International Conference on Bangla Speech and Language Processing(ICBSLP), 27-28 September, 2019

Bengali Question Answering System for Factoid Questions: A statistical approach

Sourav Sarker
Computer Science and Engineering
Shahjalal University of
Science and Technology
Sylhet-3114, Bangladesh
sourav39@student.sust.edu

Syeda Tamanna Alam Monisha
Computer Science and Engineering
Shahjalal University of
Science and Technology
Sylhet-3114, Bangladesh
alammonisha@gmail.com

Md Mahadi Hasan Nahid
Computer Science and Engineering
Shahjalal University of
Science and Technology
Sylhet-3114, Bangladesh
nahid-cse@sust.edu

***Abstract*— Question answering system in recent days is one of the most trending and interesting topics of research in computational linguistics. Bengali being among the most spoken languages in the world has yet faced difficulties in computational linguistics. This paper demonstrates an attempt to develop a closed domain factoid question answering system for Bengali language. Our proposed system combining multiple sources for answer extraction extracts the answer having the accuracy 66.2% and 56.8% with and without mentioning the object name respectively. The system also hits around 72% documents from which the answer can be extracted. Besides the sub-parts of our system, the question and document classifier provides 90.6% and 75.3% accuracy respectively over five coarse-grained categories.**

***Index Terms*— Question Answering (QA) System, Bengali Question Answering System, Factoid QA System**

I. INTRODUCTION

An automated question answering system is a program that is able to converse with the user in natural language in such a way that no one is able to differentiate it from a real human being. In today's time question answering system is one of the hot topics in Natural Language Processing (NLP) research. It can be either closed-domain based or open-domain based. Closed-domain question answering deals with questions under a specific domain whereas open-domain question answering deals with questions about nearly anything based on world knowledge. In general a question is of two types: factoid question which is satisfied by a short text and descriptive or complex question which needs to be answered in more than one line.

Being one of the hot topics the works being done on question answering system is increasing day by day. In English a lot of works have been done on question answering system and also there exists a number of working question answering systems. But although Bengali being one of the most widely spoken languages the works done for Bengali language compared to English is yet very low.

We worked towards developing a question answering system for only factoid questions for Bengali language on a closed domain. Initially we have chosen Shahjalal University of Science & Technology (SUST) as our domain because every year during the admission test of Shahjalal University of Science & Technology (SUST) the candidates have a lot of queries related to SUST. In general there are different social media groups which are formed due to helping the candidates with information, updates, queries etc. Candidates also ask questions in the official website of admission test. So we wanted to create a common platform for the candidates where they can get answers to the queries. Thus we wanted to build a Bengali question answering system which can reply to the queries instantly.

II. RELATED WORKS

A lot of researches have been done and are also ongoing on question answering system in different languages. Besides there are also works on question classification, question features, question taxonomies and answer extraction which are the sub-parts of question answering system. There have been number of question answering systems developed since the 1960s. Among the earlier question answering systems some of them are domain restricted and some are generalized.

AnswerBus is an open-domain question answering system where the information related to the answer is retrieved from the web in sentence level. The authors used five search engines (Google, Yahoo, Altavista, WiseNut and Yahoo News) to extract the web documents which contain the answers to the questions of the users. The current rate of correct answers to TREC-8's (Text REtrieval Conference-8) 200 questions is 70.5% [1]. JAVELIN is

978-1-7281-5242-4/19

another open-domain based question answering system [2]. The team suggested three QA runs JAVELIN I, JAVELIN II [3] and JAVELIN III [4] among which JAVELIN III can be used for cross-lingual task. Some of the earlier domain restricted QA systems are BASEBALL and LUNAR [5][6]. A system was described by author's Abney et. al., 2000 that handles arbitrary questions by producing a candidate list of answers ranked by their plausibility [7]. The system was evaluated on the TREC question-answering track which showed that the correct answer to queries appeared in the top five answers 46% of the time with a mean score of 0.356.

Nowadays there is a remarkable increase in the research for Bengali language in the field of Natural Language Processing (NLP). Question answering system being one of the hot topics the research works for question answering system for Bengali language is also increasing at this time.

Author's Banerjee et al., 2014 made the first attempt on building a factoid question answering system for Bengali language where the answer is processed by help of named entities [8]. They also discussed the challenges faced to develop the system for Bengali language. A question answering system was developed for Bengali using anaphora-cataphora resolution [9]. Authors experimented the system for both Bengali and English language and they used semantic and syntactical analysis. The model reduces the complexity of using noun instead of pronoun for the requested answer with respect to the given question queries for Bengali and provides 60% accuracy.

III. CORPUS CONSTRUCTION & ANALYSIS

Dataset is an important issue towards the development of a question answering system. To build a question answering system we need two types of data. One is the knowledge base and the other is question database. Then the data are prepared individually for the training of question classification and document categorization.

As we have worked on a closed domain based Bengali factoid question answering system on the domain Shahjalal University of Science & Technology (SUST) we could not get any prepared dataset for our work. So we had to prepare our own questions and knowledge base. This data collection task was one of the challenging tasks as we had no resource available. We used different sources for our data collection part. We had to collect data both for questions and documents in different ways.

A. Question Database

Various sources were used for building the question dataset such as crowd sourcing, social media, manual generation etc. But among them, we collected most of the data from crowd sourcing. We collected questions from the students of Shahjalal University of Science & Technology (SUST) as well as from the official website of SUST where we got the frequently asked questions by the candidates. We also prepared questions from different documents, articles and news on Shahjalal University of Science & Technology (SUST).

B. Knowledge Base

Knowledge base is the set of documents from where the answers of the questions are to be searched and extracted. We have built our knowledge base based on the website which carries the information solely about Shahjalal University of Science & Technology (SUST) like `www.sust.edu`, `en.wikipedia.org/wiki/Shahjalal_University_of_Science_and_Technology` etc. and news from different web portals. We also collected some of the paragraphs about SUST by crowd-sourcing.

TABLE I
Sources of Question Database and Knowledge Base

Data Type	Collection Type	Source	Amount of Data
Question	Crowd Sourcing	2nd Year students, Dept of CSE,SUST	11300
	Social Media Data Crawling	Facebook Group: SUST Admission Aid	1055
	Manual Generation	Authors	3000
Document	Crowd Sourcing	2nd Year students, Dept of CSE,SUST	40
	SUST based websites	www.sust.edu, Wikipedia sust	100
	News Portals	Bdnews24.com, eprothom-alo.com, www.sustnews24.com thedailystar.com etc.	80

IV. ANSWER TYPE TAXONOMY

As we have worked for a closed domain question answering system on the domain Shahjalal University of Science & Technology (SUST) we defined five coarse-grained classes related to SUST for question classification and document categorization. The questions and documents are classified in these five categories. Table II shows the category details.

TABLE II
Question and Document Categories

Class Name	Short form	Description
Administration	ADS	Questions that require administrative information as answer and documents of administrative type are of ADS class
Admission	ADM	Questions that require admission related information as answer and documents of admission type are of ADM class
Academic	ACD	Questions that require academic information as answer and documents of academic type are of ACD class
Campus	CAM	Questions that require campus related information as answer and documents related to campus are of CAM class
Miscellaneous	MISC	Questions that require any information other than the above four types and documents other than the four types are of MISC class

TABLE III

Sample Tagged Question

Question	Label
সাস্টে মোট কতটি সিট আছে	ADM
সাস্টে প্রতিবছর কতটি করে স্কলারশিপ দেয়	ACD
সাস্ট মানে কি	MISC
সরকারী বৃত্তি প্রাপ্তদের কি ভর্তি ফি দিতে হয়	ADS
সাস্ট ক্যাম্পাসে কি খাবারের দাম সহনীয়	CAM

TABLE IV

Sample Tagged Document

Document	Label
মানবতার জন্য শেখো স্লোগানকে প্রতিপাদ্য করে ২০১২ সালের ১১ জানুয়ারী কিছু সচেতন শিক্ষক-শিক্ষার্থীর হাত ধরে শুরু হয় শাহজালাল বিজ্ঞান ও প্রযুক্তি বিশ্ববিদ্যালয়ের একমাত্র প্রকৃতি ও পরিবেশ বিষয়ক সংগঠন গ্রিন এক্সপ্লোর সোসাইটির পথচলা। বিশ্ববিদ্যালয়ের শিক্ষার্থীদের পাশাপাশি তৃনমূল পর্যায়ে পরিবেশ সচেতনতা বৃদ্ধিতে কাজ করে যাচ্ছে সংগঠনটি।	CAM

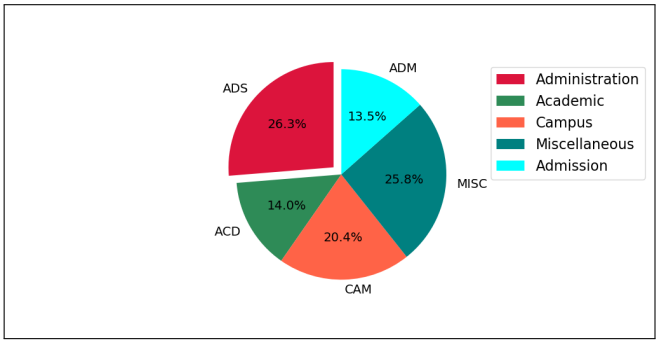

Fig. 1. Percentage of questions in each category

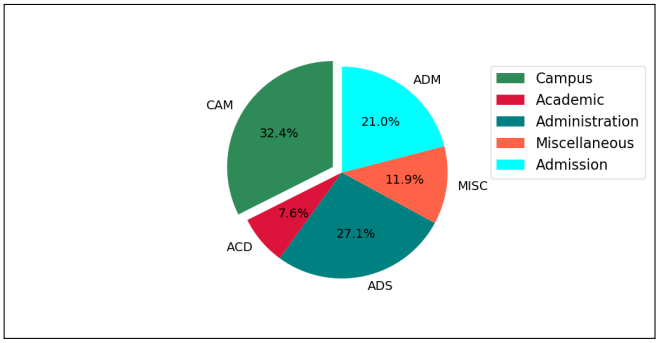

Fig. 2. Percentage of documents in each category

V. METHODOLOGY

The architecture that we have proposed for our system is shown in figure 3. As shown in the figure we have used three sources for answer extraction: mapped question, collection of documents and internet resource.

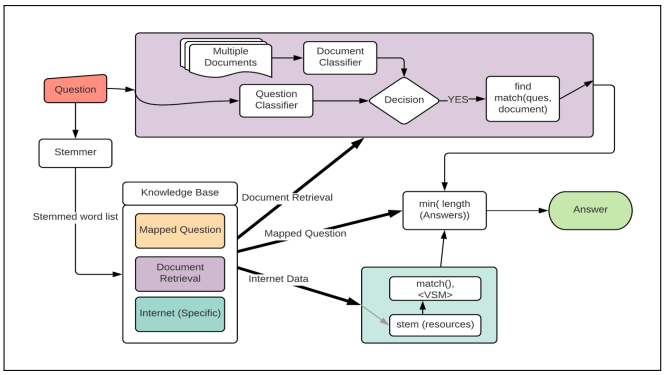

Fig. 3. Proposed architecture for our system

As depicted in the figure for extracting an answer at first a list of stem words is created from the question. We have used a Bengali rule based stemmer for the process. At first the answer is looked for in the frequently asked section. Here we have mapped the most frequently asked questions to the answers. If the answer is found here considering common tag words (শাবি, শাবিপ্রবি, শাহজালাল বিশ্ববিদ্যালয়, শাহজালাল বিজ্ঞান ও প্রযুক্তি বিশ্ববিদ্যালয়, সাস্ট) then an answer will be provided otherwise the answer will be searched in the categorized documents. In this case the asked question is first classified to the expected answer type and then searched in the documents of the same type and if the answer is not found in the categorized documents then it searches in the internet sources which are predefined for the particular domain and provides an answer.

We wanted to build a question answering system for Bengali language on a closed domain and not having enough resources to build it, we had to work from the scratch. We have divided our methodology to answer a factoid question in four basic parts:

1) Data Preparation
2) Question Classification
3) Document Categorization
4) Answer Processing

A. Data Processing

As we collected data from different sources we needed to prepare and clean the dataset. The following tasks were done to prepare the corpus for further processing.

1) Removed stop words (অবশ্য, অনেক, অনেকে, অনেকেই, অন্তত, ভাবে, মধ্যে etc.)
2) Removed sign characters
3) Tokenized the words in special case
4) Checked and made correction of spelling mistakes of raw data manually
5) Rechecked the labels that were assigned manually

B. Question Classification

Question classification is an important first step towards building a question answering system [10]. The classification of a question to expected answer type reduces the search space by a considerable amount [11]. We have used four machine learning algorithms Stochastic Gradient Descent(SGD), Decision Tree(DT), Support Vector Machine(SVM) and Naive Bayes(NB) for our question classification phase [12]. For feature extraction similar words and both bi-gram and tri-gram were used where tri-gram provides better results. Again, we have tested dynamic word clustering model as deep learning methods are getting popularity in these days [13].

C. Document Categorization

The documents that we have collected were also classified to predefined categories. For document categorization we have used a different approach other than question classification. As document classification follows passage classification we used word embedding here. Word embedding is one kind of learned representation for text where words that have the same meaning have a nearly same representation. We have used fastText as our embedding technique and implemented convolutional neural network(CNN) classifier.

D. Answer Processing

The most important and challenging part of our question answering system is the answer extraction method. As shown in figure 3 we have used three sources for answer extraction process. The following techniques have been used for our answer extraction phase.

1) Vector Space Model (VSM): In vector space model documents and queries are represented as vectors of features representing the terms that occur within the collection. A document d_j and a question q can be represented by the following vectors.

$$\vec{d_j} = (w_{1,j}, w_{2,j}, w_{3,j}, w_{4,j}, \ldots, w_{n,j})$$

$$\vec{q} = (w_{1,q}, w_{2,q}, w_{3,q}, w_{4,q} \ldots, w_{n,q})$$

Here the number of dimensions in the vector is the total number of terms in the whole collection. The match found by the question from a particular document is measured by the following similarity function:

$$sim(\vec{q}, \vec{d_j}) = \frac{\sum_{i=i}^{N} w_{i,q} * w_{i,j}}{\sqrt{\sum_{i=1}^{N} w_{i,q}^2} * \sqrt{\sum_{i=1}^{N} w_{i,j}^2}}$$

After calculating this score an answer is generated using the sentence having the highest rank.

2) Comparison of VSM and Edit distance: We have used a comparison between edit distance and VSM and received the maximum value between edit distance and VSM matching. Comparing both, the answer with the highest score is provided and if multiple answers have the same score then the one with minimum length is selected as factoid questions require single facts as answer.

E. Performance Metric

We have used the accuracy measure to evaluate our experiments. Accuracy is the number of correct predictions made by a model over all kinds of predictions made which is measured by the formula:

$$Accuracy = \frac{TP + TN}{P + N}$$

where,

P = total number of positive samples

N = total number of negative samples

TP (True Positive) = number of cases when the actual class of the sample is true and also predicted as true

TN (True Negative) = number of cases when the actual class of the sample is false and also predicted as false

VI. EXPERIMENTS & RESULT ANALYSIS

For the training purpose of question classification and document categorization we had to manually tag the questions and documents. In both cases we have used 75% of the dataset for training purpose and 25% for testing. For question classification Support Vector Machine(SVM) with linear kernel provides the best accuracy 90.6% among classifiers Stochastic Gradient Descent(SGD), Decision Tree(DT), Support Vector Machine(SVM) and Naive Bayes(NB) that we used [12].

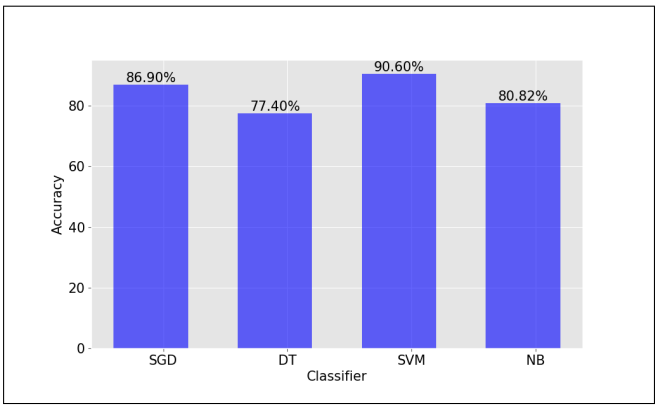

Fig. 4. Best performance for different classifier for question classification

For document categorization we have implemented convolutional neural network(CNN) using fastText skip-gram word embedding technique. Accuracy of 75.3% was obtained for document categorization over the five

categories that we defined.

Following our answer extraction technique the proposed system provides around 56.8% accuracy without mentioning the object name like শাবি, শাবিপ্রবি, শাহজালাল বিশ্ববিদ্যালয়, শাহজালাল বিজ্ঞান ও প্রযুক্তি বিশ্ববিদ্যালয়, সাস্ট and around 66.2% with mentioning the object name. We have received a document hit of around 72% for our system.

TABLE V
Sample answers to asked questions

Question	Extracted Answer
সাস্টে কয়টি অনুষদ রয়েছে	৭ টি
সাস্টের বর্তমান ভিসির নাম কি	অধ্যাপক ফরিদ উদ্দিন আহমেদ
ছেলেদের হল কয়টি	৩ টি

TABLE VI
Performance of overall system

Type	Score	Condition
Document Hits	72%	Considering object name
Answer accuracy	66.2%	Considering object name
Answer accuracy	56.8%	Without mentioning object name

VII. CONCLUSION

In our whole time of working with Bengali question answering system we have tried to build a generic factoid question answering system that is if we just provide the system with the knowledge base and question database, it will be able to extract answers from them. We have built a corpus consisting of 15355 questions and 220 documents on our domain Shahjalal University of Science & Technology for our system. The dataset or corpus creates the main hazard for building a question answering system or any language tool. No well-established Bengali language tool has been released till date. So for the better performance of a question answering system natural language processing tools like named entity recognizer, parts of speech tagger, stemmer etc. and also the corpus which is base of the question answering system need to be developed.

So far we have worked with factoid questions only. We have future plans to forward the process by taking complex and descriptive questions into consideration. We also want to implement more techniques to enhance the performance of the system.

References

[1] Z. Zheng, "Answerbus question answering system," in *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 2002, pp. 399–404.

[2] E. Nyberg, T. Mitamura, J. G. Carbonell, J. P. Callan, and K. Collins-Thompson, "The javelin question-answering system at trec 2002," 2002.

[3] E. Nyberg, R. E. Frederking, T. Mitamura, M. W. Bilotti, K. Hannan, L. Hiyakumoto, J. Ko, F. Lin, L. V. Lita, V. Pedro *et al.*, "Javelin i and ii systems at trec 2005." in *TREC*, vol. 2, no. 1, 2005, p. 1.

[4] T. Mitamura, F. Lin, H. Shima, M. Wang, J. Ko, J. Betteridge, M. W. Bilotti, A. H. Schlaikjer, and E. Nyberg, "Javelin iii: Cross-lingual question answering from japanese and chinese documents." in *NTCIR*, 2007.

[5] 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.

[6] W. A. Woods, "Progress in natural language understanding: an application to lunar geology," in *Proceedings of the June 4-8, 1973, national computer conference and exposition*. ACM, 1973, pp. 441–450.

[7] S. Abney, M. Collins, and A. Singhal, "Answer extraction," in *Proceedings of the sixth conference on Applied natural language processing*. Association for Computational Linguistics, 2000, pp. 296–301.

[8] S. Banerjee, S. K. Naskar, and S. Bandyopadhyay, "Bfqa: A bengali factoid question answering system," in *International Conference on Text, Speech, and Dialogue*. Springer, 2014, pp. 217–224.

[9] S. Khan, K. T. Kubra, and M. M. H. Nahid, "Improving answer extraction for bangali q/a system using anaphora-cataphora resolution," in *2018 International Conference on Innovation in Engineering and Technology (ICIET)*. IEEE, 2018, pp. 1–6.

[10] S. Banerjee and S. Bandyopadhyay, "Bengali question classification: Towards developing qa system," in *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, 2012, pp. 25–40.

[11] E. Haihong, Y. Hu, M. Song, Z. Ou, and X. Wang, "Research and implementation of question classification model in q&a system," in *International Conference on Algorithms and Architectures for Parallel Processing*. Springer, 2017, pp. 372–384.

[12] S. T. A. Monisha, S. Sarker, and M. M. H. Nahid, "Classification of bengali questions towards a factoid question answering system," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT 2019)*. IEEE, 2019, pp. 660–664.

[13] Z. S. Ritu, N. Nowshin, M. M. H. Nahid, and S. Ismail, "Performance analysis of different word embedding models on bangla language," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. IEEE, 2018, pp. 1–5.