ABSTRACT

The overall performance of question answering system depends on Question classification. Question classification is the first part of question answering system, if the classification of question will be wrong then obviously the system will give the wrong answer. The two main important features of question classification are: 1) Different processing strategies can be chosen to find an answer for example if the question belongs to Description : Definition category, the QA system search for possible answers in encyclopedia, Wikipedia. 2) Narrow down the number of possible answers for example question belongs to the category Location : City, it means that the system only consider cities as possible answers.

The main objective of question answering (QA) is to provide a laconic answer, certain question posed in natural language. Most systems treat question answering as three distinct sub-tasks: processing of question, processing of the document, and answer processing.

Question processing consists of an assembly of question representation that is the classification of question, derivation of expected answer type, and keyword extraction. An important component of question answering systems is question classification. The focus of our research is on the question classification. The job of question classification is to forecast the entity type of the answer of a natural language question.

More specifically, classification of questions with respect to expected answer types. The research is motivated by the fact that previous research has shown to correctly predict the expected answer type. It is imperative for a question answering system as a whole to be successful. It is but obvious that if the categorization of question will be wrong then it will lead to finding out the wrong answer so it is very important for question answering system that it will classify the exact problem domain of a question. So definitely the proper classification of a question will also provide us a reduction of steps in two ways like first suppose if a question is classified as a category of CITY then the system will find out only cities no need to find out anything else and if a question belongs to category DEFINITION then the QA system will find out the
answer for explanation part using or connecting Wikipedia. The data for this research work have been taken from the Cognitive Computing Group at University of Illinois at Urbana-Champaign (UIUC) which is also known as TREC10 data set. Which consists of a training set of 5500 questions, and a test set with 500 questions (http://l2r.cs.uiuc.edu/~cogcomp/Data/QA/QC/).

Question classification is typically completed using supervised machine learning techniques because the training datasets of properly labeled questions has been used. The learning function is used for mapping the questions into question categories whereas unsupervised learning is used for clustering type of problems. Special types of features that are: lexical, syntactical and semantic features can be taken out from a question. In this research work one new syntactic feature has been introduced that is question prototype which improves the accuracy of classification. Headwords are occasionally not capable to extract the actual information of the statement from questions so some question descriptions have been taken and on the basis of that the exact category of question has decided. Furthermore, development of an information replica has been performed to optimally combine different features. This thesis introduces an optimization technique that is particle swarm optimization to the support vector machine classifier for improving the classification accuracy with appropriate selection of parameter value for Support Vector Machine.

The thesis discusses three different classifiers: Naive Bayes, Back-Propagation-Feed-forward Artificial Neural Networks (BPFFANN) and Support Vector Machine (SVM). It is observed that applying Naïve Bayes Classifier on question classification is easy to implement, but it cannot handle the correlation/dependence between input variables. Furthermore, the testing has been performed by the Back Propagation Feed Forward Artificial Neural Network classifier with the combination of a rich set of features and found that it takes extensive time for training and ANNs frequently converge on local minima rather than global minima. Classification accuracy is quite good as compared to naïve Bayes but ANNs repeatedly suffers the problem of over-fitting, further testing is performed by using Support Vector Machine classifier on the well-known TREC 10 dataset and the result of classification accuracy performs better than naïve bayes and Back Propagation Feed Forward Artificial Neural Network.
The primary objective of the proposed thesis work is to implement the hybrid algorithm which is based on the optimization of Support Vector Machine by using Particle Swarm Optimization. The question classification accuracy obtained by this research work is: 1) for Naïve Bayes classifier 80.6% for coarse grain and 78.6% for fine grain. 2) for BPFFANN 93.2% for coarse grain and 89.2% for fine grain. 3) for Linear SVM 96.25% for coarse grain and 91.1% for fine grain. 4) for PSO based Linear SVM 97.3% for coarse grain and 92.5 % for fine grain. The testing has been performed for the proposed research work on the well-known TREC 10 dataset and succeeded to achieve a new record on the accuracy of classification on this dataset.