2019 22nd International Conference on Computer and Information Technology (ICCIT), 18-20 December 2019

Bangla Intelligence Question Answering System Based on Mathematics and Statistics

Md. Kowsher
Dept. of Applied Mathematics
Noakhali Science and Technology University, Noakhali-3814,Bangladesh.
ga.kowsher@gmail.com

M M Mahabubur Rahman
Dept. of CSTE
Noakhali Science and Technology University Noakhali-3814, Bangladesh
toufikrahman098@gmail.com

Sk Shohorab Ahmed
Dept. of Information and Communication Engineering,
University of Rajshai, Rajshai-6205, Bangladesh
shohorab.ahmed.it@gmail.com

Nusrat Jahan Prottasha
Dept of CSE
Daffodil International University
Dhaka,Bangladesh
nuaratjahan1234561234@gmail.com

***Abstract*— The Bangla Informative Question Answering System (BIQAS) is a significant Machine Learning (ML) technique that helps a user to trace relevant information by Bengali Natural Language Processing (BNLP). In this research paper, we have applied three mathematical and statistical procedures for BIQAS based on question answering data. These procedures are cosine similarity, Jaccard similarity, and Naive Bayes algorithm. The cosine similarity has interacted with dimension reduction technique SVD on user questions and questions answering data in order to reduce the space and time complexity. These procedures of this research are separated into two parts: pre-processing data and establishment of a relationship between user's questions and contained informative questions. We have got 93.22% accurate answer by using cosine similarity, 84.64% by Jaccard similarity and 91.31% by Naïve Bayes algorithm.**

Keywords— BIQAS, BNLP, Information retrieval, Machine Learning, Mathematics, Statistics.

I. INTRODUCTION

The present time is the era of information. The information is increasing day by day and the world is being more informative, so the virtual information retrieval system keeps its significant that is artificial question answering system. Users often have specific questions in mind, that's why they want to obtain replies. They would like the replies to be easy and precise, and they always prefer to express the questions in their native language without being restricted to a specific query language, query formation rules, or even a specific knowledge domain. The new approach taken to matching the user needs is to carry out actual analysis of the question from a linguistic point of view and to attempt to understand what the user really means.

In Bangla NLP, the BIQAS has been formed by three main modules: data collection, information and user questions processing, and making the relationship between them. Different techniques are held for the sake of pre-processing of BNLP, e.g., anaphora, cleaning special character and punctuation, stop words removing, verb processing, lemmatization, and synonyms word's processing. In order to obtain the perfect anaphora resolution, the famous Hobbs' algorithm is imparted. In lemmatization action, we have described three procedures with a strong system in the lowest time and space complexity. To reduce the dimension of a question and information, we have used the SVD that also minimizes the execution time of program. It also helps understand and calculation in a simple way. The TF-IDF is used to find out the influence of words in documents and constructed the perfect vector.

In order to generate reply of users' questions, we have used Cosine similarity, Jaccard similarity and Naïve-Bayes. These methods aid to establish the relations between users' questions and information.

The contributions are summarized as follows:

• We have introduced mathematical and statistical procedures for BIQAS for information retrieval.

• For the pre-processing of data, we have applied Hobbs' algorithm, Edit Distance, Trie and DBSRA.

• We have used the SVD with cosine similarity to reduce time and space complexity as well as instant answering of questions.

• In order to generate answer of BIQAS Cosine similarly, Jaccard similarity, and Naïve-Bayes algorithms are used.

• To make easy of pre-processing steps, we developed BLTK module that contains all of the pre-processing steps and mathematical techniques.

For the easiest explanation, we consider two users questions as example in every term of this paper, these are

User Question-1: বাংলাদেশের সবচেয়ে বড় ছাত্রী হল কোথায় অবস্থিত? [Where is the largest female hall situated in Bangladesh?]

User Question-2: এটির অডিটোরিয়ামের নাম কি? [What's the name of its auditorium?]

II. RELATED STUDY

Welbl et al. formed the WikiHop dataset which contained questions that needs more than one Wikipedia document to answer [1]. Asiaee et al. devised an ontology-based QA system, Onto NLQA. It has five primary parts such as Linguistic preprocessing, Entity recognition, Ontology element matching, Semantic association discovery and Query formulation and answer retrieval [2]. Xie et al. proposed a question answering system which was based on

978-1-7281-5842-6/19/$31.00 ©2019 IEEE

ontology. From the course of "Natural Language Processing", the ontology data were extracted [3]. Kowsher, Md, et al. proposed a Bangla chatbot based on Bangla language processing. This framework follows three basic steps: question processing, information retrieval (from the web) and answer extraction [4]. Lee et al. proposed an ontology-based QA system. They defined sixteen types of queries in this system. Then the corresponding inferring approach was defined and implemented for each query [5]. Lopez et al. devised an ontology-based question answering system, AquaLog. The input questions are processed and categorized into 23 groups. If the input question falls within one of these groups, the system will process it accurately [6]. Raj proposed a QA system for a specific domain based on the ontological information. This system has four main parts: the question analysis, which analysis the user's question [7]. ROBERT F. SIMMONS et al. devised a Natural Language Question Answering Systems which mainly focused on syntactic, semantic, and logical analysis of English strings [8]. Boris Katz invented the world's first Web-based question answering system called START. It was created by InfoLab Group at the MIT Computer Science and Artificial Intelligence Laboratory which aims to provide with "just the right information," in its place of providing a number of hits [9]. Moldovan et al. utilized syntax-based natural language understanding technique and question classification technique to get better accuracy in the question answering task named TREC[10] and Kowsher, Md, et al. discovered an information-based Bangla Automatic Question Answering System which can provide informative knowledge from users asking [11]. Cai Dongfeng and Cui Huan discovered a web-based Chinese Automatic Question Answering System which uses Google Web services API [12]. Liu Hongshen, Qin Feng, Chen Xiaoping, Tao Tao, et al. proposed a teaching mode software which can keep the attendance of the students and also can keep the answer of any student [13]. Jeon et al. evaluated the question retrieval action for four famous retrieval methods, the vector space model, the Okapi model, the language model, and the translation model [14].Wei Wang, Baichuan Li, Irwin King proposed an improved question retrieval model that can detect users' intentions connected with the former question retrieval [15]. Unlike these works, we have introduced intelligence question answering system of Bangla with the help of mathematics and statistics.

III. Background Study

A. Lemmatization

Lemmatization is a simplification process for finding out the extract root-word in natural language understanding. Lemmatization has been used in a variety of real world applications such as text mining, Chat bot, questions and answering etc.

In this research, we have used an effective lemmatization algorithm for the BNLP. At first we have slightly modified the Trie algorithm based on prefixes. After that we have used a mapping based new algorithm titled as Dictionary-Based Search by Removing Affix (DBSRA).

B. TF-IDF

TF-IDF is the abbreviation of the Term Frequency-Inverse Document Frequency. It is a numerical technique to find out the importance of a word to a sentence and is mathematical-statistically significant. Basically, TF determines the frequency of a term in a sentence and IDF determines the importance of a word to its documents. Mathematically,

$$tf - id = \frac{n_{i,j}}{\sum_k n_{i,j}} * [1 + \log(\frac{N}{df_t})$$

C. Cosine Similarity

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of two non-zero vectors can be derived as:

$$\vec{A}.\vec{B} = ||A||||B|| \cos\theta$$

D. Jaccard Similarity

The Jaccard index also called the Jaccard similarity is a statistical method for determining the similarity of distinct sample sets. If P and Q be two sets the Jaccard index formula is given

$$J_{index}(P,Q) = \frac{|P \cap Q|}{|P \cup Q|} * 100 = \frac{|P \cap Q|}{|P|+|Q|-|P \cap Q|} * 100$$

E. Naive-Bayesian Theory

The Bayesian hypothesis is the most common application of Bayes' theorem that is a statistical hypothesis. Testing of the drug, productions and materials as well as computing the entire output of a company based on machines are the one kinds of example. This theorem states mathematically in the following equation:

M(A | B) = (M(B | A) M(A))/(M(B))

where A and B are events and M(B) ≠0 and M(A | B) is a conditional probability of event A happening given that B is true. Similarly, M(B | A) is also a conditional probability happening given that A is true.

IV. Proposed Work

In this research paper, we have presented a Bengali Intelligence Question Answering System (BIQAS) based on mathematics and statistics using Bengali Natural Language Processing (BNLP).

The procedure is isolated in three parts that are: informative documents collection, pre-processing data and relationships between information and user questions. Corpora have been attached for the pre-processing inserted data. The action of Cosine Similarity, Jaccard similarity, and Naïve Bayes algorithms are urged to obtain the relationship between the questions and answers. But Cosine Similarity deals with vectors. In this case, the documents and questions transmit to vectors using the TF-IDF model. In order to minimize the execution time and space complexity, we have used SVD techniques.

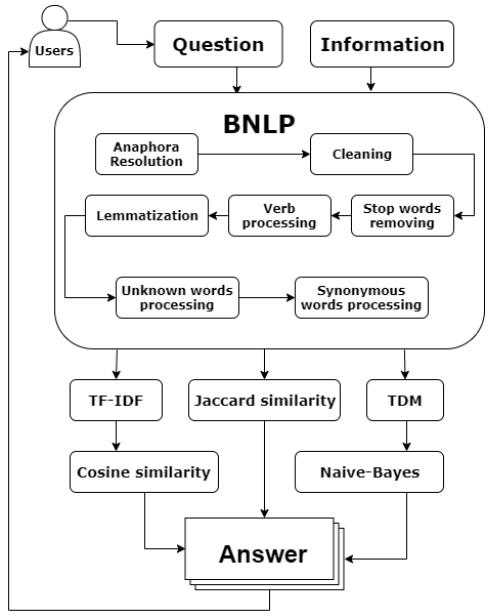

Fig. 1. Proposed Work

V. Pre-Processing

In order to run algorithm, the dataset must be pre-processed. There are several pre-processing techniques are held in BIQAS. The first one is Anaphora which refers to a word that is used earlier in a sentence to avoid repetition, e. g., the pronouns. It requires a successful identification and resolution of NLP. In the proposed BIQAS, we have described a review of work done in the field of anaphora resolution which has an influence on pronouns, mainly personal pronoun using The Hobbs' algorithm.

Table.1: Workflow of Anaphora Resolution.

বাংলাদেশের	সবচেয়ে	বড়	আবাসিক	ছাত্রী	হল	কোথায়	অবস্থিত?
নোবিপ্রবিতে	অবস্থিত।						
এটির	অডিটোরিয়ামের		নাম		কি?		
বীর	মুক্তিযোদ্ধা	হাজী	মোহাম্মদ	ইদ্রিস	অডিটোরিয়াম।		

Here, 'এটি'(It) is the pronoun of the name of 'নোবিপ্রবি'(NSTU). So we used 'এটি' instead of 'নোবিপ্রবি' from the previous example as Anaphora Resolution.

Cleaning word refers to remove an unwanted character which does not have any sentiment on informative data; for example: colon, semicolon, comma, question mark, exclamation point, and other punctuations.

Stop words refer to the words that do not have any influence on documents or sentences. Instances of Bengali stop words are এবং (and), কোথায় (where), অথবা (or), তে (to), সাথে (with), etc. Since our BIQAS is an algorithm based data, the stop words need to be dismissed.

Here we have removed the Bengali stop words 'সবচেয়ে (most)', 'কোথায় (where)' and 'কি (what)' from the previous questions.

In BNLP, there are few verbs that cannot be lemmatized by any system because of ignoring all kinds of lemmatization algorithms. For example, (গেলে, went) and (গিয়ে, going) generate from the root word (যাওয়া, go). There are no relations of character between (গেলে, went) and (যাওয়া, go). So processing with algorithms to these words is not a good choice. That is why these types of verbs are converted into their root verbs for easily accessing as lemmatization.

In our project paper, we used different kinds of lemmatization techniques like DBSRA and Trie. Sometimes there are few words in the Bangla language which do not work in Trie but work in DBSRA or work in Trie but not in DBSRA. So we have used Levenshtein distance to find out the best lemma word between DBSRA and trie.

Lemmatization algorithms are not a good choice for unknown words. Here unknown words refer to the name of a place, person or name of anything. Levenshtein distance assists to determine which word is known or unknown. We count the probability of edit between lemma and word (before lemmatization). If the probability P (lemma |word) is greater than 50% [P (lemma |word) > 50%], then it is counted as unknown words.

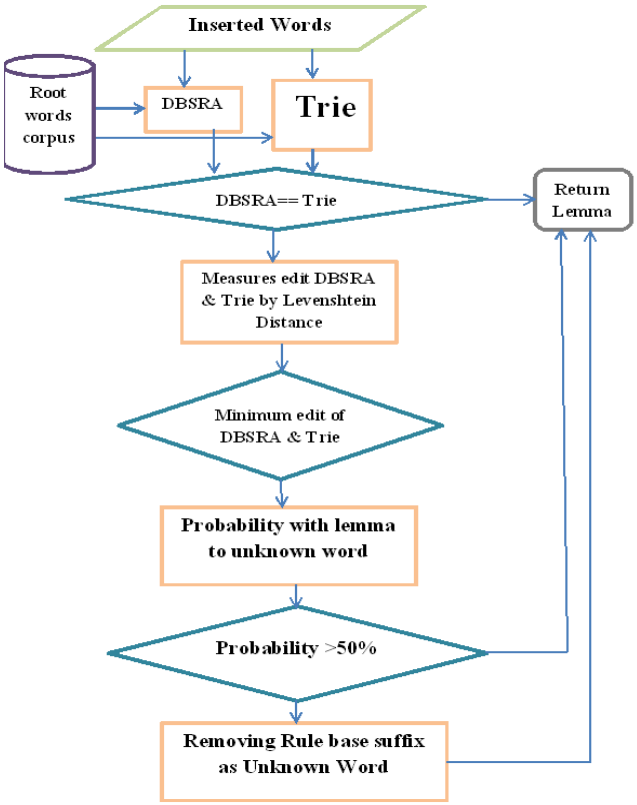

Fig. 2. Workflow of lemmtiazation

In order to process the unknown words, we have established a corpus of the suffix of Bengali language; for instances: তে (te), ছে (che), য়ের(yer), etc. The longest common suffix has been removed from the last position of an unknown word. Thus we obtain the lemma or root of an unknown word.

Synonym words indicate the exact or nearly the same meaning of different words. Users ask questions having words that are not available in information data but its meanings do. In this sense, BIQAS may fail to answer correctly. So synonym words processing has a significant

action in the BIQAS as well as in Natural Language Understanding (NLU).

Here 'বৃহৎ (large)' is the synonym of 'বড় (big)' and 'বৃহ' (large) is considered as a common word.

After preprocessing the question will be

User Question-1: বাংলাদেশ বৃহৎ আবাসিক ছাত্রী হল অবস্থিত [Where is the largest female hall situated in Bangladesh?]

User Question-2: নোবিপ্রবি অডিটোরিয়াম নাম [What's the name of its auditorium?]

VI. Establishment of Relationship

A. Cosine Similarity

In order to represent the words of user questions or informative questions as numerically, we have used the Vectorization method TF-IDF model. For the simplification for our task, we have taken two examples from the considered corpus. The TF-IDF value of the informative questions and user question have been shown in the table

Table 2: Cosine Similarity Calculation

Terms	Words				IDF	TF*IDF			
	IQ-1	IQ-2	q1	q2		IQ-1	IQ-2	q1	q2
নোবিপ্রবি	1/3	0	0	1/3	1.301	0.4337	0	0	0.4337
অডিটোরিয়াম	1/3	0	0	1/3	1.301	0.4337	0	0	0.4337
নাম	1/3	0	0	1/3	1.301	0.4337	0	0	0.4337
বাংলাদেশ	0	1/6	1/6	0	1.301	0	0.217	0.217	0
বিশ্ববিদ্যালয়	0	1/6	0	0	1.301	0	0.217	0	0
বৃহৎ	0	1/6	0	0	1.301	0	0.217	0	0
ছাত্রী	0	1/6	1/6	0	1.301	0	0.217	0.217	0
হল	0	1/6	1/6	0	1.301	0	0.217	0.217	0
থাকা		1/6	0	0	1.301	0	0.217	0	0

Let us set the term weights and construct the term-document matrix C and query matrix:

$$C = \begin{bmatrix} 0.4337 & 0 \\ 0.4337 & 0 \\ 0.4337 & 0 \\ 0 & 0.217 \\ 0 & 0.217 \\ 0 & 0.217 \\ 0 & 0.217 \\ 0 & 0.217 \\ 0 & 0.217 \end{bmatrix}, q_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0.217 \\ 0 \\ 0.217 \\ 0.217 \\ 0.217 \\ 0 \end{bmatrix}, q_2 = \begin{bmatrix} 0.4337 \\ 0.4337 \\ 0.4337 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Now we will use SVD in matrix C and will find U, Σ and V matrices, where

$$C = U\Sigma V^T$$

$$\begin{bmatrix} -0.57735 & 0 \\ -0.57735 & 0 \\ -0.57735 & 0 \\ 0 & -0.40824 \\ 0 & -0.40824 \\ 0 & -0.40824 \\ 0 & -0.40824 \\ 0 & -0.40824 \\ 0 & -0.40824 \end{bmatrix} \begin{bmatrix} 0.75119 & 0 \\ 0 & 0.53153 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}^T$$

Row of V holds the eigenvector value. These are the coordinates of individual document vectors. Hence

$IQ_1 = (-1,0)$ And $IQ_2 = (0,-1)$

To find the new query vector co-ordinates, we have

$$q = q^T U_k \Sigma_k^{-1}$$

Now for first question:

$$= \begin{bmatrix} 0 & 0 & 0 & 0.217 & 0 & 0.217 & 0.217 & 0.217 & 0 \end{bmatrix} \begin{bmatrix} -0.57735 & 0 \\ -0.57735 & 0 \\ -0.57735 & 0 \\ 0 & -0.40824 \\ 0 & -0.40824 \\ 0 & -0.40824 \\ 0 & -0.40824 \\ 0 & -0.40824 \\ 0 & -0.40824 \end{bmatrix} \begin{bmatrix} \frac{1}{0.75119} & 0 \\ 0 & \frac{1}{0.53153} \end{bmatrix}$$

$$q_1 = \begin{bmatrix} 0 & \frac{-11073727}{66442375} \end{bmatrix}$$

Similarly, $q_2 = \begin{bmatrix} \frac{-150238017}{150238000} & 0 \end{bmatrix}$

Now we will find the cosine similarities.

$$\cos\theta_1 = \text{sim}(q_1, IQ_1) = \frac{q_1 \cdot IQ_1}{|q_1|\,|IQ_1|}$$

$$= \frac{(0)(-1)+(-\frac{11073727}{66442375})(0)}{\sqrt{(0)^2+(-\frac{11073727}{66442375})^2}\sqrt{(-1)^2+(0)^2}}$$

$$= 0$$

Similarly,

$$\cos\theta_2 = sim(q_1, IQ_2) = \frac{q_1 \cdot IQ_2}{|q_1|\,|IQ_2|}$$

$$= \frac{(0)(0)+(-\frac{11073727}{66442375})(-1)}{\sqrt{(0)^2+(-\frac{11073727}{66442375})^2}\sqrt{(0)^2+(-1)^2}}$$

$$= 1$$

We can see that $sim(q_1, IQ_2) > sim(q_1, IQ_1)$. So user question1 can be found in informative question2, i.e. IQ_2. The user question-2 can be found in informative question1, i.e. $IQ1$.

B. Jaccard Similarity

After pre-processing the questions and data can be revealed as sets

Now, let an example using set notation and Venn-diagram

P= {বাংলাদেশ, বড়, আবাসিক, ছাত্রী, হল, অবস্থিত}
Q= {বাংলাদেশ, বিশ্ববিদ্যালয়, বৃহৎ, ছাত্রী, হল, থাকা}

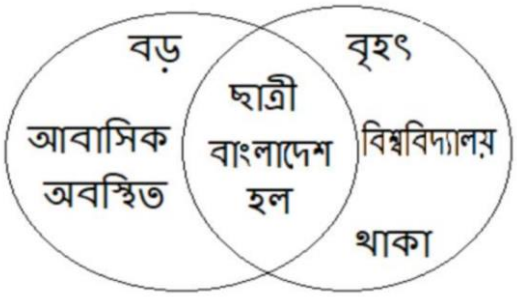

Fig.3: Jaccard Similarity

$$J_{index}(P,Q) = \frac{|P \cap Q|}{|P \cup Q|}$$

$$= \frac{|\{\text{বাংলাদেশ, ছাত্রী, হল,}\}|}{|\{\text{বাংলাদেশ, বড়, আবাসিক, ছাত্রী, হল, অবস্থিত, বিশ্ববিদ্যালয়, বৃহৎ, থাকা}\}|} * 100$$

$$= \frac{3}{9} = 0.33$$

In this way, we can find $sim(q_1, IQ_2) = \frac{1}{3}$ and $sim(q_1, IQ_1) = 0$ So, the reply of question-1 can be founded in informative question-2 $sim(q_2, IQ_1) = 1$ And $sim(q_2, IQ_2) = 0$ so, the reply of question-2 can be founded in informative question-1.

C. Naive Bayes Experiments

After pre-processing data, we have converted every word to term-document matrix then calculated probability

Table 3: Naïve Byes Calculation

	Doc	Questions	Class
Training	1	নোবিপ্রবি অডিটোরিয়াম নাম	IQ1
	2	বাংলাদেশ বিশ্ববিদ্যালয় বৃহৎ ছাত্রী হল থাকা	IQ2
Test	3	বাংলাদেশ বৃহৎ ছাত্রী হল অবস্থিত	q1

Here,

P(বাংলাদেশ | IQ1)= $\frac{|0+1|}{|3+10|} = \frac{1}{13}$ P(বাংলাদেশ | IQ2)= $\frac{|1+1}{|6+10|} = \frac{1}{8}$

P(বৃহৎ| IQ1)= $\frac{|0+1|}{|3+10|} = \frac{1}{13}$ P(বৃহৎ | IQ2)= $\frac{|1+1}{|6+10|} = \frac{1}{8}$

P(ছাত্রী | IQ1)= $\frac{|0+1|}{|3+10|} = \frac{1}{13}$ P(ছাত্রী | IQ2)= $\frac{|1+1}{|6+10|} = \frac{1}{8}$

P(হল | IQ1)= $\frac{|0+1|}{|3+10|} = \frac{1}{13}$ P(হল | IQ2)= $\frac{|1+1}{|6+10|} = \frac{1}{8}$

P(অবস্থিত | IQ1)= $\frac{|0+1|}{|3+10|} = \frac{1}{13}$, P(অবস্থিত | IQ2)= $\frac{|0+1|}{|6+10|} = \frac{1}{16}$

Now Probability

P(IQ_1)= $\frac{1}{2}$, P(IQ_2)= $\frac{1}{2}$

P($IQ_1|Q_1$)= $\frac{1}{2} * (\frac{1}{13})^5 = \frac{1}{742586}$,

P($IQ_2|Q_1$)= $\frac{1}{2} * (\frac{1}{8})^4 * \frac{1}{16} = \frac{1}{131072}$

Since P($IQ_2|Q_1$) > P($IQ_1|Q_1$) so the reply of question Q_1 can be found in (IQ_2. Similarly the reply of Q_2 can be found.

VII. Experiments

We have described a range of experiments to measure our proposed model the mathematical and statistical procedures for BIQAS. In this section, first, we present the questions that we target to reply to the experiments and describe the experimental setup. Then, we discuss the performance and result of our propounded work.

A. Corpus

For the implication of BIQAS, we describe mainly five types of corpus. In the first corpus, there are 28,324 Bengali root words. The main aim of this corpus is to lemmatize Bengali words. The second one that contains 382 Bengali stop words is to remove the stop words from the inserted documents and questions. We have compiled 74 topics as informative documents like as hall information, department information, teacher information, library, NSTU nature, bus schedules etc. of Noakhali Science and Technology University (NSTU) that is our third corpus as questions with its relevant answer of document's information. In this work, we have originated 3127 questions from our inserted documents as our fourth corpus. Every question contains its corresponding answer. To test our data, we have created 2852 questions from relevant 74 topics.

1 Categories:
2 Noakhali Science and Technology University.
3 Information :
4 - - নোবিপ্রবি কোথায় অবস্থিত?
5 – নোয়াখালী শহর থেকে আট কিলোমিটার দক্ষিণে সোনাপুর–সুবর্ণচর সড়কের পশ্চিম পাশে।
6 -- নোবিপ্রবি কত একর জায়গা নিয়ে গঠিত?
7 – ১০১ একর।
8 -- নোবিপ্রবি বাংলাদেশের কততম পাবলিক বিশ্ববিদ্যালয়?
9 - ২৭ তম।
10 -- নোবিপ্রবির ওয়েবসাইট এড্রেস কি?
11 - nstu.edu.bd
12 - - নোবিপ্রবির অডিটোরিয়ামের নাম কি?
13 - বীর মুক্তিযোদ্ধা হাজী মোহাম্মদ ইদ্রিস অডিটোরিয়াম।
14 - - বাংলাদেশের কোন বিশ্ববিদ্যালয়ে সবচেয়ে বৃহত্তম ছাত্রী হল রয়েছে?
15 – নোয়াখালী বিজ্ঞান ও প্রযুক্তি বিশ্ববিদ্যালয়ে।

Fig. 4. Questions and Answers(Data)

B. Experimental Setup

We implemented our propounded model and in Anaconda distribution with Python 3.7 programming language and executed them on a Windows 10 PC with an Intel Core i7, CPU (3.20GHz) and 8GB memory.

Python is a high-level object-oriented language (OOP) which is suitable for scientific examination and tools development. We have used the Anaconda as the apportionment of Python. Anaconda creates the best stage for open source data science which is powered by Python. In order to preprocess, we have used the BLTK tool [3, 11] that is the Bengali Language Toolkit. The BLTK also provides the TF-IDF, SVD, cosine similarity, Jaccard similarity, and Naive Bayes algorithm.

C.Result and Analysis

To test our proposed BIQAS, We have created 2852 questions as testing data from selected 74 topics of Noakhali Science and Technology University (NSTU) and we obtained 93.22% accuracy in cosine Similarity, 82.64% in Jaccard Similarity and 91.34% in Naive Bayes classifier.

D. Comparison between English Chatbot and BIQAS

Since there are no chatbot in Bengali like BIQAS, so there is any other Bengali chatbot to compare with our work. Our work is at present the state of the art for the Bengali Intelligence Bot. So we compare BIQAS with two English chatbots which are Neural Conversational Machine (NCM) and Cleverbot.

What's your mobile number? (তোমার মোবাইল নাম্বার কি?)
Mitsuku : That information is confidential.
BIQAS: আমার কোন মোবাইল নেই। আমার সাথে যোগাযোগ করতে আমার প্রোগ্রামটি রান করুন।

How old are you?(তোমার বয়স কত?)
Mitsuku: I am 18 years old.
BIQAS: আমি এখনও তরুন।

What is your address?(তোমার ঠিকানা কি?)
Mitsuku: I am in Leeds.
BIQAS:আমার প্রোগ্রামটি যে কম্পিউটারে ইনস্টল থাকবে সেটিই আমার ঠিকানা।

What's your mobile number?(তোমার মোবাইল নাম্বারকি?)
Mitsuku: That information is confidential.
BIQAS: আমার কোন মোবাইল নেই। আমার সাথে যোগাযোগ করতে আমার প্রোগ্রামটি রান করুন।

How old are you?(তোমার বয়স কত?)
Mitsuku: I am 18 years old.
BIQAS:আমি এখনও তরুন।

What is your address?(তোমার ঠিকানা কি?)
Mitsuku: I am in Leeds.
BIQAS: আমার প্রোগ্রামটি যে কম্পিউটারে অনুভূতি থাকবে সেটিই আমার ঠিকানা।

VIII. Conclusion & Future Works

The main challenge of this project paper is to implement a Bengali intelligence bot for information retrieval. We have shown the theoretical and experimental methodology of our proposed work. In this scientific paper, we have described three procedures using machine learning, mathematics, and statistics. To establish the full methodology, we have followed some procedures like pre-processing, time and space reduction, and established the relation between information and questions.

In the future, the proposed BIQAS system can be enabled for the purpose of educations, industry, business and personal tasks with voice replying system. An advance it can be shaped with the assist of Deep Learning algorithms such as Recurrent Neural Network (RNN) by processing the BNLP.

References

[1] Clark, Peter, et al. "Think you have solved question answering? try arc, the ai2 reasoning challenge." arXiv preprint arXiv:1803.05457 (2018).

[2] Asiaee, Amir Hosein. A framework for ontology-based question answering with application to parasite data. Diss. University of Georgia, Athens, GA, USA, 2013.

[3] Kowsher, Md, et al. "Doly: Bengali Chatbot for Bengali Education." 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT). IEEE, 2019.

[4] Shah, Urvi, et al. "Information retrieval on the semantic web." Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002.

[5] Abdi, Asad, Norisma Idris, and Zahrah Ahmad. "QAPD: an ontology-based question answering system in the physics domain." Soft Computing 22.1 (2018): 213-230.

[6] Lopez, Vanessa, et al. "AquaLog: An ontology-driven question answering system for organizational semantic intranets." Web semantics: science, services and agents on the world wide web 5.2 (2007): 72-105.

[7] Raj, P. C. "Architecture of an ontology-based domain-specific natural language question answering system." arXiv preprint arXiv:1311.3175 (2013).

[8] Simmons, Robert F. "Natural language question-answering systems: 1969." Communications of the ACM 13.1 (1970): 15-30.]

[9] Yu, Zheng-Tao, et al. "Answer extracting for chinese questionanswering system based on latent semantic analysis." CHINESE JOURNAL OF COMPUTERS-CHINESE EDITION29.10 (2006).

[10] Whittaker, Edward, Sadaoki Furui, and Dietrich Klakow. "A statistical classification approach to question answering using web data." 2005 International Conference on Cyberworlds (CW'05). IEEE, 2005. 11.

[11] Kowsher, Md, Imran Hossen, and SkShohorab Ahmed. "Bengali Information Retrieval System (BIRS)." International Journal on Natural Language Computing (IJNLC) 8.5 (2019).

[12] Dongfeng, Cai, et al. "A Web-based Chinese automatic question answering system." The Fourth International Conference onComputer and Information Technology, 2004. CIT'04.. IEEE, 2004.

[13] Stalin, Shalini, Rajeev Pandey, and Raju Barskar. "Web based application for hindi question answering system." International Journal of Electronics and Computer Science Engineering 2.1 (2012): 72-78.

[14] Jeon, Jiwoon, W. Bruce Croft, and Joon Ho Lee. "Finding semantically similar questions based on their answers." Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005.

[15] Wang, Wei, Baichuan Li, and Irwin King. "Improving question retrieval in community question answering with label ranking." The 2011 International Joint Conference on Neural Networks. IEEE, 2011.