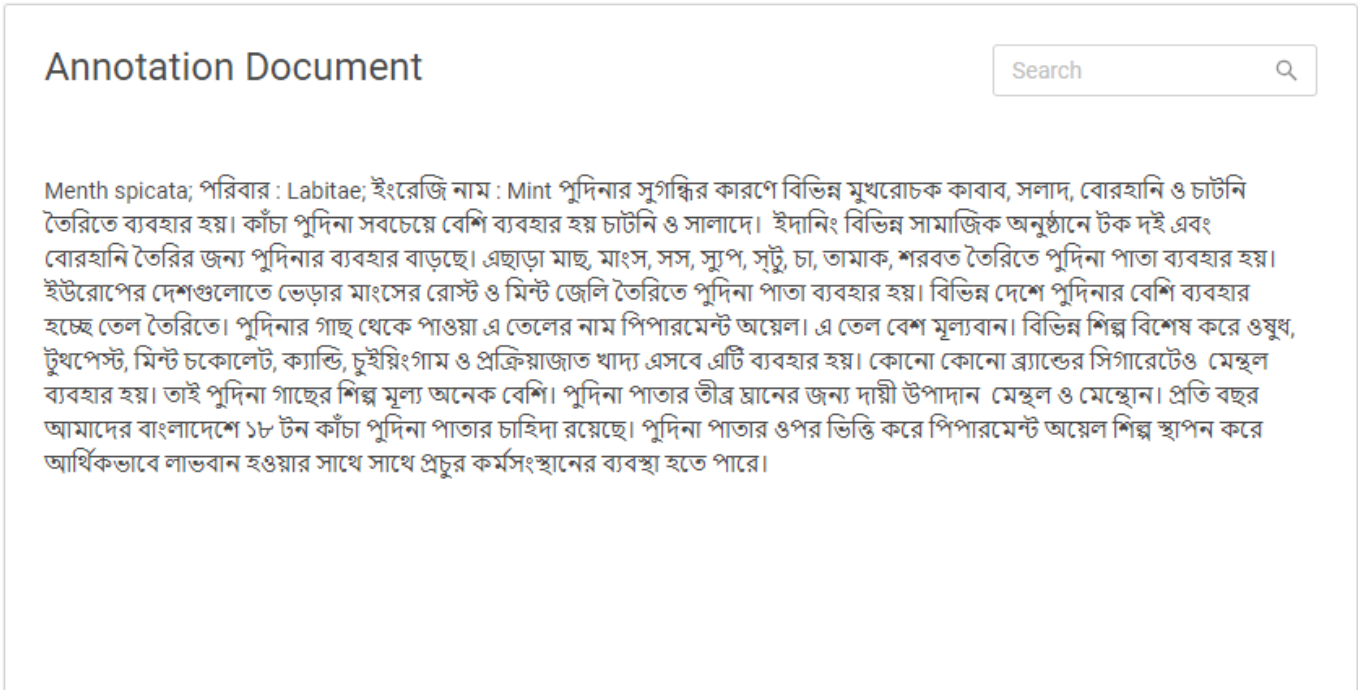


Figure 3.2 (a) : Data annotation

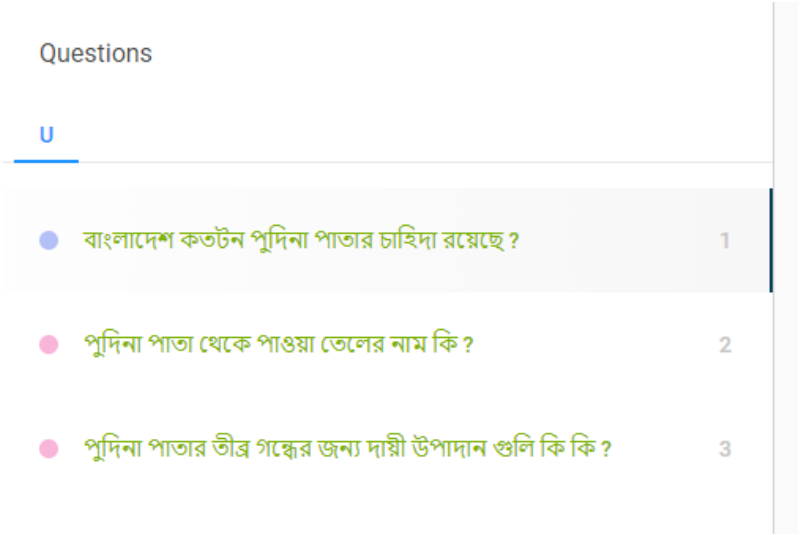


Figure 3.2 (b) : Data annotation

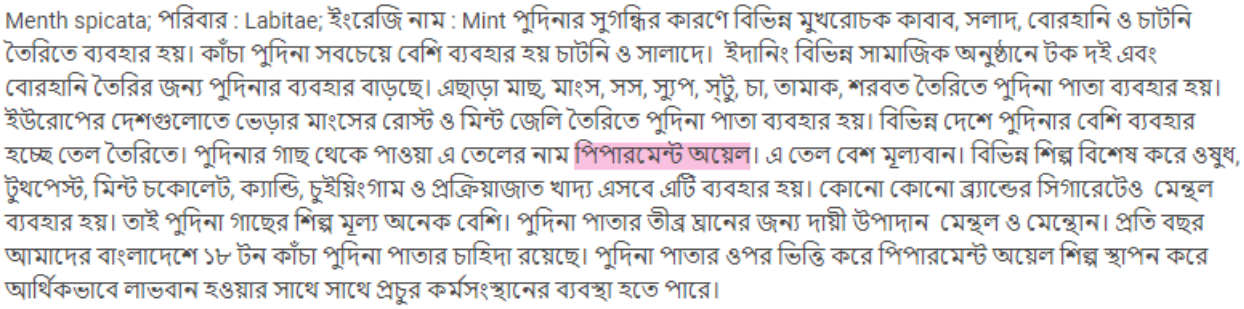

Menth spicata; পরিবার : Labitae; ইংরেজি নাম : Mint পুদিনার সুগন্ধির কারণে বিভিন্ন মুখরোচক কাবাব, সলাদ, বোরহানি ও চাটনি তৈরিতে ব্যবহার হয়। কাঁচা পুদিনা সবচেয়ে বেশি ব্যবহার হয় চাটনি ও সালাদে। ইদানিং বিভিন্ন সামাজিক অনুষ্ঠানে টক দই এবং বোরহানি তৈরির জন্য পুদিনার ব্যবহার বাড়ছে। এছাড়া মাছ, মাংস, সস, স্যুপ, স্টু, চা, তামাক, শরবত তৈরিতে পুদিনা পাতা ব্যবহার হয়। ইউরোপের দেশগুলোতে ভেড়ার মাংসের রোস্ট ও মিন্ট জেলি তৈরিতে পুদিনা পাতা ব্যবহার হয়। বিভিন্ন দেশে পুদিনার বেশি ব্যবহার হচ্ছে তেল তৈরিতে। পুদিনার গাছ থেকে পাওয়া এ তেলের নাম পিপারমেন্ট অয়েল। এ তেল বেশ মূল্যবান। বিভিন্ন শিল্প বিশেষ করে ওষুধ, টুথপেস্ট, মিন্ট চকোলেট, ক্যান্ডি, চুইয়িংগাম ও প্রক্রিয়াজাত খাদ্য এসবে এটি ব্যবহার হয়। কোনো কোনো ব্র্যান্ডের সিগারেটেও মেন্থল ব্যবহার হয়। তাই পুদিনা গাছের শিল্প মূল্য অনেক বেশি। পুদিনা পাতার তীব্র ঘ্রানের জন্য দায়ী উপাদান মেন্থল ও মেন্থোন। প্রতি বছর আমাদের বাংলাদেশে ১৮ টন কাঁচা পুদিনা পাতার চাহিদা রয়েছে। পুদিনা পাতার ওপর ভিত্তি করে পিপারমেন্ট অয়েল শিল্প স্থাপন করে আর্থিকভাবে লাভবান হওয়ার সাথে সাথে প্রচুর কর্মসংস্থানের ব্যবস্থা হতে পারে।

Figure 3.2 (c) : Data annotation

From the above 3 figure we can see the three most important thing in our data annotation part. Figure 3.2(a) is the context or passage that we upload in our annotation tool. Then next one figure 3.2 (b) is the selected questions from the passage and the third one figure 3.2 (c) is highlighted one of the answer that are annotated by human.

3.5 Implementation Requirements

To complete this study and develop the model we have used some hardware instrument

and software instrument. Those are

Software:

Google Colab: It's is an open source software that utilized the idea of making environment. It is an

Allocation of Python and full of hundreds of packages related to scientific programming,

data science, development and more.

JSON editor: We have converted the data set from text format to json by using an annotation tools and then download via json format. Then for simplicity we find the error present here or not we use this software. Then we have put the data set into colab environment for implementation.

Python 3.8: We've used a python version 3.8 for our work. It's an open-source by the python software foundation.

Hardware/Software Requirements:

1. Ram(more than 4 GB)
2. Hard Disk (minimum 4 GB)
3. Key-board, mouse and laptop.

3.6 Model

In NLP & computer vision there a lot of models out there. For especially working with the text previously used machine learning techniques such as SVM, Naive Bayes, etc. For tokenization different vector representations used like bag of words, TF-IDF. Later, comes deep learning for improving performances as the number of data is reached. CNN, LSTM, RNN used here. Previously used vector representations are unable to perform when the data is more & more. These representations waste memory & sometimes not good for quality work. Later, FastText, GLoVE and Word2Vec perform better with deep learning & text representations. For question answering like complex task later BERT came into the picture. With Squad 1.1 dataset question answering tasks make easier. BERT uses vector representations called WordPiece. It increases the performance & predicts answers very correctly.

For our question-answering task, we fine-tune BERT from previously pretrained architecture of bert -multilingual case, uncased & distilbert. As there is no such bigger dataset for Bengali languages that's why we choose transfer learning techniques. It requires limited data to train a new model with the new dataset. BERT-multilingual already trained on a huge number of corpus so, we just transfer the learning & then fine-tune with our dataset.

3.7 BERT

BERT stands for Bidirectional Encoder Representation from Transformers [23] is a neural network for text data purposes. Its performance is excellent in terms of several NLP tasks on GLUE including question-answering of SQuAD 1.1 and SQuAD 2.0 dataset. Many variants of BERT out there. Recently, google use it in their search engine & find impressive results. BERT is inspired by transformers which processes words in relation to all the other words in a sentence instead of looking at them one by one. This allows BERT to look at contexts both before and after a particular word and helps it to pick up features of a language

BERT is inspired by transformer architecture where it uses self-attention mechanism to learn from

the contextual meanings of words, sub-words. It uses encoder - decoder architecture where encoder use for input sequences & decoder for output predictions. In general, seq2seq models like LSTM or RNN reads the sentences from right to left or left to right. But the encoder of transformer read the entire input sequence at once. BERT uses encoder part from transformer & 12 block of transformer, hidden size of 768, attention heads of 12 & 110M parameters in total. BERT large has 24 block of transformer, hidden size of 1024, attention heads of 16 & 340M parameters in total.

3.8 Input sequence of BERT

BERT input can be handle single sentence e.g.(text classification) or sentences of pair e.g.(question, answer) in a single token sequence. A sentence of BERT is considered to be an arbitrary span of contiguous text rather than a linguistic sentence. A sequence can be one sentence or two sentences concatenated together.

BERT uses a tokenizer called workpiece which has 30K vocabulary. Each sentence start with a [CLS] token & end with a [CLS] token. In the middle for divided a sentence pair is uses [SEP] token.

3.9 BERT Question-Answering

For question-answering task BERT is modified for the previous text classification or other purposes. The input question and the context is consider together a whole sentence divided with [SEP] token. Segment A is assigned for question & segment B for context. Start and end vector added to the two new vector as output. The probability of a word being the start of answer span is computed as the dot product between the output representation of the word and the start vector followed by a softmax over the set of words in the passage. The probability of a word being the end of span is also calculated by Equation (1). $P_i = \frac{e^{S.Ti}}{\sum e^{S.Tj}}$ .. …….. .(1)

The score of a candidate span from i_{th} word to j_{th} word is defined as $S.T_i + E.T_j$ and the maximum scoring span where j>=i is used as a prediction. The training objective is the sum of the log-likelihoods of the correct start and end positions. For SQuAD 2.0 for questions that do not have any

answer start and end spans are considered to be at the CLS token.

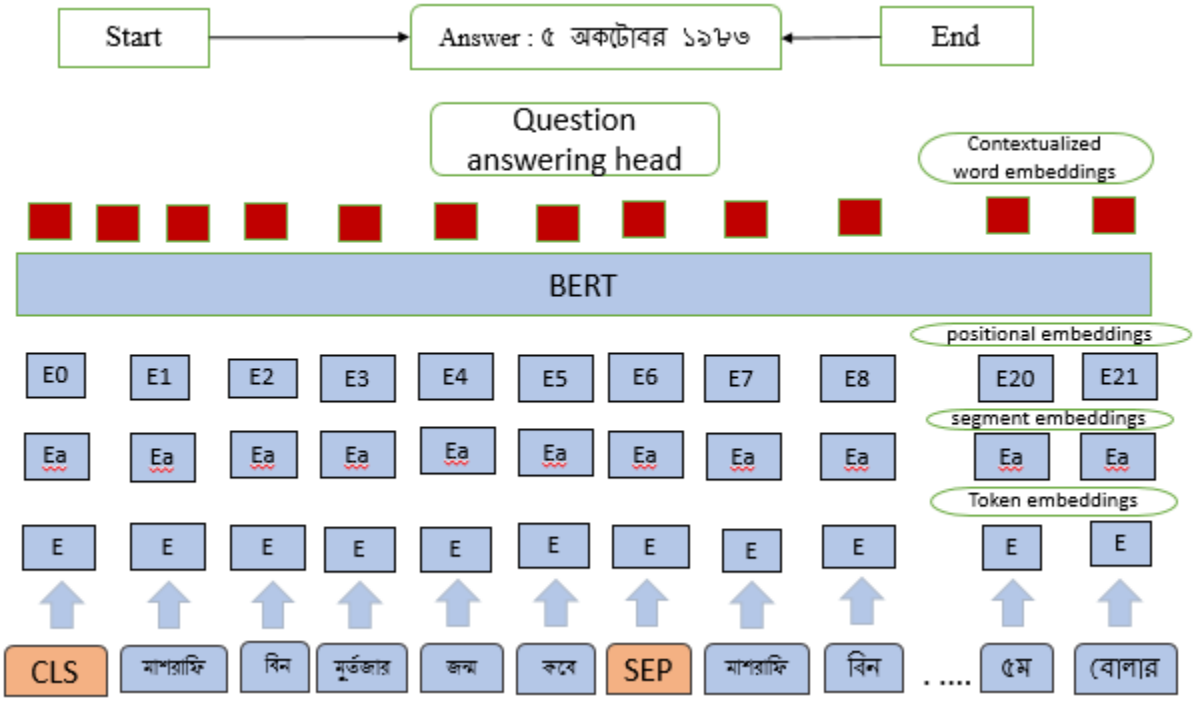


Figure 3.9: Our model architecture

Figure 3.9 shows our model architecture. In this model input question and the context is considered to be together in a whole sentence divided with [SEP] token.

CHAPTER 4

EXPERIMENTAL RESULT

AND ANAYSIS

4.1 Overview:

In this chapter we will discuss about our experimental result that we found after implementing different deep learning algorithm in this system. The experimental setup & result analysis is discussed in this section. First, we introduced the hyper-parameter settings is in experimental setup. Next we discussed about the result analysis part.

4.2 Experimental Setup:

First, the experiment is setup with renowned pytorch framework. We've used another library called Transformers to implement our BERT model along with top of pytorch. For BERT the number of hyper-parameters as follows –

Table 2: Hyperparameter settings

Hyper-parameters	**Value**
learning_rate	3e-5
max_seq_length	354
docs_stride	128
max_answer_length	30
train_batch_size	8
gradient_accumulation_steps	8
Activation	GELU (Gaussian Error Linear Unit)

Next, divide into 85% training & 15% for validation. Then we go for training with our 3 architectures.

4.3 Result analysis:

For question-answering the metrics we follow that is EM(exact match) & f1-score.

Exact Match

Exact match is true/false or binary metric that represents the percentage of predictions that match of any of the true answers only. Like if a ground truth answer and a predicted answer of a question is same then the score is 1, else 0. For example the answer is "Bangla" and the predicted is "Bangla" then score is 1, else 0.

F1-score

F1 Score is measured by multiplication of precision and recall divided by addition of precision and recall and multiply by 2. For the 'Nazrul' example, the precision is 100% as the answer is a subset of the ground truth, but the recall is 50% as out of two ground truth tokens ['Nazrul','Islam'] only one is predicted, so resulting F1 score is 66.67%.

If a question has multiple answer then the answer that gives the maximum F1 score is taken as ground truth. F1 score is also averaged over the entire dataset to get the total score.

After training our BERT model with only 5 epochs our result & answer prediction is given as below.

Table 3: Prediction of BERT-base-multilingual-cased

Context	মাশরাফি বিন মর্তুজা (জন্ম ৫ অক্টোবর ১৯৮৩) হলেন একজন বাংলাদেশী ক্রিকেটার ও রাজনীতিবিদ, যিনি বাংলাদেশ জাতীয় ক্রিকেট দলের সাবেক টেস্ট, ওয়ানডেতে ও টি-টোয়েন্টি অধিনায়ক ছিলেন এবং বর্তমানে নড়াইল-২ আসন থেকে নির্বাচিত জাতীয় সংসদ সদস্য। ইএসপিএন কর্তৃক পরিচালিত \"ওয়ার্ল্ড ফেইম ১০০\" এ বিশ্বের সেরা ১০০ জন ক্রীড়াবিদের মধ্যে অন্যতম। অধিনায়ক হিসেবে ওয়ানডেতে ১০০টি উইকেট নেওয়া বোলারদের মধ্যে তিনি ৫ম বোলার।
Question	মাশরাফি বিন মুর্তজার জন্ম কবে ?
Actual Answer	৫ অক্টোবর ১৯৮৩
Predicted Answer	জন্ম ৫ অকটোবর ১৯৮৩

From the above table 3, we found that the actual & predicted answer is almost same. As our dataset is not big that's why can predict in correct way.

Table 4: Prediction of BERT-base-multilingual-uncased

Context	মাশরাফি বিন মর্তুজা (জন্ম ৫ অক্টোবর ১৯৮৩) হলেন একজন বাংলাদেশী ক্রিকেটার ও রাজনীতিবিদ, যিনি বাংলাদেশ জাতীয় ক্রিকেট দলের সাবেক টেস্ট, ওয়ানডেতে ও টি-টোয়েন্টি অধিনায়ক ছিলেন এবং বর্তমানে নড়াইল-২ আসন থেকে নির্বাচিত জাতীয় সংসদ সদস্য। ইএসপিএন কর্তৃক পরিচালিত \"ওয়ার্ল্ড ফেইম ১০০\" এ বিশ্বের সেরা ১০০ জন ক্রীড়াবিদের মধ্যে অন্যতম। অধিনায়ক হিসেবে ওয়ানডেতে ১০০টি উইকেট নেওয়া বোলারদের মধ্যে তিনি ৫ম বোলার।
Question	মাশরাফি বিন মুর্তজার জন্ম কবে ?
Actual Answer	৫ অক্টোবর ১৯৮৩
Predicted Answer	৫ অকটোবর ১৯৮৩

From the above table 4, we found that the actual & predicted answer is same. As our dataset is not big that's why can predict in correct way.

Table 5: Comparison of two architecture

Model	Loss	Exact match (em)	f1-score
Bert-base-Multilingual-cased	1.80	41.98	57.07
Bert-base-Multilingual-uncased	1.58	43.58	59.89

By using Bert-base-Multilingual-cased model for our context and question analysis while generating answers the loss is 1.80, Exact match (em) is 41.98 and f1-score is 57.07. On the other hand, while using Bert-base-Multilingual-uncased model for our context and question analysis while generating answers the loss is reduced to 1.58, Exact match (em) is 43.58 and f1-score is 59.89.

4.4 Summary

This is the implementation part of our research. In This chapter we implement 2 different deep learning algorithms for our work. The main goal of this thesis is to find out the best model that can performed more accurately. The expected model is found by comparing different performance parameters of algorithm such as EM score and F1-score. After comparing by different parameter, we found that mBERT-case gives us the better performance with both EM and f1-score.

CHAPTER 5

CONCLUSION AND FUTURE RESEARCH

5.1 Overview

In this thesis we tried to find out the best model which can be able to predict the best possible answer more accurately. So that we applied different multilingual BERT architecture. After applying and comparing we found that multilingual BERT-base-cased performed very well than all other. The work flow of this thesis is described below step by step.

Step 1: Data collection from Wikipedia pages

Step 2: Data preprocessing

Step 3: Separated training and testing data.

Step 4: Trained model by different algorithm.

Step 5: Predicting Test result

Step 6: Compare model.

Step 7: Found the Best model for answer prediction.

It's very important to predict answer so that our work is based on this.

5.2 Conclusion

We have worked on developing an automated question answering system in which if we provide an informative passage in Bengali and answer any kind of Bengali question related to the information provided in the passage then our automated question answering system will extract the answers from the provided information. To develop such automated question answering system the main key is having a quality dataset. As not much quality work has been done in this field of Bengali language we had to work from the scratch. Hopefully we got the desired success with a remarkable accuracy and it will open new doors in the ways of Bengali Language based researches. This research work will help others in research in Bengali language.

5.3 Limitations

Every research work has some limitations. We also have some limitations in our work as we worked on Bengali dataset. We have figured out our limitations a lot so that it is much easier to work on these limitations later. Basically Dataset is our first and main limitation. Then it must be said that the model because there are different types of models out there, it is a matter of time to find out which of them performs well. Even then we used a few architectures. Next limitation is that it requires a lot of computational resources to run on the deep learning model. Although we have used Google's free service for computation but in the future our local pc will give a very good performance.

5.4 Implication for Future Research

We ventured to attain the best achievable outcome but there are still few barriers to our work. One of the significant limitations of our work is the inadequate amount of data. Our accumulated data was collected from Wikipedia sources which isn't considered a credible source. Additionally, we have operated with only 1 years of data. In the future, we will try to enhance our data collection by assembling more and more data including a longer time range.

REFERENCE

[1] B. F. Green, Jr., A. K. Wolf, C. Chomsky, and K. Laughery, “Baseball: An automatic question-answerer,” in Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference. ACM, 1961, pp. 219–224.

[2] J. Falconer, “Google: Our new search strategy is to compute answers, not links,” 2011. [Online]. Available: https://thenextweb.com/google/2011/06/01/ google-our-new-search-strategy-is-to-compute-answers-not-links/

[3] Abdallah A, Hamada M and Nurseitov "Automated Question-Answer Medical Model based on Deep Learning Technology" In:Attention-Based Fully Gated CNN-BGRU for Russian Handwritten Text. Journal of Imaging. 10.3390/jimaging6120141, 6:12, (141)

[4] Mourad Sarrouti, Said Ouatik El Alaoui,SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions,Artificial Intelligence in Medicine,Volume 102,2020,101767,ISSN 0933-3657

[5] A. Srivastava, S. Shinde, N. Patel, S. Despande, A. Dalvi and S. Tripathi, "Questionator-Automated Question Generation using Deep Learning," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.212.

[6] Khot, T., Clark, P., Guerquin, M., Jansen, P., & Sabharwal, A. (2020). QASC: A Dataset for Question Answering via Sentence Composition. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 8082-8090.

[7] Weiguo Zheng, Jeffrey Xu Yu, Lei Zou, and Hong Cheng. 2018. Question answering over knowledge graphs: question understanding via template decomposition. Proc. VLDB Endow. 11, 11 (July 2018), 1373–1386. DOI:https://doi.org/10.14778/3236187.3236192

[8] Mio Kobayashi, Ai Ishii, Chikara Hoshino, Hiroshi Miyashita, Takuya Matsuzaki, " Automated Historical Fact-Checking by Passage Retrieval, Word Statistics, and Virtual Question-Answering".

[9] Md. Mohsin Uddin, Nazmus Sakib Patwary, Md. Mohaim, " End-To-End Neural Network for Paraphrased Question Answering Architecture with Single Supporting Line in Bangla Language "

[10] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. V. Merriënboer, A. Joulin, and T. Mikolov, “Towards ai-complete question answering: A set of prerequisite toy tasks,” Computer Science, 2015.

[11] Tablan, V., Bontcheva, K., Roberts, I. and Cunningham, H., 2015. Mímir: An open-source semantic search framework for interactive information seeking and discovery. Journal of Web Semantics, 30, pp.52-68.

[12] Farouk, M., Ishizuka, M. and Bollegala, D., 2018, October. Graph Matching Based Semantic Search Engine. In Research Conference on Metadata and Semantics Research (pp. 89-100). Springer, Cham.

[13] Sayed, A. and Al Muqrishi, A., 2017. IBRI-CASONTO: Ontology-based semantic search engine. Egyptian Informatics Journal, 18(3), pp.181-192

[14] Allot, A., Peng, Y., Wei, C.H., Lee, K., Phan, L. and Lu, Z., 2018. LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. Nucleic acids research, 46(W1), pp.W530-W536.

[15] J. W. Ely, A taxonomy of generic clinical questions: classification study, BMJ 321 (7258) (2000) 429–432. doi: 10.1136/bmj.321.7258.429. URL https://doi.org/10.1136%2Fbmj.321.7258.429

[16] M. Lee, J. J. Cimino, H. R. Zhu, C. Sable, V. Shanker, J. W. Ely, H. Yu, Beyond information retrieval-medical question answering, in: AMIA Annual Symposium Proceedings, Vol. 2006, 2006, pp. 469–473.

[17] J. L. Vicedo, J. Gómez, TREC: Experiment and evaluation in information retrieval, Journal of the American Society for Information Science and Technology 58 (6) (2007) 910–911. doi:10.1002/asi.20583. URL https://doi.org/10.1002%2Fasi.20583

[18] S. K. Ray, S. Singh, and B. P. Joshi, "A semantic approach for question classification using wordnet and wikipedia," Pattern Recognition Letters, vol. 31, no. 13, pp. 1935–1943, 2010.

[19] Z. Huang, M. Thint, and Z. Qin, "Question classification using head words and their hypernyms," in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008, pp. 927–936. 40

[20] C. Kwok, O. Etzioni, and D. S. Weld, "Scaling question answering to the web," ACM Transactions on Information Systems (TOIS), vol. 19, no. 3, pp. 242–262, 2001.

[21] U. Hermjakob, "Parsing and question classification for question answering," in Proceedings of the workshop on Open-domain question answering-Volume 12. Association for Computational Linguistics, 2001, p 1-7

[22] https://annotate.deepset.ai/projects

[23] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.

[24] https://rajpurkar.github.io/SQuAD-explorer/

Jiggasha: A Bengali Question Answering System using Fine-Tuned BERT

ORIGINALITY REPORT

9%	8%	9%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.tandfonline.com Internet Source	5%
2	www.coursehero.com Internet Source	2%
3	arxiv.org Internet Source	1%
4	Sourav Sarker, Syeda Tamanna Alam Monisha, Md Mahadi Hasan Nahid. "Bengali Question Answering System for Factoid Questions: A statistical approach", 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), 2019 Publication	1%
5	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	1%

Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020

Publication

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		