Abstract—Question classification is very important for question answering system. The features include words, bigrams, and trigrams, as well as the parts of speech in each sentence. Many approaches to question classification have been proposed and have achieved reasonable results. Naïve bayes machine learning approach is most common approach for question classification. The dominant approaches are machine learning and context based classification. This paper presents our research work on automatic question classification through naïve bayes algorithm machine learning approaches. In this paper we also discuss the research work done in this field.

Index Terms: Algorithms, Experimentation Question Answering, Text Classification, Machine Learning, Naïve Bayes.

I. INTRODUCTION

With the emergence of question answering systems the rate of posting and answering questions by people increased. As a result of which system accumulated large number of questions. Hence, it is necessary to organize these questions in a good way. Question categorization is a technique used for this purpose. Question Categorization, is a useful technique in Web-based Question Answering system. On the basis of the questions, it will be associated to the corresponding category. Earlier approaches for the creation of automatic document classifiers consisted of manually building, by means of knowledge engineering techniques, an expert system capable of taking Document Categorization decisions. [1], the major disadvantage of which was that it required rules manually defined by a knowledge engineer with the aid of a domain expert. Another problem is that when the classifier is ported to a completely different domain, concerned domain expert need to intervene and the work has to be done in its entirety. To overcome the pitfalls associated with rule-based classification „Machine Learning” techniques are currently applied for these purposes. In this approach set of pre classified questions are fed to the classifier. This acts as the training example for the classifier. Based on these examples the classifier will classify the future samples. The common text classifiers which employ these approaches include probabilistic classifiers, decision tree classifiers, decision rule classifiers, regression based classifiers, neural network based classifiers, and SVM based classifiers [7]. Another approach that can be taken is context based interpretation. It takes advantage of tracking the contextual meaning of words and phrases during (and after) the development of ontology for that context, and subsequently uses this information as knowledge base for interpretation of free text sentences. The classifiers will be trained by set of training examples for each category, which are predefined. Hence forth the classifiers will be used to classify set of questions. The performance evaluation of classifiers using both approaches, SVM and Naïve Bayes, is conducted. The paper describes about Naïve Bayes Classifier and the algorithm which has been employed for question categorization using this approach.

Fig.1: High-Level Architecture of a Question Answering System [2]

II. CHALLENGES IN QUESTION ANSWERING SYSTEM

Building a question answering system is not an easy task. QA systems typically require a great deal of human effort, in order to create linguistic rules that can cope with the vast variety of questions that can be asked, and the many different ways in which they can be formulated. For example, consider the question “What is the capital of India?”, and the different reasonable 1 formulations of the same question:

- Name the capital of India.
- What is India capital city?
- What’s the name of the capital of India?
- Which city is the capital of India?

As can be observed, such a simple question can be formulated in many different ways, making it a very laborious task to manually build specific rules for each different formulation. Moreover, to aggravate the problem, the
passages where the answer to any of the above questions may be found can, themselves, be phrased in many different ways:

• Delhi, the capital of India.
• The capital of India is Delhi.
• Delhi became the capital city of India

Thus, systems that follow a rule-based strategy tend to target a specific language, resulting in systems that are very difficult to port to different languages. Made use of a great deal of hand-crafted linguistic patterns, which are specific to the Portuguese language, resulting in a system that is difficult to extend and adapt for different languages. Furthermore, there are also challenges inherent to question answering for the Web. For example, although the Web contains vast amounts of information, it is crucial to provide the right queries to the Web information source, or else we can receive noisy information that will lead the system to extract wrong answers.

III. QUESTION CLASSIFICATION

Question classification can loosely be defined as the task of given a question (represented by a set of features), assign the question to a single or a set of categories (answer types). Adopting the formal definition of text categorization to the problem of question classification, the task can be defined as follows: Question classification is the task of assigning a Boolean value to each pair \( q_j, c \), i.e., \( 2 \times Q \times C \), where \( Q \) is the domain of questions and \( C = \{ c_1, c_2, \ldots, c|C| \} \) is a set of predefined categories. The task therefore requires taxonomy of answer types according to which questions should be categorized on the one hand, and a means for actually making this classification on the other.

Table 1. Li & Roth’s two-layer taxonomy for question classification [13].

<table>
<thead>
<tr>
<th>Coarse</th>
<th>Fine</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABBREVIATION</td>
<td>abbreviation, expansion</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>definition, description, manner, reason</td>
</tr>
<tr>
<td>ENTITY</td>
<td>animal, body, color, creative, currency, medical disease, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word</td>
</tr>
<tr>
<td>HUMAN</td>
<td>description, group, individual, title</td>
</tr>
<tr>
<td>LOCATION</td>
<td>city, country, mountain, other, state</td>
</tr>
<tr>
<td>NUMERIC</td>
<td>code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight</td>
</tr>
</tbody>
</table>

IV. NAÏVE BAYES CLASSIFIER

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes’ theorem with strong independence

Assumptions [1]. Bayes theorem can be stated as follows

\[
p(c_j|q) = \frac{p(c_j)p(q|c_j)}{p(q)} \tag{1}
\]

\[ P(q) \]

Where \( P(C_k|q_j) \) is the posterior probability,
\[ P(C_k) \]

is the prior probability,
\[ P(q|C_k) \]

is the likelihood and \( P(q) \) is the evidence.

A naive Bayes classifier follows conditional independence since it 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. Thus terms are given a weight value which is independent of its position and presence of other terms. Naive Bayes classifier is trained by set of labeled training examples. Thus it is said to undergo supervised learning.

In the experiments carried out by (Zhang & Lee, 2003), a naive Bayesian classifier was used for the task of question classification, trained on the standard data set for question classification of Li & Roth. A bag-of-words model was used to represent a question as an unordered collection of words. Within this model, a dictionary is created, whose length is equal to the number of distinct words present in the training set. After that, questions can be encoded as binary feature vectors \( x \), in such a way that if the question contains the \( i \)-th word of the dictionary, then \( x_i \) is set to 1; otherwise, \( x_i \) is set to 0. To exemplify this model, consider the questions “What is the capital of India?” and “Where was Mahatma Gandhi born?”, as well as the resultant word dictionary:

[What, was, is, where, the, Mahatma Gandhi, born, capital, of, india]

A binary feature vector representation 15 for the question “What is the capital of India?”, would then be:

\[ [1, 0, 1, 0, 1, 0, 1, 1, 1] \]

Another kind of feature – bag-of-n grams – was also used in the experiments, in order to try to capture some dependencies between words, such as the order in which they appear. With the bag-of-n grams features, the classifier showed a slight improvement of 5.8% in terms of accuracy, over the bag-of-words features. The experimental results showed that naive Bayesian classifiers trained on 5,500 examples of labeled questions, perform relatively well on the task of question classification, with an accuracy of 83.2% under the coarse grained category. However, the results also showed that in order to achieve this accuracy, the training sets need to be quite large, since when the classifier was trained with just 1,000 examples, the accuracy dropped to merely 53.8%. Additionally, it is important to keep in mind that these results were obtained by using very superficial text features, which leaves some room for improvement.

V. NAÏVE BAYES ALGORITHM

Naive Bayes algorithm proposed by us proceeds by finding out the feature vectors associated with each category. Feature vector includes common terms occurring in questions pertaining to one particular category expressed in terms of
their weight or relevance in particular question. Common terms are found out by stop-word elimination, stemming and then pruning (eliminating words with frequency below a particular range and frequency above a particular range). Weight of a term in a category, associated with a particular question can be found out using weight calculation methods. Here the weight is calculated by finding the entropy associated with the term [9]. This is given by

\[ n_i = 1 + \log \left( \frac{N}{n_i} \right) \log f_i \]  

(2)

Where \( n_i \) is the weight associated with word \( i \), \( N \) denotes total number of categories, \( f_i \) frequency of word \( i \) in question \( t \), \( n_i \) total number of occurrences of word \( i \) in all questions.

The next phase of algorithm is to classify a new question. The probability of the question to belong to all categories are found out and the category for which it has maximum posterior probability is the one to which the question is assigned to.

\[ C_k = \text{argmax}_k p(c_k | q_j) \]  

(3)

Where \( C_k \) is the category with maximum posterior probability. The probability of a question \( q_j \) to belong to a category \( c_k \) is given by eq.(1). Since Naive Bayes classifier assumes conditional independence \( P(q_j | c_k) \) can be given as

\[ p(q_j/c_k) = \prod p(a/c_k) \]  

(4)

\[ p(a/c_k) = p_i \cdot (1 - p_i)^{1-a} \]  

(5)

The weight calculation followed here is different from ordinary tf-idf method followed for text classification. The classification using the above said approach of weight calculation showed increased accuracy. The accuracy was found to increase with increase in number of training data. To speed up the calculation of weight the terms have been saved in database. MySQL database has been used. Index has been provided which further improved the ease of retrieval thus reducing the classification time.

**VI. RELATED WORK**

Zhang & Lee (2003)[7] performed a number of experiments on question classification using the same taxonomy as Li & Roth as well as the same training and testing data. In an initial experiment they compared different machine learning approaches with regards to the question classification problem: Nearest Neighbors (NN), Naïve Bayes (NB), Decision Trees (DT), SNoW, and SVM. NN, NB, and DT are by now fairly standard techniques and good descriptions of them can be found in for instance Mitchell (1997). The feature extracted and used as input to the machine learning algorithms in the initial experiment was bag-of-words and bag-of-n-grams (all continuous word sequences in the question). Questions were represented as binary feature vectors since the term frequency of each word or \( n \) gram in a question usually is 0 or 110. The results of the experiments are shown in table 6.1.[7]

### Table 2: Results from Zhang & Lee (2003)[7].

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>bag-of-words</th>
<th>bag-of-n grams</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coarse</td>
<td>fine</td>
</tr>
<tr>
<td>NN</td>
<td>75.6</td>
<td>68.4</td>
</tr>
<tr>
<td>NB</td>
<td>77.4</td>
<td>58.4</td>
</tr>
<tr>
<td>DT</td>
<td>84.2</td>
<td>77.0</td>
</tr>
<tr>
<td>SNoW</td>
<td>66.8</td>
<td>74.0</td>
</tr>
<tr>
<td>SVM</td>
<td>85.8</td>
<td>80.2</td>
</tr>
</tbody>
</table>

The results for the SVM algorithm presented in table 6.1 are when the linear kernel is used. This kernel had as good performance as the polynomial, RBF, and sigmoid kernels (Zhang & Lee 2003b).

The paper entitled “An empirical study of naïve bayes classifier” proposes naïve bayes machine learning algorithm for question answer system and finds that problem in this technique is unrealistic independence assumption.

João Pedro Carlos Gomes Silva (2009)[2]

In this project three machine learning algorithm(KNN,Naive bayes,SVM) used and with the use of different parameter conclude that SVM is best method for question classification also suggested potential improvements that could be made to the system developed in this work.

Lotfi A. Zadeh the paper entitled “From Search Engines to Question Answering Systems”(2006)[11]

- Developed Existing search engines, with Google at the top, have many truly remarkable capabilities. But there is a basic limitation search engines do not have deduction capability
- In this perspective, a search engine may be viewed as a semi-mechanized question-answering system.
- To achieve upgrading, new concepts, ideas and tools are needed to address difficult problems which arise when knowledge has to be dealt with in an environment of imprecision, uncertainty and partial truth.

BY Kadri Hacioglu & Wayne Ward

The paper entitled “Question Classification with Support Vector Machine and Error Correction Codes” (2003)[13] proposes a comparison with related work has shown, also find that SVM learning approach is very promising yielding comparable performance with minimal and even with no linguistic analysis.

Dell Zhang & Wee Sun Lee(2003)[7]

The paper entitled “Question Classification using Support Vector Machines” proposes research work on automatic question classification through machine learning approaches. also show that with only surface text features SVM outperforms four other machine learning methods (KNN, NB, DT, SNoW) for question classification.
ACKNOWLEDGMENT

I thank Prof. Megha Mishra for several invaluable suggestions and the entire SSCET team for help with various Components, feature suggestions and guidance.

REFERENCES


AUTHOR BIOGRAPHY

Rishika Yadav – BE(CSE) from MPCCET ,currently works as asst prof. in SSCET also pursuing ME(CTA) from SSCET. Published 1 paper in national level paper presentation.

Megha Mishra-BE (CSE), ME (CTA), PhD(pursuing) currently works as sr.asst. Prof in SSCET, she published more than 10 papers in national and international journals.