Bengali Question Classification: Towards Developing QA System

Somnath Banerjee *Sivaji Bandyopadhyay*
Department of Computer Science and Engineering
Jadavpur University, India
s.banerjee1980@gmail.com, sivaji_cse_ju@yahoo.com

ABSTRACT

This paper demonstrates the question classification step towards building a question answering system in Bengali. Bengali is an eastern Indo-Aryan language with about 230 million total speakers and one of the most spoken languages in the world. An important first step in developing a question answering system is to classify natural language question properly. In this work, we have studied suitable lexical, syntactic and semantic features to classify the Bengali question. As Bengali question classification is at early stage of development, so for simplicity we have proposed single-layer taxonomy which consists of only nine course-grained classes. We have also studied and categorized the interrogatives in Bengali language. The proposed automated classification work is based on various machine learning techniques. The baseline system based on Naïve Bayes classifier has achieved 80.65% accuracy. We have achieved up to 87.63% accuracy using decision tree classifier.

KEYWORDS : Bengali Question Classification, Question Classification, Machine Learning.

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 25–40,
COLING 2012, Mumbai, December 2012.

1 Introduction

Because of the high level of information overload on the Internet, research into question answering is becoming increasingly important. Question answering systems focus on how to respond to users' queries with exact answers. In recent years, many international question answering contests have been held at conferences and workshops, such as Text REtrieval Conference (TREC), Cross Language Evaluation Forum (CLEF) and NII Test Collection for IR Systems (NTCIR). Although Bengali is the sixth most spoken languages in the world, no QA contest in Bengali has been conducted so far.

Bengali (Bengali: বাংলা (Bangla)) is an eastern Indo-Aryan language. It is native to the region of eastern South Asia known as Bengal, which comprises present day Bangladesh, the Indian state of West Bengal, and parts of the Indian states of Tripura and Assam. Besides this region, there are also significant Bengali-speaking communities in: the Middle East (namely, UAE, Saudi Arabia, Bahrain, and Kuwait), Europe, North America, South-East Asia and Pakistan. It is written using the Bengali script. With about 193 million native and about 230 million total speakers, Bengali is one of the most spoken languages (ranked sixth[1]) in the world. The National song and the National anthem of India, and the National anthem of Bangladesh were composed in Bengali.

Along with other Eastern Indo-Aryan languages, Bengali evolved from the Magadhi Prakrit and Sanskrit languages. It is now the primary language spoken in Bangladesh and is the second most commonly spoken language in India. All the Indo-Aryan languages including Bengali, Hindi, Marathi, Gujrati are called the daughters of Sanskrit.

Question Classification (QC) is an important component of Question Answering System (QAS). The task of a question classifier is to assign one or more class labels, depending on classification strategy, to a given question written in natural language. For example for the question "What London street is the home of British journalism?", the task of question classification is to assign label "Location" to this question, since the answer to this question is a named entity of type "Location". Since we predict the type of the answer, question classification is also referred as answer type prediction. The set of predefined categories which are considered as question classes usually called question taxonomy or answer type taxonomy. Question classification has a key role in automated QA systems. Although different types of QA systems have different architectures, most of them follow a framework in which question classification plays an important role (Voorhees, 2001). Furthermore, it has been shown that the performance of question classification has significant influence on the overall performance of a QA system (Ittycheriah et. al., 2001; Hovy et. al., 2001; Moldovan et. al., 2003).

Basically there are two main motivations for question classification: locating the answer and choosing the search strategy. Knowing the question class not only reduces the search space need to find the answer, it can also find the true answer in a given set of candidate answers. For example, knowing that the class of the question "who was the president of U.S. in 1934?" is of type "human", the answering system should only consider the name entities in candidate passages which is of type "human" and does not need to test all phrases within a passage to see whether it can be an answer or not.

[1] http://en.wikipedia.org/wiki/Bengali_language

On the other hand, question class can also be used to choose the search strategy when the question is reformed to a query over information retrieval (IR) engine. For example, consider the question “What is a pyrotechnic display?”. Identifying that the question class is “definition”, the searching template for locating the answer can be for example “pyrotechnic display is a ...” or “pyrotechnic displays are ...”, which are much better than simply searching by question words.

The remaining part of the paper is organized as follows- section 2 describes different approaches being used in question classification. Section 3 describes available language resources for Bengali language. Section 4 describes the various interrogatives present in Bengali questions. Section 5 describes taxonomies for question types. Section 6 explains the various features used in our work. Section 7 describes the classifiers used in the present work. Section 8 details the experiments conducted on our work and outlines the results. The last section concludes this work and its future work.

2 Related Work

A lot of researches on factoid question classification, question taxonomies, question features and question classifiers have been published continuously until now. Question classification in TREC QA has been intensively studied during the past decade. There are basically two different approaches used to classify questions- one is rule based and another is machine learning based. However, a number of researchers have also used some hybrid approaches which combine rule-based and machine learning based approaches (Huang et. al., 2008; Roy et. al., 2010; Silva et. al., 2011).

Rule based approaches used some manually handcrafted grammar rules to analyze the question to determine the answer type (Hull, 1999; Prager et. al., 1999). Though handcrafted rules have been used successfully for question classification, these approaches however, suffer from the need to define too many rules to determine specific types (Li et. al., 2004). Furthermore, while rule-based approaches may perform well on a particular dataset, they may have quite a poor performance on a new dataset and consequently it is difficult to scale them (Li et. al., 2004). So it is difficult to make a manual classifier with a limited amount of rules.

On the other hand, machine learning-based approaches perform the question classification by extracting some features from questions, train a classifier and predicting the question class using the trained classifier. Many successful learning-based classification approaches have been proposed. Many researchers have employed machine learning methods (e.g., maximum entropy and support vector machine) by using different features, such as syntactic features (Zhang et. al., 2003; Nguyen et. al., 2008) and semantic features (Moschitti et. al., 2007). However, these methods mainly focused on English factoid questions and confined themselves to classify a question into two or a few predefined categories (e.g., "what","how", "why", "when", "where" and so on).

There are also some notable studies that have used both rule-based and machine learning based approaches together. The most successful study (Silva et. al., 2011) that works on question classification, first match the questions with some pre-defined rules and then use the matched rules as features in the machine learning-based classifier. The same approach is used in the work by (Huang et. al., 2008). Machine learning-based and hybrid methods are the most successful approaches on question classification and most of the recent works are based on these approaches.

But in Bengali there is no such Question-Answering system available and this motivates us to classify questions for developing Bengali question-answering system in which user will pose a question in Bengali and also get answer in Bengali.

3 Overview of Language Resource

Compared to the question answering systems in English, because of the specificity on writing and grammar; and the lack of basic language processing resources Bengali question answering system is in development stage. Also, the availability of the experimentation corpus is very rare in the web.

Our classification work in Bengali uses Bengali Shallow Parser which is developed as part of the IL-ILMT Consortium [2]. The shallow parser gives the analysis of a sentence in terms of morphological analysis, POS tagging, Chunking, etc. Apart from the final output, intermediate output of individual modules is also available. All outputs are in Shakti Standard Format (SSF)[3].

4 Interrogatives in Bengali

People could determine the question type by the interrogative present in the question, such as the word 'why' in 'Why are you late?' describes that someone asks the reason. But not all questions type can be determined only by the interrogative. Bengali interrogatives not only describe important information about expected answer but also indicate the Number representations, i.e.- singular or plural.

Unlike English language there are many interrogatives present in the Bengali language. We have been classified it in three categories-

a) Simple Interrogative(SI) or Unit Interrogative(UI)

b) Dual Interrogative(DI)

c) Compound/Composite Interrogative(CI)

4.1 Simple Interrogatives or Unit Interrogatives

It is made up of a single interrogative word which can be considered an Interrogative unit. Further, a SI can be classified into two cases according to answer indication Number Representation. A SI can be indicating a Singular Answer (SA), Plural Answer (PA). If it indicates a SA then it is considered Singular Simple Interrogative (SSI) or Singular Unit Interrogative (SUI), otherwise it indicates a PA and it is considered Plural Single Interrogative (PSI). Sometimes SI indicates both (SA and PA) and sometimes it plays a neutral role. So, SI also can be considered as BSI (both) and NSI (neutral). Therefore, we have found four sub-categories of SI, i.e., SSI, PSI, BSI and NSI.

For example, SI/UI: কে('ke'), কারা ('kara'), কাদের ('kader'), কাহাকে ('kahake').

SSI/SUI: কে ('ke'), কাহাকে ('kahake') ;

PSI/PUI: কারা('kara'), কাদের('kader').

[2] http:// ltrc.iiit.ac.in/analyzer/bengali/
[3] http:// ltrc.iiit.ac.in/mtpil2012/Data/ssf-guide.pdf

BSI: কোন (‘*kon*’), কত(‘koto’), কয়টি(‘koiti’);

NSI:কিভাবে(‘kivabe’),কেন (keno).

4.2 Dual Interrogatives

Each dual interrogative (DI) is made up of using an SI/UI twice. But, all the SI/UI cannot be used to make DI. All the SSI/SUI can be used twice in a question to make DI.

For example,

DI: ‘কে কে’ (‘ke ke’); using SSI কে (‘ke’)

DI: ‘কার কার’ (‘kar kar’); using SSI কার (‘kar’)

DI: ‘কি কি’ (‘ki ki’); using SSI কি (‘ki’)

Although a DI is consisting of one SSI twice, but each DI indicates Plural Answer (PA) only. So, কে(‘ke’)indicates SA , but ‘কে কে’ (‘ke ke’) indicates PA. This implies that all DIs are implicitly PA.

4.3 Compound / Composite Interrogatives

Each compound interrogative (CI) is made up of using multiple Simple Interrogatives. As the CI is formed for getting multiple answers, so it is difficult to categorize it into SA or PA. Also, for this sort of questions simplification is needed. We have found only six CIs from corpus.

CI = { কে কবে, কারা কবে, কে কার, কবে কার, কে কখন, কে কোন }

Bengali Interrogatives are shown in Table-1.

Sl. No	Interrogative (Bengali)	Category	Number Representation
1	কে (ke)	SSI	Singular
2	কাকে (kake)	SSI	Singular
3	কাহাকে (kahake)	SSI	Singular
4	কে কে (ke ke)	PDI	Plural
5	কারা (kara)	PSI	Plural
6	কার (kar)	SSI	Singular
7	কার কার (kar kar)	DI	Plural

8	কাদের (kader)	PSI	Plural
9	কোন (kon)	BSI	Singular/Plural
10	কোন কোন (kon kon)	DI	Plural
11	কি (ki)	NSI	Neutral
12	কি কি (ki ki)	DI	Plural
13	কত (koto)	BSI	Singular/Plural
14	কয়টি (koiti)	BSI	Singular/Plural
15	কখন (kokhon)	NSI	Neutral
16	কোথায় (kothai)	NSI	Singular
17	কবে (kobe)	NSI	Neutral
18	কেন (keno)	NSI	Neutral
19	কিভাবে (kivabe)	NSI	Neutral
20	কেমন (kemon)	NSI	Neutral
21	কে কবে (ke kobe)	CI	Singular
22	কারা কবে (kara kobe)	CI	Plural
23	কে কখন (ke kokhon)	CI	Singular
24	কে কার (ke kar)	CI	Singular
25	কবে কার (kobe kar)	CI	Singular
26	কে কোন (ke kon)	CI	Singular

Table 1-Bengali Interrogatives

5 Question Type Taxonomies

The set of question categories (classes) are referred as question taxonomies or question ontology. Though different question taxonomies have been proposed in different works, but most of the recent learning-based and hybrid approaches are based on two layer taxonomy proposed by Li and Roth (Li et. al., 2002). This taxonomy consists of six course-grained classes and fifty fine-grained classes. The taxonomy proposed by Hermjakob (Hermjakob et al., 2002) consists of 180 classes which is the broadest question taxonomy proposed until now.

As Bengali question classification is at early stage of development, so for simplicity we used single-layer taxonomy for Bengali question type which consists of only eight course-grained classes and no fine-grained classes. Also, we do not consider two more classes namely list and yes-no-explain which have been introduced by Metzler and Croft (Metzler et. al., 2005). Table-2 lists this taxonomy.

Type	Description
PER	Person name i.e., name of human beings
ORG	Organization e.g., office, company etc.
LOC	Location related questions e.g., country , district, place etc.
TEM	Temporal e.g., date, time, year i.e., time related
NUM	Numerical e.g., statistical related questions
METH	Method e.g., procedure related questions
REA	Reason e.g., why related questions
DEF	Definition related questions
MISC	Miscellaneous ; river, mountain, hormone, bird, metal etc.

Table 2- Bengali Question Taxonomies

6 Features

In the task of question classification, there is always an important problem to decide the optimal set of features to train the classifiers. Different studies extracted various features with different approaches and the features in question classification task can be categorized into 3 different types: lexical, syntactical and semantic features (Loni, 2011). We also used three types of features used for question classification.

Loni and others (Loni et. al., 2011) also represented a question in question classification task similar to document representation in vector space model, i.e., a question is a vector which is described by the words inside it.

Therefore a question Q can be represented as:

$$Q= (W_1, W_2, \dots , W_N)$$

Where,

W_I = frequency of term I in question Q;

N = total number of Term

Due to sparseness of feature vector only non-zero valued features are kept in feature vector. Therefore the size of samples is quite small despite the huge size of feature space. All lexical, syntactical and semantic features can be added to feature space and expand the above feature vector. The next subsections describe the features used for Bengali question classification.

6.1 Lexical Features

Lexical features of a question are generally extracted based on the context words of the question, i.e., the words which appear in a question. We have used five lexical features as below-

wh-word, *wh-word position* and *wh-type*: Question's *wh-word* or interrogative is one of the important lexical features and Huang (Huang et. al., 2008; Huang et. al., 2009) has shown that considering question *wh-words* as a feature can improve the performance of classification for English. As the free-word-order" nature of the Bengali language, the position of the wh-word has also been considered as another lexical feature. We considered the value of this feature according to the position {first, middle, last} in given question. We have also considered the interrogative type (*WH-type*) as another lexical feature.

Question length: (Blunsom et. al., 2006) introduced question's length as an important lexical feature which is simply the number of words in a question. We also considered this feature for Bengali classification.

End marker: End marker plays an important role in Bengali question classification that is either "?" or "।" in Bengali. If the end marker is "।", then it has been observed from the experimental corpus that the given question is definition question.

Word shape: Word shapes refer to apparent properties of single words. (Huang et. al., 2008) introduced five categories for word shapes: all digits, lower case, upper case, mixed and other. Word shapes alone is not a good feature set for question classification, but when they combined with other kind of features they usually improve the accuracy of classification(Huang et. al., 2008; Loni et. al., 2011). Capitalization feature is not present in Bengali; so we have considered the other three categories i.e., all digit, mixed and other.

Example: কে(*ke*) গৌড়(*goura*) প্রতিষ্ঠা(*protistha*) করেন(*Karen*) ?

Lexical features: *wh-word*: কে ; *wh-word position*: first ; *wh-type*: SSI; *question length*: 5; *end-marker*: ?

6.2 Syntactical Features

Different works extracted several syntactical features with different approaches. The most common syntactical features are Part of Speech (POS) tags and head words (Loni et. al., 2011).

POS tags: This indicate the part-of-speech tag of each word in a question such as NN (Noun), NP (Noun Phrase), VP (Verb Phrase), JJ (adjective), and etc.

We have added all POS tags of question in feature vector. Similar approach has been successfully used for English (Li and Roth, 2004; Blunsom et. al., 2006). This feature space sometimes referred as bag-of-pos tags. (Loni et. al., 2011) introduced a feature namely tagged unigram which is simply the unigrams augmented with pos tags. Considering the tagged unigrams instead of normal unigrams can help the classifier to distinguish a word with different tags as two different features (Loni et. al., 2011).

Head words: A head word is usually defined as the most informative word in a question or a word that specifies the object that question seeks (Huang et. al., 2008). Correctly identified headword can significantly improve the classification accuracy since it is the most informative word in the question. For example for the question "What is the oldest city in Canada?" the headword is "city". The word "city" in this question can highly contribute the classifier to classify this question as "LOC:city".

Extracting question's headword is quite a challenging problem and there is no research has been conducted so far for Bengali. But, we have considered three cases based on the position of question-word or interrogative in the question-

Case I: if *question-word* appears at beginning, then the first NP chunk after the question-word will be considered as head-word. For example-

কে(*ke*) গৌড়(*goura*) প্রতিষ্ঠা(*protistha*) করেন(*Karen*) ?

WQ NNP NN VM SYM

So, in the above example গৌড়(*goura*) is the head-word.

Case II: if the position of the *question-word* is in between of the question, then the immediate NP-chunk before the question-word will be considered as head-word. For example-

গৌড়*(goura)* কোথায়*(kothai)* অবস্থিত*(obosthita)* ?

NNP WQ JJ SYM

So, in the above example গৌড়(*goura*) is the head-word.

Case III: if *question-word* appears at last i.e., just before end marker, then the immediate NP-chunk before the question-word will be considered as head-word. For example-

বাংলাদেশে*(bangladeshe)* অর্থনীতি*(arthoniti)* কলেজ*(kolege)* কয়টি*(koiti)* ?

NNP (NN NN NN) WQ SYM

So, in the above example বাংলাদেশে*(bangladeshe)* অর্থনীতি*(arthoniti)* কলেজ*(kolege)* is the head-word

Now, if we consider the following example-

কে(*ke*) গৌড়(*goura*) প্রতিষ্ঠা(*protistha*) করেন(*Karen*) ?

Then, the syntactic features will be: [{WQ, 1}, {NNP, 1}, {NN, 1}, {VM, 1}]

6.3 Semantic Features

Semantic features can be extracted based on the semantic meaning of the words in a question. We have used *related word* and *named entities* as semantic features.

Related word: In the absence of Bengali WordNet a Bengali to Bengali dictionary[4] has been used to retrieve the related words. We have manually prepared three *related word* categories by analyzing the training data. The lists are as below-

date :{ জন্মদিন, জন্মতারিখ, দিন, দশক, ঘন্টা, সপ্তাহ, মাস, বছর... etc}

food :{ খাবার, মাছ, খাদ্য, মাখন, ফল, আলু, মিষ্টি, স্বাদ...etc}

human_authority :{নরপতি, রাজা, প্রধানমন্ত্রী, বিচারপতি, মহাপরিচালক, চেয়ারম্যান, জেনারেল, সুলতান, সম্রাট, মহাধ্যক্ষ...etc}

If a question word belongs to any category, then its category name will be added in the future vector.

কে(*ke*) গৌড়ের(*gourer*) স্বাধীন(swadhin) নরপতি(narapoti)[*human-authority*] ছিলেন(*chilen*) ?

For the above example the semantic feature can be added to the feature vector as: [{human-authority, 1}]

Named entities: Some studies (Li and Roth, 2004; Blunsom et. al., 2006) successfully used *named entities* as semantic feature. To identify the Bengali named entities in question text a Hidden Markov Model Based Named Entity Recognizer (NER) System (Ekbal et. al., 2007) has been used as Bengali NER system.

কে(*ke*) গৌড়(*goura*)[*Location*] প্রতিষ্ঠা(*protistha*) করেন(k*aren*) ?

For the above example the semantic feature can be added to the feature vector as: [{Location, 1}]

7 Classification Module

Though many supervised learning approaches have been proposed for question classification (Li et. al., 2002; Blunsom et. al., 2006; Huang et. al., 2008), but these approaches mainly differ in the classifier they use and the features they extract (Loni, 2011). We assume that a Bengali question is unambiguous, i.e., a question has only one class. So, we assign one label to a given question and can be described as follows-

C = {c1, c2, c3, c4, c5, c6, c7, c8};

Where, C is the set of possible classes

Q = {Q1, Q2, Q3… QN-1, QN};

Where, Q is the set of N given questions

The task of our question classifier is to assign the most likely class C_k to a question Q_m . Recent studies (Zhang et. al., 2003; Huang et. al., 2008; Silva et. al., 2011) also consider one label for one question.

[4] http://dsal.uchicago.edu/dictionaries/biswas-bangala/

We have used Naive Bayes (NB), Kernel Naïve Bayes (KNB), Decision Tree (DT) and Rule Induction (RI) and DT has been performed the best among them.

7.1 Naïve Bayes (NB)

Naïve Bayes (NB) classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions, i.e. assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable.

Using the simplest assumption of a constant prior distribution, Bayes theorem leads to a straightforward relationship between conditional probabilities. Given a class label C with *m* classes, $c_1, c_2, ..., cm$, and an attribute vector *x* of all other attributes, the conditional probability of class label c_i can be expressed as follows:

$$P(C = c_i \mid x) = \frac{P(x \mid C = c_i)P(C = c_i)}{P(x)}$$

Where $P(C=c_i)$ is the probability of class label c_i and can be estimated from the data directly. The probability of a particular unknown sample, $P(x)$, does not have to be calculated because it does not depend on the class label and the class with highest probability can be determined without its knowledge.

7.2 Kernel Naïve Bayes (KNB)

Kernel Naïve Bayes classifier is modified version of NB classifier that uses estimated kernel density. Conditional probability $P(x \mid C = c_i)$ can be written as a kernel density estimate for class c_i

$$P(x \mid C = c_i) = f_i(x) \qquad \text{And} \qquad f_i(x) = \sum_{t=1}^{n} K_i(x, x_t);$$

Where, x_t are training points and $K_i(x,x_t)$ is a kernel function.

7.3 Rule Induction

Rule Induction (RI) learns a pruned set of rules with respect to the information gain. It works similar to the propositional rule learner named Repeated Incremental Pruning to Produce Error Reduction (RIPPER, Cohen 1995). Starting with the less prevalent classes, the algorithm iteratively grows and prunes rules until there are no positive examples left or the error rate is greater than 50%.

In the growing phase, for each rule greedily conditions are added to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain.

In the prune phase, for each rule any final sequences of the antecedents is pruned with the pruning metric p/(p+n).

7.4 Decision Tree

Decision trees are powerful classification methods which often can also easily be understood. In order to classify an example, the tree is traversed bottom-down. Every node in a decision tree is labelled with an attribute. The example's value for this attribute determines which of the out coming edges is taken. For nominal attributes, we have one outgoing edge per possible attribute value, and for numerical attributes the outgoing edges are labelled with disjoint ranges. This decision tree learner works similar to Quinlan's C4.5 or CART.

8 Experimentations and Results

8.1 Corpus

Though Bengali is one of the most spoken languages in the world, but there is no standard questions data available. So, we had to collected questions data from the web and we had selected the questions of different domains e.g., education, geography, history, science etc. available in BCSTAT.COM[5]. 1100 questions have been selected and processed to extract the features. Bengali shallow parser has been used to obtain the part of speech (POS). Two high qualified human annotators have been labelled the questions with an agreement score of 95.93%. We have used 770 questions (70%) for training and rest 330 questions (30%) to test the classification models.

8.2 Experiments

We have used four models i.e., Naive Bayes (NB), Kernel Naïve Bayes (KNB), Decision Tree (DT) and Rule Induction (RI) and used well known widely used Rapid Miner[6] Tool for experimentation. Performance of any classifier needs to be tested with some metric to assess the results. In our study, *classification accuracy* has been used to evaluate the results of the experiments.

Accuracy and *error* are widely used metrics to determine class discrimination ability of classifiers, and calculated using the following equation-

$$accuracy(\%) = \frac{TP + TN}{P + N}$$

$$error(\%) = 100 - accuracy$$

Where, TP = true positive samples; TN = true negative samples

P = positive samples; N = negative samples

It is a primary metric in evaluating classifier performances and it is defined as the percentage of test samples that are correctly classified by the algorithm.

Initially we have been only considering the lexical features of the questions. Naïve Bayes (NB) has been used as Baseline system for our experiment with classification accuracy of 80.65%. It has been found from the experiments that performance of baseline system drastically fall on ORG class (Precision-14.41%, Recall-41.18%) and MTHD class (Precision-34.62%, Recall-75.27%).

5 http://www.bcstest.com
6 http://www/rapid-i.com

Though, KNB classifier increased the accuracy but failed to produce better performance on ORG and MTHD classes. Rule Induction classifier not only increased the accuracy (83.31%) but also performed well on ORG and MTHD classes. Decision Tree has been performed the best among all classifiers (*accuracy* 84.19%) and has been exceptionally performed well on ORG, MTHD and others classes. The detail results have been shown on Table-3.

Features	Classifier	Accuracy	Error
Lexical	NB (*Baseline*)	80.65%	19.35%
	KNB	81.09%	18.91%
	RI	83.31%	16.69%
	DT	84.19%	15.81%

Table 3- Classifiers Performance

Next we have used lexical and semantic features together and applied NB, KNB, RI and DT classifiers respectively. It has been noted from experimental results that inclusion of semantic features improves the performance of all the said classifiers. The experiment results have been illustrated in table-4.

Features	Classifier	Accuracy	Error
Lexical + *Syntactical*	NB	81.34%	18.66%
	KNB	82.37%	17.63%
	RI	84.23%	15.77%
	DT	85.69%	14.31%

Table 4- Classifiers Performance

Finally, we have used lexical, syntactical and semantic features altogether and applied the four classifiers. Use of semantic features improves the performance of K-NB, NB classifiers on handling ORG and MTHD classes.

After inclusion of three features, NB classifier has been outperformed DT classifier handling NUM classes and RI classifier has been outperformed DT classifier handling TEMP classes. But overall DT classifier (*accuracy* 87.63%) has been performed well on classifying Bengali Questions. The detail results have been shown on Table-5.

Features	Classifier	Accuracy	Error
Lexical + *Syntactical* + *Semantic*	NB	81.89%	18.11%
	KNB	83.21%	16.79%
	RI	85.57%	14.43%
	DT	87.63%	12.37%

Table 5- Classifiers Performance

Conclusion and perspectives

This paper presents our research work on automatic question classification through machine learning approaches. The main contributions of this paper are as follows-

- We have studied the interrogatives and categorized them into three categories. We have also extracted the probable number representation i.e., singular or plural for each Bengali interrogative. 26 interrogatives have been identified from the experimented corpus.
- The baseline system based on Naïve Bayes classifier (using only lexical features) has achieved 80.65% accuracy. We have investigated the lexical, syntactic and semantic features for Bengali questions and the identified features performed well (achieved accuracy up to 87.63%) on Bengali Questions.
- We have experimented on four machine learning classifiers and shown that overall Decision Tree outperforms NB, KNB, RI methods for Bengali question classification.

The main future direction of our research is to exploit other lexical, semantic and syntactic features for Bengali question classification. In future an investigation can be performed on including new interrogatives using a large corpus. It may increase the count of Bengali interrogatives, particularly DI and CI. It is also worth investigating other types of machine learning algorithms. In the current work, we have prepared only three related word categories. So, the model performance can be improved in future by identifying new suitable categories.

References

Voorhees, E. M.(2001). Overview of the trec 2001 question answering track. *In Proceedings of the Tenth Text REtrieval Conference (TREC)*, pages 42–51.

Ittycheriah A., Franz, M., Zhu, W. J., Ratnaparkhi, A. and Mammone, R. J. (2001).*IBM's statistical question answering system. In Proceedings of the 9th TREC*, NIST.

Hovy, E., Gerber, L., Hermjakob, U., Lin, C. and Ravichandran, D. (2001). Toward semantics-based answer pinpointing .

Moldovan, D., Pa¸sca, M., Harabagiu, S. and Surdeanu, M. (2003). Performance issues and error analysis in an open-domain question answering system.*ACM Trans. Inf. Syst.*, 21:133–154.

Huang, Z., Thint, M. and Qin. Z. (2008). Question classification using head words and their hypernyms. *In Proceedings of EMNLP*, pages 927–936.

Ray, S. K., Singh, S. and B. P. Joshi. (2010). A semantic approach for question classification using wordnet and wikipedia. Pattern Recogn. *Lett.*,31:1935–1943.

Silva, J., Coheur, L., Mendes, A. and Wichert, A. (2011). From symbolic to sub-symbolic information in question classification. *Artifcial Intelligence Review*, 35(2):137–154.

Hull, D. A. (1999). Xerox TREC-8 question answering track report. *In Voorhees and Harman.*

Prager, J., Radev, D., Brown, E. and Coden, A. (1999). The use of predictive annotation for question answering in trec8. *In NIST Special Publication 500-246:TREC8*, pages 399–411.NIST.

Li, X. and Roth, D. (2004). Learning question classifiers: The role of semantic information. *In Proc. International Conference on Computational Linguistics (COLING)*, pages 556–562.

Zhang,D. and Lee,W.S. (2003). Question classification using support vector machines *In SIGIR.*

Nguyen, T., Nguyen, L., Shimazu, A. (2008). Using semi-supervised learning for question classification. *Journal of Natural Language Processing*, 15(1):3–21.

Moschitti, A., Quarteroni, S., Basili, R. and Manandhar, S. (2007). Exploiting syntactic and shallow semantic kernels for question/answer classification. *In ACL*, pages 776–783.

Li, X., and Roth, D. (2002). Learning question classifiers. *In Proceedings of the 19th international conference on Computational linguistics, COLING '02*, pages 1–7. Association for Computational Linguistics.

Hermjakob,U., Hovy, E. and Lin, C. (2002). Automated question answering in webclopedia - a demonstration. *In Proceedings of ACL-02.*

Metzler, D. and Croft, W. B. (2005). Analysis of statistical question classification for fact-based questions. *Inf. Retr.*, 8:481–504.

Loni, B., Tulder,G., Wiggers, P., Loog, M. And Tax, D. (2011).Question classification with weighted combination of lexical, syntactical and semantic features. *In Proceedings of the 15th international conference of Text,Dialog and Speech*.

Huang, Z., Thint, M. and Celikyilmaz, A. (2009). Investigation of question classifier in question answering. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, *(EMNLP '09)*, pages 543–550.

Blunsom, P., Kocik, K. and Curran, J. R. (2006). Question classification with log-linear models. *In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '06*, pages 615–616, NY, USA, ACM.

Loni, B. (2011). A Survey of State-of-the-Art Methods on Question Classification, *Literature Survey, Published on TU Delft Repository*.

Ekbal, A. and Bandyopadhyay, S. (2007). A Hidden Markov Model Based Named Entity Recognition System: Bengali and Hindi as Case Studies. *PReMI 2007*: 545-552.