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 1 shows some example of Bengali questions

<table>
<thead>
<tr>
<th>Question in Bengali</th>
<th>Answer in Bengali</th>
</tr>
</thead>
<tbody>
<tr>
<td>কত?</td>
<td>হয়?</td>
</tr>
<tr>
<td>আয়তন কত?</td>
<td>আয়তন কত হয়?</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

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.
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.

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.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. KNN’s 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.
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.

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>ও</td>
<td>482</td>
<td>য়</td>
<td>162</td>
</tr>
<tr>
<td>326</td>
<td></td>
<td>158</td>
<td></td>
</tr>
<tr>
<td>243</td>
<td></td>
<td>152</td>
<td></td>
</tr>
<tr>
<td>204</td>
<td></td>
<td>147</td>
<td></td>
</tr>
<tr>
<td>হয়</td>
<td>193</td>
<td>কত</td>
<td>147</td>
</tr>
<tr>
<td>184</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. 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


