A language based comparison of different similarity functions and classifiers using web based Bilingual Question Answering System developed using Machine Learning Approach

Krishma Singla
Department of Computer Science and Engineering, Banasthali Vidyapith, Rajasthan
krishma32@gmail.com

Mohit Dua
Department of Computer Engineering,National Institute of Technology, Kurukshetra, Haryana
er.mohitdua@gmail.com

Garima Nanda
Department of Computer Science and Engineering, Banasthali Vidyapith, Jaipur, Rajasthan
nanda.garima1004@gmail.com

ABSTRACT
Contemporary information techniques and services offered by the Internet are going through the dilemma of determining and managing an increasing amount of textual information, to which ingress is often difficult. But recently Machine Learning approaches have shown their outstanding performance and elasticity in many applications such as Artificial Intelligence and Pattern Recognition. Question Answering (QA) System is an Information Retrieval system in which the expected response given is directly the answer as requested by the user instead of list of references which have some probability of being the answer. The main intention of this research is to present the knowledge and fetch the answer for a given query by employing machine learning approach. The query will be matched to the knowledge database by computing their similarity. The stated research portrays the Web Based Bilingual Question Answering system constituting of Hindi and English language employing by machine learning approach.

Keywords
Question Answering System; Natural Language; Vector Space Model; Information Retrieval; Classification; Machine Learning; Naïve Bayes; Random Forest.

1. INTRODUCTION
Along with the vibrant growth of amount of readily available knowledge resources on to the search engines, we have invaded a stage where an effectual QA mechanism will become a vital part of our life to ingress the information.

Question Answering Systems are assessed more complicated than Information Retrieval system and require substantial Natural Language Processing techniques to offer a precise answer to a natural language question.

A new practice of designing Natural Language Question Answering was recommended, It comprises of 3 stages; these are Question processing, Document processing and Answer Extraction. Question Processing is a task to examine and classify user query and articulate user requests. Document processing is used to gather someappropriate document set and it will be

attained into one paragraph to answer user's query. Answer Extraction is accountable to select acknowledgement based on relevant fragment from the data.

Classification is well stated as bringing out the exact "Class Label" from the specified input (Query). Essentially in Classification quest, deck of labels is annotated in advance and each input is judged in isolation from all the other inputs given. Queries given by the user in NL are entertained by the QAS, hunts for the accurate answers from a stockpile of catalogue and reflects back with the concise and accurate one. As demonstrated in the given figure 1,Adecisive encouragement is monitored in knowledge based Question Answering systems, that is creating a massive knowledge base with practical and precise facts & domains.

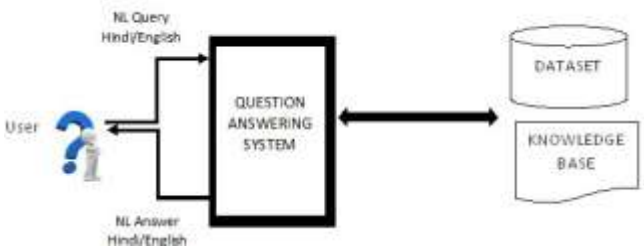


Figure 1. QA System

2. RELATED WORK

Discrete approaches and procedures are used by distinct system to upgrade the performance and to diminish the cost of evolution of Question Answering Systems. User assign different enquiries in natural language so that the precise answers can be revealed. Many researchers engaged into QA system from past many years in different languages just as Chinese, English, Japanese, etc.

Garima et.al developed a Question Answering system based on Machine Learning approach [1], which indicated the opinion of similarity and classification that assign preferably better results by clarifying the overall accuracy of detecting the relevant answers of the particular questions proposed by the user.

Entirely, QA system is a technique of exposing the accurate answers of the user asked questions over an immense collection.

Jovita et.al explained a QA layout based on Vector Space Model [3], in which a document is supposed as a vector that has magnitude and direction. In VSM term is represented by using dimension from vector space. It is clear by this research that an accurate answer can be retrieved from a given question by effectively using VSM.

Number of QA systems worked on the abstract of machine learning such as Support Vector Machine. Machine learning techniques need similarity functions that can calculate analogy. Similarity Computations require mainly for clustering.

Many unfavorable situations were faced by the researchers and users while manipulating the question answering Systems which seem to be decreased along with the proposed approaches.

3. ARCHITECTURE AND PROPOSED WORK

The architecture of Bilingual QA System consists of three Phases as in figure 2:

(i) Accessing NL Query phase.
(ii) Feature Extraction Phase
(iii) Classification Phase

The module of developed QA System composed by two knowledge bases that of English and Hindi Natural Language separately. Each of them holds:

- SWD(Stop Words Database)- this database carries those words which are scrutinized before or after dealing with natural language query. Stop words are basically the ordinary words in a language. For example; is, at, on, of, etc.
- Entities- Entities here refer to the known terms, where they have their own distinct individuality.
- Trained Data- Trained data is a set of queries that are trained for testing the system by the user.

3.1 Accessing NL Query Phase

Firstly the user will decide its language in which the system will receive the query. For accessing the input query the decision making will be done initially; i.e. Accessing NL query Phase. Now the NL query is read, accessed and preprocessed. While moving through the preprocessing level, stop words are dropped out and task of tokenization takes place resulting in tokens.

Example: भारतकीराजधानीक्याहै

Tokens: भारतराजधानीक्या

Example: what is the capital of India
Tokens: what capital India

Where the stop words are: "is, the, of " and in Hindi: "की, है"

3.2 Feature Extraction Phase

At this level, Entity prediction takes place. After reading the query prescribed by the user as input, the preprocessed query is fed and revert tokens. Tokens from the earlier level are further given into the upcoming level where it goes into two different sub phases.

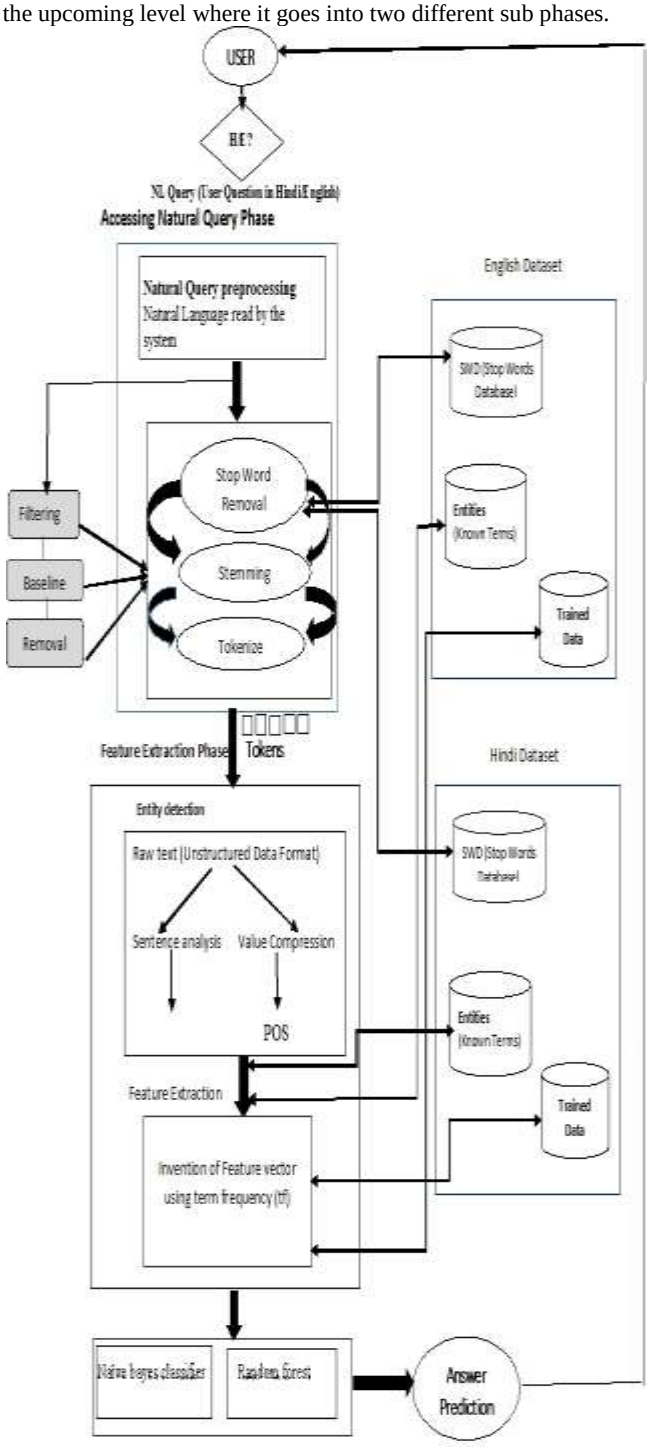


Figure 2. Architecture Module of QA System

3.2.1 Entity detection

Using similarity measures entity Detection takes place. Basically similarity functions defines the similarity between two objects, resulting into entities. The research manages various similarity functions among of which Smith Waterman algorithm is performing best in Hindi and Jaro Winkler in English Natural Language queries **[5].**Instead of observing the total sequence, smith waterman algorithm optimizes the similarity function by measuring the segments of all possible lengths where on the other side Jaro Winkler explains how much the two string are similar among each other.

3.2.2 Feature Vectors

If the data given to an algorithm or any system referring NL systems is immense and it is reckon to be inessential, it is then needed to be transfigured to reduced set of features (merely known as features vector). Regarding expression related to "Term Frequency", tf(t,d), the easiest way is taking general frequency of a term in a text document; that is number of times t is occurring in d; where t is term and d is document.

3.3 Classification Phase

As a result of above level, Feature Vector is input of this phase of Classification where the task of uprooting the correct label of class over the specified query is attained. Naive Baye's Classifier and Random forest is appraised in this QA system. It is keenly required to train it with some data. After training, testing of the data needs to be done and finally performance evaluation is taken in consideration. If the classifier is based on training it must contain the correct label for each input.

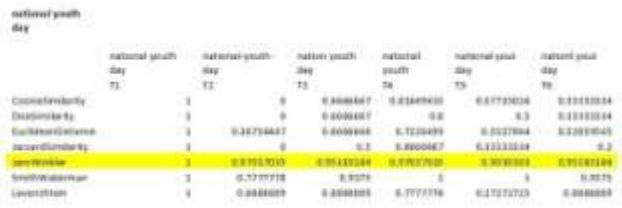
In performance evaluation of NB classifier, some notations for probabilities are used; these are p(c)-this is the prior probability of class c, p'(c) - it is the posterior probability of class c which is returned by the specific classifier.

Random forest is conception of the general technologies of prevalent decision forests. It is efficient for large databases and is magnificent in accuracy among other algorithms.

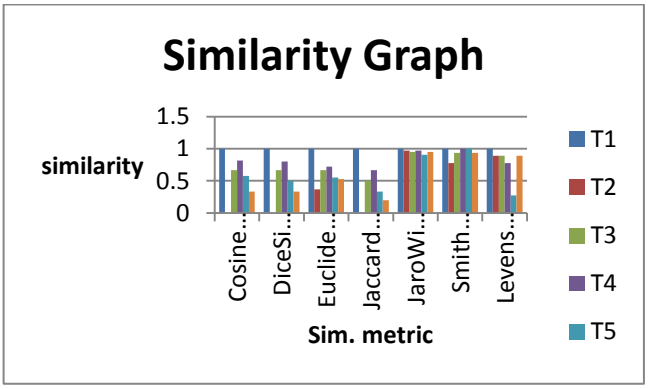
4. EXPERIMENTS AND RESULTS

Accomplishment of QA System is self-sustaining of platform; it can be easily performed on windows as well as Linux. The research has given a user friendly GUI for the Web based QA system. Being web based and bilingual, it is easily operated and manageable by different users. The classifier used in the demonstration of the system is Naïve Bayes Classifier and Random Forest. Text based dataset is merely used for Testing where numerous similarity measures are used. Results are presented here in the analysis concluding it with a similarity graph.

Testing is done on datasets of English and Hindi language. The dataset consists of 100 questions in General Knowledge domain. Queries are fired randomly by the user. Performance is measured by using different test sets and the similarity functions are compared on different basis showing the concise results. Test set is a pile of questions which are tested for checking performance.

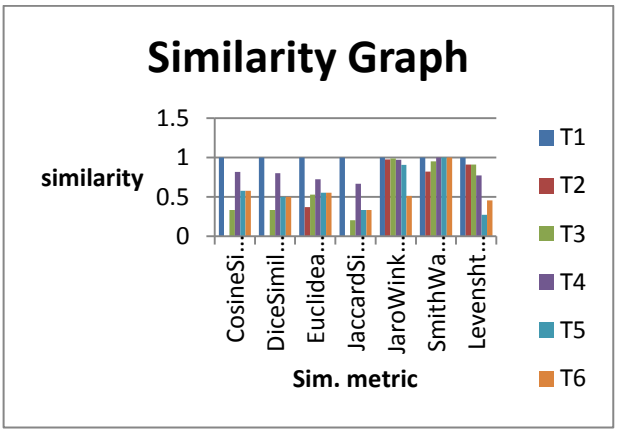


From T1 to T6 in the table are the queries given by the user to check how much the fired query is similar to the actual phrase stored in the knowledge base. Different similarity functions such as cosine similarity, Jaro Winkler, Smith Waterman, etc, are examined on different queries to perceive which among these is magnificent.



The graph is showing that among various similarity functions, jaro winkler is performing at its best. Overall accuracy is calculated.

Similarly, a query fired in hindi language perform differently in similarity functions. Among Cosine, Dice similarity, Jaro Winkler, Smith Waterman, Levensthein; Smith Waterman is efficient.

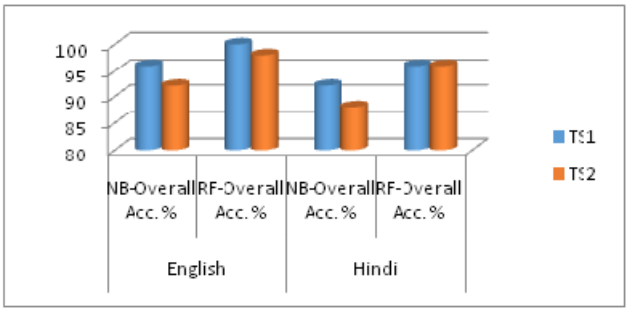


Now the classifiers are compared in both languages showing ultimate results in the above graph.

Table 1. Comparison of Classifiers in terms of Accuracy

	English		Hindi	
Test Set	**NB-Overall Acc %**	**RF-Overall Acc %**	**NB-Overall Acc %**	**RF-Overall Acc %**
TS1	96	100	92	96
TS2	92	98	88	96

TS1 and TS2 are the two Test sets which are used for comparing and finding out the accuracy of different classifiers in both languages. Since Naïve Bayes and Random Forest classifiers are used, their performance is compared individually on English and Hindi data set. The table reflects that the random forest is more accurate than Naïve Bayes showing 100% accuracy in Englishand 96% in Hindi Language on TS1.



Following is the illustration study being done regarding the comparison of time complexity of both stated classifiers in the form of table and graph.

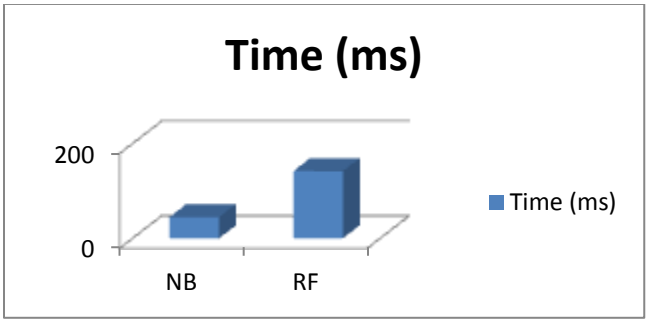
Naïve Bayes classifier takes less time than Random Forest as it consuming 45 ms.

This comparison is done on the test set.

Table 2. Time Comparison of Classifiers

	Naïve Bayes	**Random Forest**
Time (ms)	45	142

The graph reflects the comparison of the time taken by the classifiers.



5. CONCLUSION AND FURTURE WORK

Web based Bilingual Question Answering System for Hindi and English Natural Language presents enormous idea of QA System with Overall accuracy and threshold of 0.9. Overall Accuracy and similarity concepts are used here giving an extensive platform to the user to interrogate a query in natural language and receiving the relevant result in same language. Concepts used here are far better than the beliefs used in previous stated systems. As the future work is scrutinized, this QA system can be implemented as multilingual using different languages using other discrete classification techniques along with different datasets. Further if feature extraction is modified, accuracy would be increased.

6. REFERENCES

[1] Garima Nanda, Krishma Singla and Mohit Dua "A Hindi Question Answering System using Machine Learning approach" (ICCTICT 2016) IEEE, (unpublished).

[2] Sunil A. Khillare, Bharat A. Shelke, and C. NamrataMahender," Comparitive Study on Question Answering Systems and Techniques," International Journal of Advanced Research in Computer Science and Software Engineering, pp. 775-778, Vol. 4, Issue 11, November 2014.

[3] Jovita, Linda, Andrei Hartawan, DerwinSuhartono," Using Vector Space Model in Question Answering System,"International Conference on Computer Science and Computational Intelligence (ICCSCI 2015), ScienceDirect, pp. 305-311.

[4] Asma Ben Abacha a, Pierre Zweigenbaum," MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies," Information Processing and Management, ScienceDirect, pp. 570-594, 2015.

[5] Show-Jane Yen, Yu-ChiehWu, Jie-Chi Yang, Yue-Shi Lee, Chung-Jung Lee, Jui-Jung Liu," A support vector machine-based context-ranking model for question answering," Information Sciences, ScienceDirect, pp. 77-87, 2013.

[6] Rajender Kumar, MohitDua, Shivani Jindal," D-HIRD: Domain-Independent Hindi Language Interface to Relational Database," International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC), IEEE (2014).

[7] Sanjay K Dwivedi, Vaishali Singh, "Research and Reviews in Question answering system," International Conference on Computational Intelligence: Modelling Techniques and Applications, ScienceDirect, pp. 417-424, 2013.

[8] Smith Mahboob Alam Khalid, Valentin Jijkoun and Maarten de Rijke, "Machine Learning for Question Answering from Tabular Data," 18th International Workshop on Database and Expert Systems Applications, IEEE, 2007.

[9] Er. Amit Chaudhary, Er. AnnuBattan," Natural Language Interface to Databases-An Introduction," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 7, July 2014.

[10] Sneha Bagde, Mohit Dua, and Zorawar Singh Virk, " Comparison of Different Similarity Functions on Hindi QA System," International conference on ICT for Sustainable Development(ICT4SD), Springer, pp. 657-663, vol. 408, February 2016.