Machine Learning Approaches for Bengali Automated Question Detection System

Saifuddin Al Azad Sagor
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet-3114, Bangladesh

Nurul Azim Rizvi
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet-3114, Bangladesh

Md Mahadi Hasan Nahid
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet-3114, Bangladesh

ABSTRACT
Question detection is the initial tasks of question answering system. Question detection, the first step of QA system has been done applying SVM, Logistic Regression, K-Neighbor classifier, Multilayer Perceptron and LSTM. We achieved best performance for SVM with linear kernel trick. For 2000 feature words we got 1.4% error rate which is best among the approaches we used for Bengali language. For LSTM algorithm we got 3.22% error rate. LSTM looks promising with large and better dataset.

General Terms
Bengali Question Detection, Bengali Question Answering (QA) System

Keywords
Bengali Question Detection, Bengali Question Answering (QA) System, Question Detection, SVM, LSTM.

1. INTRODUCTION
Various questions are asked worldwide in every second. For finding answer of those questions we have to have information regarding that topic. These information are in books, papers and most importantly on web. With the improvement of Internet and web, more and more information's became available and easier to find for all. When we need to find answer of a specific question then we actually get that answer from some information available to us regarding that question. We manually read available information in that topic and try to find the answer. As more information are available on Internet and web if we can automate the process of reading and getting the answer from some documents then we'll be able to answer more questions automatically. In this paper we'll describe some initial steps taken and method applied for building automated question answering system (QA) in Bengali language. Question detection serves great purpose in QA system [1]. As, people often ask questions in natural language they don't follow grammatical rules. So, a sentence is a question or not should be determined first. Different pattern of questions by people makes it more difficult to detect [2]. For fetching answer from documents for asked question the system needs to process that question and the first step of that processing is question detection thus it's the first step of question answering system [3]. Table I shows some example of Bengali questions

Table 1. Bengali question example

□□□□□□□□□□□□ □□□□□□ কত □□□□ □□□□□□□□□□ □□□□□?
□□□□□□□ □□□□□□ □□□□ □□□□□□□ হয় ?
□□□□□□□□□□□□ □□□□□□ □□□□□ □□□□ □□?
□□□□□□□□□□□□ আয়তন কত?
□□□□□□□□□□□□ □□□□ □□□□□□□□□□ কত?
…

After the question detection part the system needs to classify the question into a category from some predefined categories. This classification helps to guess expected answer type and shrink the search domain. Classification hence is a crucial part of QA system.

2. METHODOLOGY
2.1 Related works in this topic
BFQA [4] has three components and these are question analysis, sentence extraction and answer extraction. The question analysis step has five stages and these are question type identification, expected answer type identification, named entity identification, question topical target identification and keyword identification. Figure 1 shows the BFQA architecture in detail.

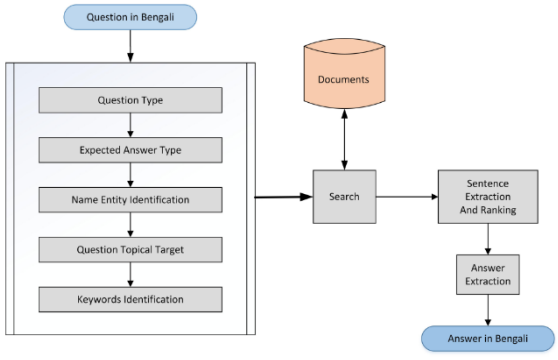

Fig.1. BFQA architecture [4]

BASEBALL [5] is a QA system which was developed in very early phase. This system is domain restricted and provides information associated with a baseball league played in USA. It was one of the first attempt of developing a working QA

system. Another domain restricted QA system LUNAR [6] was developed for the mission Apollo. It worked on geological analysis documents collected from Apollo lunar exploration. JAVELIN is a system which is independent and works for open domain question answering. It has very well defined and modular architecture [7].

AskMSR QA system is different than traditional QA system in architecture. With tremendous redundancy it uses Web as a

Gigantic data repository. These data can be exploited for question-answering [8].

Vast data is available on Web in different languages. AnswerBus question answering system is a very effective sentence level web information retrieval technique or system [9]. Based on RDF architecture a QA system was built [10]. QAKis [11] is built based on Relational Patterns. It's an open domain QA system.

For traditional biomedicine task HPI QA system was developed that didn't use any latest technique or idea [12]. Bengali language question was categorized into some predefined coarse grained categories that represents expected answer type. Support Vector Machine was used there [13].

2.2 Dataset Preparation and Analysis

We prepared a dataset of total 8000 lines of Bengali text. In our dataset, there were 3100 questions and 4900 non-questions. These data were collected from different websites. Some articles about famous poets of Bangladesh were collected from Wikipedia. We also collected question answer pair from tatoeba project. These data were prepared manually. Then we had to tag these data for our question detection purpose. We classified the data in two categories which are question and non-question.

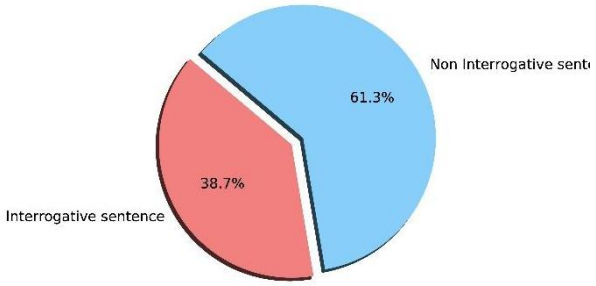

Fig.2. Overview of Dataset

We manually cleaned the whole dataset for our work. Then we did sentence tokenization for the further processing. Then we did word tokenization in the sentences we got. Then we converted the words to vec using countvectorizer. After that, using tfidf we scored the words and got the highest valuable words.

2.3 Feature Extraction

1 Sentence Tokenization: First, we separated each sentence from the passage we collected for further processing. In sentence tokenization many things were considered for getting better result.

2 Word Tokenization: After sentence tokenization we applied word tokenization in those sentences. By word tokenizing we separated each word of the sentences and stored them for further processing.

3 *Wh-word:* After word tokenization our task got easier for us. In Bengali language there is some word just like wh-word in English. These words became very handy for us. They are mostly found in questions. So, it's obvious that these Wh-word are the most frequent words after word tokenization.

We gave these words higher priority than other words. After giving higher priority to wh-word, a sentence's probability to be a question becomes higher if these words occur. Table II shows some example of Bengali Wh-words.

Table 2. Bengali Wh-word example

□□?
□□□?
□□?
□□□□□□□?
…

4) Most Frequent Words: Though we saw that wh-word's frequency were higher than other words but we got some other words which had higher frequencies too along with wh words. We took these special words as our feature also. We did unigram analysis in our dataset to find these most frequent words which played a significant role in our experiment. Table III shows the top 13 most frequent words in our dataset along with frequencies.

2.4 Training and Recognition

For training purpose, we used SVM, MLP, KNN, Logistic Regression and LSTM algorithm. These algorithms were run in our large dataset of 8000 Bengali text. At first, we ran SVM algorithm and used 3 different kernels of SVM. The kernels are Linear, Polynomial and RBF.We used MLP where hidden layers size were 12 and solver was lbfgs. KNN is another machine learning algorithm we used. The leaf size were 30 and algorithm used is auto. KNNs metric was minkowski and weights used was distance. We ran machine learning algorithm LR (Logistic Regression) to classify our text dataset according to our goal. LSTM is one of the most used algorithm for classifying text data. The last algorithm we used is LSTM. When implementing LSTM model our number of feature words were 20000. Sequential model was used and input length was 42. Total layers used in our model are 21 and activation function is sigmoid. The best Result is achieved for SVM. The performance will be better with LSTM if the dataset is improved.

Table 3. Bengali question most frequent words

Word	Frequency	Word	Frequency
□□□	482	ও	162
□□	326	□□□□□	158
□□□□	243	□□	152
□□□	204	কত	147
হয়	193	□□□□	147
□□	184		

3. RESULTS AND ANALYSIS

After successful training of our dataset we got higher accuracy and lesser error which was quite satisfying. In every algorithm except LSTM, our data train test split was 80/20 that is 80% of total data was used for training and 20% was for testing. For LSTM algorithm our training and testing data were 66.67% and 33.33% respectively.

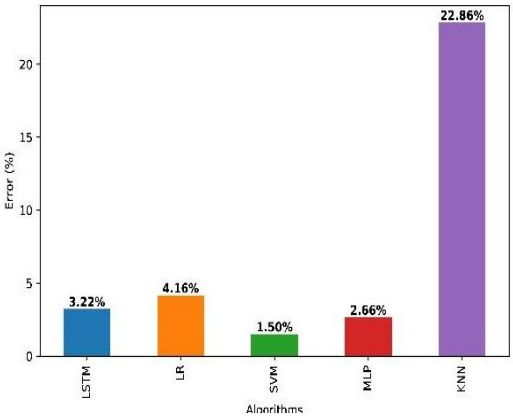

Fig. 3. Performance of Different algorithms.

The error percentage we got is quite satisfying. We applied SVM with different kernel trick, Logistic Regression, MLP, KNN and LSTM. The least error rate we got by applying SVM with linear kernel and it is 1.50%. Worst error rate we got was for KNN and the error rate is 22.86%. Logistic Regression, MLP and LSTM's error rates are 4.16%, 2.86% and 3.22% respectively. One can't say which the best algorithm of machine learning is as it depends on many things like dataset, feature set etc. So, we had to experiment with several algorithms to get the best result. Figure 3 represents the error rate of used classifiers. We performed LSTM algorithm but didn't get the better result compared with SVM. The reason behind this is shortages of labeled data. We expect to get better result for LSTM by increasing dataset. Figure 4 presents the best, average and worst error rate for LSTM algorithm. SVM is the best performing algorithm we applied. We applied SVM with Linear, Polynomial and RBF kernel trick. For 500 feature words error rate of Linear, Polynomial and RBF kernel is 2.5%, 3.1% and 2.5%. For 1000 feature words error rate of Linear, Polynomial and RBF kernel is 1.6%, 3.1% and 2%. For 1500 feature words error rate of Linear, Polynomial and RBF kernel is 1.7%, 3% and 1.7%. For 2000 feature words error rate of Linear, Polynomial and RBF kernel is 1.4%, 3.1% and 1.5%. So, we can easily see from the observation that SVM with Linear kernel trick for 2000 feature words gives the best result. Figure 5 represents the performance of SVM with different kernels.

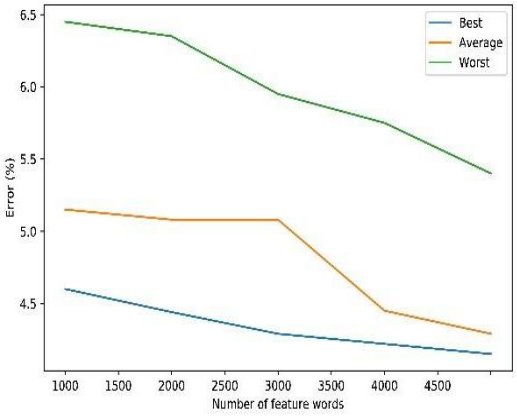

Fig. 4. Performance of LSTM algorithm

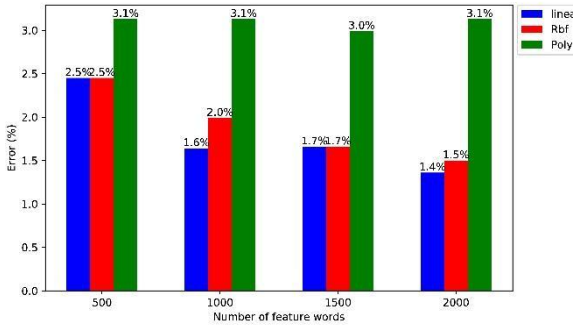

Fig. 5. Performance of SVM with different kernels.

4. CONCLUSION

We have done some of the initial works of QA system for Bengali language. Further steps need to be done for a complete QA system. It needs lots of effort to do that and good enough dataset. If dataset is increased we hope to get a better performance by using LSTM. Other methods can also be tried for better result. Question classification, Retrieving information and Answer extraction are the parts of QA system that can be a huge domain of research in Bengali Natural Language Processing. Question detection is a subpart of question answering system and many other problems. It can open door for many other research fields. Described question detection system will give better performance with better dataset.

5. REFERENCES

[1] K. Wang and T.-S. Chua, "Exploiting salient patterns for question detection and question retrieval in community-based question answering," in Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010, pp. 1155–1163.

[2] K. Boakye, B. Favre, and D. Hakkani-Tür, "Any questions? Automatic question detection in meetings," in Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on. IEEE, 2009, pp. 485–489.

[3] S. Azzam, N. Caldwell, and D. P. Del Carpio, "Automatic question and answer detection," Oct. 15 2013, uS Patent 8,560,567

[4] 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.

[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] E. Nyberg, T. Mitamura, J. G. Carbonell, J. Callan, K. CollinsThompson, K. Czuba, M. Duggan, L. Hiyakumoto, N. Hu, Y. Huang et al., "The javelin question-answering system at trec 2002," Proceedings of TREC 11, 2002.

[8] E. Brill, S. Dumais, and M. Banko, "An analysis of the askmsr question-answering system," in Proceedings of the ACL-02 conferenceon Empirical methods in natural language processing-Volume 10.Association for Computational Linguistics, 2002, pp. 257–264.

[9] 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.

[10] L. Zou, R. Huang, H. Wang, J. X. Yu, W. He, and D. Zhao, "Natural language question answering over rdf: a graph data driven approach," in Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014, pp. 313–324.

[11] E. Cabrio, J. Cojan, A. P. Aprosio, B. Magnini, A. Lavelli, and F. Gandon, "Qakis: an open domain qa system based on relational patterns," in Proceedings of the 2012th International Conference on Posters & Demonstrations Track-Volume 914. CEUR-WS. org, 2012, pp. 9–12.

[12] F. Schulze, R. Schüler, T. Draeger, D. Dummer, A. Ernst, P. Flemming, C. Perscheid, and M. Neves, "Hpi question answering system in bioasq 2016," in Proceedings of the Fourth BioASQ workshop at the Conference of the Association for Computational Linguistics, 2016, pp. 38–44.

[13] S. M. H. Nirob, M. K. Nayeem, and M. S. Islam, "Question classification using support vector machine with hybrid feature extraction method," in Computer and Information Technology (ICCIT), 2017 20th International Conference of. IEEE, 2017, pp. 1–6.