A Fuzzy Logic Based Improved Keyword Extraction From Meeting Transcripts

J.I.Sheeba, Assistant Professor,
Department of Computer Science & Engineering,
Pondicherry Engineering College, Puducherry, India, sheeba@pec.edu.

Dr.K.Vivekanandan, Professor,
Department of Computer Science & Engineering,
Pondicherry Engineering College, Puducherry, India, k.vivekanandan@pec.edu

***Abstract*—Keyword Extraction is the process of assigning keywords to a document where the important words are selected by the system automatically. This proposed frame work is used to extract the keywords using Fuzzy logic method from Meeting Transcripts. At first, the given input is preprocessed. Subsequently, the preprocessed data will be sent to the features extraction method. In this method three features are introduced namely Frequency Calculation, Noun Extraction and Clustering using Fuzzy C-Means. Each feature is categorized into three sets like Low, Medium and High followed by Fuzzy rules which can be applied in all these features to extract the certain set of principal keywords from the given input. Now the extracted keywords are reduced by solving Synonym, Homonym, Hyponymy and Polysemy (SHHP) problems method. Here, this method is performed by existing methods, the result leads to a better extraction compared to the available approaches**

Keywords- ***Keyword extraction; Fuzzy logic; Fuzzy rules; Features Extraction; Text mining***

I. INTRODUCTION

There are several hindrances for the users to obtain and manage effective information on the web. One of the best solutions to solve those difficulties is the method of representing documents by using the small major set of words called Keywords. A Keyword is a simple word that characterizes the theme and content of a document. It can be identified by frequency, length and part of speech of a particular document. Keywords are particularly used in many applications specifically in the field of searching and managing information for the users, such as query refinement of search engine, summarization, social tagging, and feature selection for classification, clustering, text indexing, categorization and topic detection for improvement of search results. These keywords can be assigned manually or automatically. Manually assigned keywords are hard to implement for large documents, hence automatic keyword extraction is needed to select essential keywords automatically in the meeting transcripts [1].Keyword extraction attempts to extract keywords according to their statistical properties. Many appropriate keywords may not be statistically frequent or even not appear in the document especially for short documents. For instance, in a news article discussing about “iPad” and “iPhone”, the word “Apple” may not be mentioned, but both “iPad” and “iPhone” are the products of “Apple”, so the word “Apple” may thus be the essential keyword for the document[2].This type of keywords are called low frequency keywords. In existing methods are not extracting for low frequency keywords and they not reducing keywords from extracted keywords.

In this proposed method, keyword extraction will be enhanced using Fuzzy logic technique. Here many features are added for extracting the accurate keywords list. Certain features are more important while others are less important, hence there should be a balanced weight in computation. Fuzzy logic is used to rectify this problem by adopting membership function for each feature. In this process, Frequency calculation, Noun Extraction, Clustering of similar words are used as word features. Fuzzy logic technique in the form of approximate reasoning provides decision support and expert system with powerful reasoning capabilities. Human reasoning is not only a traditional two-valued or multi-valued logic but also logic with Fuzzy truths, Fuzzy rules of inference and Fuzzy connectives. Fuzzy logic also deals with the concept of partial truth where the truth value indicates the range between completely true and completely false (between 0 and 1).The Fuzzy logic system consists of four components: Fuzzifier, Defuzzifier, Fuzzy knowledge base and Inference engine. Fuzzifier can be defined as crisp inputs which are translated into linguistic values by using a membership function, Inference engine will do the rule base containing fuzzy IF THEN rules to derive the linguistic values and finally the output linguistic variables from the inference engine are converted to the final crisp values by the Defuzzifier using membership function to represent the final sentence score[3].

In this proposed framework, Fuzzifier has three features for words and they are specified as, Frequency Calculation, Noun Extraction and clustering using Fuzzy C-Means. In this function, each feature is categorized into three sets like Low, Medium and High. The extracted keywords are reduced by solving Synonym, Homonym, Hyponymy and Polysemy (SHHP) problems method. This method focused on solving Synonym and

Hyponym problems by reducing the similar keywords and provides better results. Solving Homonym and Polysemy problems gives more accurate meaning to the keyword.

II. RELATED WORKS

Zhiyuan Liu and Maosong sun proposed a system for incorporating the prior knowledge about graph-based methods for keyword extraction. By the combination of this prior knowledge teamed up with the neighborhood knowledge, the performance of graph-based keyword extraction method can be managed[1] .Weinan zhang and Dingquan wang introduced a novel algorithm for advertising keywords which was recommended for short-text Web pages using page rank on the Wikipedia graph. This method verifies the content and advertisement on the web pages[4].Soheila Karbasi proposed on the normalization of term frequency in weighting schemes and the term importance degree has been identified. It can be implemented to term-weighting schemes by using several parameters[5].Shady Shehata and FakhriKarray proposed concept-based retrieval model, which consists of conceptual ontological graph (COG) representation and concept-based weighting scheme. The COG representation in this method captures the semantic structure of each term within a sentence where all the terms are arranged in the COG representation according to their contribution to the meaning of the sentence. It is based on both sentence and document level[6].Long Thanh Ngo and Dinh Sinh Mai introduce a method to improve the computational efficiency of the interval type-2 Fuzzy C-Means Clustering(IT2-FCM) based on GPU platform and applied it to land-cover classification from multi-spectral satellite image. In this process the performance of GPU is more faster than CPU[7].Nayana Mariya Varghese proposed a cluster optimization methodology based on fuzzy logic which is used for eliminating the occurrence of redundancies in data after the clustering was completed by the web page mining methods. For Clustering Fuzzy C-Means algorithm is used in this method [8].G. Tadayon Tabrizi proposed a new algorithm that avoids being trapped in local peaks using fuzzy logic and PSO algorithm. It also finds a global optimal response or optimal place of cluster centers[9]. Tushar introduced Fuzzy logic based approach for determining the input-output relationship of some manufacturing processes. Here, fuzzy logic controller is used to determine regression equations and also developed two types of clustering algorithm, namely entropy-based approach and fuzzy C-means algorithm[10].S.Selva Kumar introduced a new technique mixed C-means clustering .This method is mainly used to test against a brain tumour gene expression[11].

Maciej Piasecki proposed a new tool WordNetLoom an application for WordNet development it mainly consists of two methods namely a form-based and graph-based and it has also discussed the role of the application in WordNet development[12]. Feifan liu and Fei liu proposed an automatic keyword extraction from meeting transcripts by applying bigram method compared to unsupervised TFIDF selection with POS filtering bigram expansion method well performed[13]. Feifan liu and Fei liu introduced keyword extraction method using TFIDF and POS and SRICM tool kit method which incorporates other methods like POS and word clustering and it also evaluates the importance of a word by applying graph based method[14]. Feifan liu and Fei liu proposed another method Single-loop feedback strategy for keyword extraction . In addition to traditional frequency or position-based clues, term specificity features, decision-making sentence-related features, including a group of features which are derived from summary sentences and to generate better system summaries, they proposed a feedback loop mechanism under a supervised framework to leverage the relationship between keywords and summary sentences[15]. Xuan-Hieu Phan proposed a method to find Hidden Topic by implementing proposed Framework towards Building Applications with Short Web Documents using LDA model. In this model the low frequency keywords are easily identified and they rectified two problems which are data sparseness and synonyms problem using LDA method through MaxEnt classifier [16]. Metzeler proposed a large variety of similarity measures for short queries from Web search logs[17]. Yih and Meek discussed for how to improve the Web-relevance similarity and its method[18]. Sahami and Heilman also calculated the relatedness between text snippets with the help of search engines and a similarity kernel function[19]. Gabrilovich and Markovitch discovered a semantic relatedness using Wikipedia concepts[20]. Sheeba and K.Vivekanandan proposed a method to extract the low frequency keywords and keyphrases for every sentence in the meeting transcripts. This recommended method consists of additional features to upgrade the keyword extraction function such as, N-Gram based method, Capital letters, Double quotes method, LDA method, Sequential pattern mining and also it includes TFIDF method[21]. Another method is suggested by Sheeba and K.Vivekanandan which particularizes the combination of both keyword extraction and sentiment classification into a single model to perform both the works at a time[22].Sheeba and K.Vivekanandan proposed one more method which mainly focus on extracting hidden topics from meeting transcripts. This proposed framework will be focused on solving Synonym, Homonym, Hyponymy and Polysemy problems in meeting transcripts[23].J.I.Sheeba and K. Vivekanandan proposed the another method for Improved Unsupervised Framework for solving Synonym, Homonym, Hyponymy and Polysemy Problems from Extracted Keywords and also it identify topics in Meeting Transcripts[24].In this proposed method additional features are included for improving the keyword extraction methods such as Noun Extraction, Frequency Calculation and

Clustering using Fuzzy C-Means algorithm based on Fuzzy logic technique and also the keywords are reduced by using SHHP method .

III. PROPOSED FRAMEWORK

The below framework is proposed to extract keywords from the given input using Fuzzy logic method. The Figure1 represents framework for the proposed system. In this framework Meeting Transcripts (Example :Supreme Court Dialogs Corpus) has been taken as input. The data preprocessing has been implemented for this text file and the output will be sent to input to the feature extraction step. Here, the features are named as Frequency Calculation, Noun Extraction and Clustering of similar words are derived from each word. Thus, each word is meant to be associated with vector of three features. All these three features are fuzzified to acquire Fuzzy set and they are categorized as Low, Medium and High. Fuzzy Rules have to be applied to these Fuzzy sets to extract all the keywords. Ultimately, the final keywords are extracted by using Defuzzifier method. The extracted keywords are reduced by using Synonym, Homonym, Hyponymy and Polysemy (SHHP) problems method. This method focused on solving Synonym Hyponymy, Homonym and Polysemy problems by reducing the similar keywords.

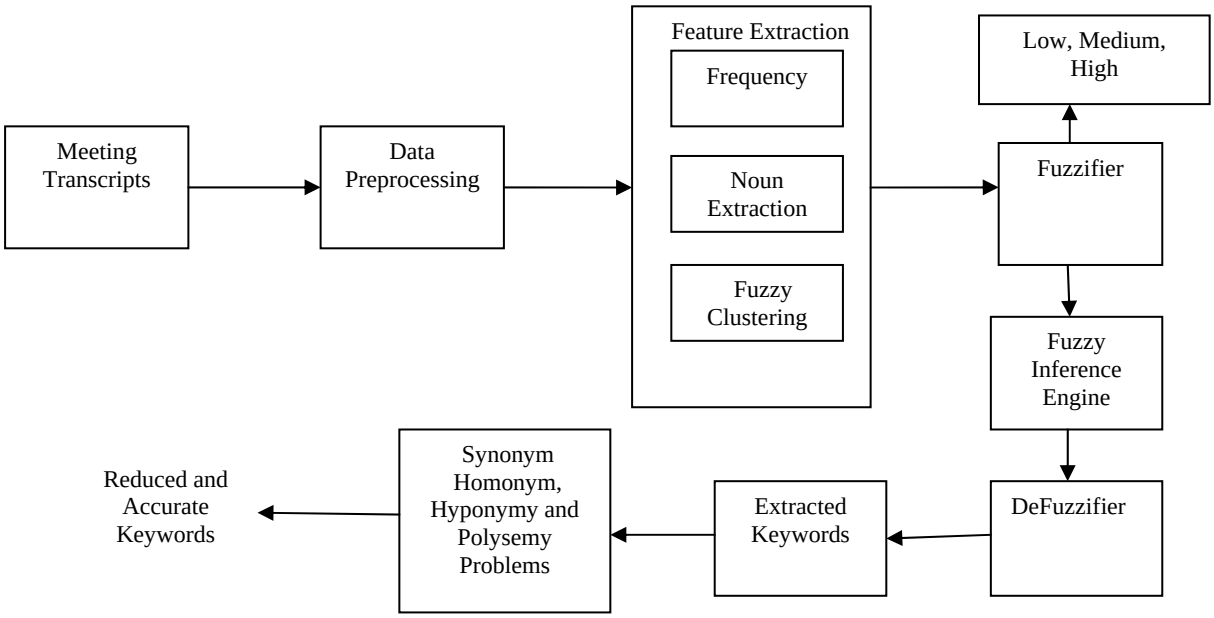

Figure1: Fuzzy logic based for solving Synonym, Hyponymy, Homonym and Polysemy (SHHP) problems from the extracted keywords using proposed Keyword Extraction Framework.(FL-SHHP-KEF)

The following steps are adopted in this proposed framework

A. Data Preprocessing.

In this technique, stop words and stem words present in the input data must be removed. Stop Words can be defined as removal of the words which appear frequently in the document providing lesser meaning while identifying the essential content of the document. Example: a, an, from, become, with etc.

Stem Words are specified as, the process of removing prefixes and suffixes from each word which refers to the end of the word that signifies different tenses.

Example: ed, es, ing and so on.

B. Feature Extraction.

The Feature extraction techniques are used to obtain the important words in the text. After preprocessing, each word from the document is represented by a vector of features. For each word , following three features are implemented such as Frequency Calculation, Noun Extraction and Clustering using Fuzzy C-Means algorithm.

1) Frequency Calculation: In the first feature, it is calculated by counting the number of each word occurring in the preprocessed data. The frequency is calculated by the occurrence of the word in which the highest frequency is calculated and divided in order to split the words into three following categories like High and Medium and Low. The valuation for this feature is derived from the ratio of the number of occurrences of the word with the total number of words in the preprocessed text.

2) Noun Extraction: Nouns are extracted using the Q-tag which is also classified into three category namely Low, Medium and High. The extracted Nouns are compared with the trained data set. The extracted Nouns which are present in the trained data set are assigned to the High category, in that case, the Nouns which are not present in the trained data set will be assigned to the Medium where as the words other than the Nouns will be assigned to the Low category.

3) Clustering using Fuzzy C- Means: The Fuzzy C-Means algorithm is mainly used for clustering. A cluster brings together instances in the data that share a common set of attributes .For each of these clusters a central value representative of the cluster's principal value is calculated .This is called centriod of the cluster. An array of these centroids for a collection of data elements from a database provides the cluster centers, it shows which set of rows in the database are closely related. This algorithm depends on the selection of the initial cluster center and the initial membership value. It involves two processes, the calculation of cluster centers and the assignment of points to these centers using a form of Euclidean distance.This process is repeated until the cluster centers have stabilized .The main goal of this algorithm is the assignment of data points into clusters with varying degrees of membership. This membership reflects the degree to which the point is more representative of one cluster than another [25].Initially, the similarity of a word and the compared words are calculated. The similarity of each word is derived by using Word Net similarity tool. With the measures of similarity, words are clustered into three groups by applying C-means algorithm. In this algorithm, each cluster is meant to be assigned with the centroid value. This represents that the first cluster includes centroid as the smallest value of similarity measures, where as the second cluster maintains the mid value as centroid and the last cluster consists of the highest value.

Subsequently, the similarity measure of each word will be summed up to any one of these three clusters in which the measures are closely related. The above task is accomplished by calculating the Euclidean distance formula. In this course of action, the distance will be derived between each similarity measure and the centroid of each cluster. The least value of the three distance values are recovered and subsequently measurement of clusters are added with the smallest distance value. The measure is added to the cluster repeatedly, where the centroid of that cluster is modified in order to discover a new average value. Similarity of all the words are identified with the use of Word Net, consequently the average of each word is calculated. The above method denotes that the words are clustered into three categories Low, Medium and High. Finally, the distance formula is used to find the centroid for each cluster. This procedure demonstrates that the words which contain closer similarity and they are assigned to the High cluster, the words which contain moderate similarity compared to other words are assigned to the Medium cluster and the words which do not have similarity with any other words are shifted to the Low category. This algorithm improves the accuracy of the cluster under noise .It will converge very quickly and also reduces the processing time .

C. Fuzzification.

Fuzzifier can be defined as crisp inputs which are translated into linguistic values by using a membership function. In this system, three word features are extracted where each word is associated with the vector of three features. These three features are operated as input to the fuzzifier separately and tends to manage the vector set of each feature as follows Low, Medium and High. In this function, each features are categorized into three sets like Low, Medium and High. Here the output returns two lists. The first list shows based on the frequency values the words are assigned as Low, Medium and High and the second list shows based on the Noun extraction and these words are ranked as Low ,Medium and High.

D. Fuzzy inference engine.

After Fuzzification, the Fuzzy inference engine is activated and it indicates to the rule base that contains Fuzzy rules and they are applied to derive the linguistic values. Fuzzy rules are implemented to this process in order to assign linguistic values to all the three features. The inference engine is linked with all the category values which are later converted into a single group that is eventually categorized as Low, Medium and High. The vital role of the fuzzy inference engine is to assign the fuzzy rules. The important sentences are extracted from these rules according to the criteria of the features. In a fuzzy rule ,the set of fuzzy propositions associated with the if condition of the rule is known as the premise or the antecedent .

In the rule if x is Low and y is High then z is Small the premise consists of the two fuzzy propositions x is Low and y is High connected by the And operator . Fuzzy rules are assigned based on user's wish. For example one of the Fuzzy rules is shown as below. Similarly the remaining Fuzzy rules are separately created for other words.

For Example : If(Frequency is Low), (Noun is Medium) and (Similarity is Low) then the result is Low

E. Defuzzification

The output of the linguistic variables from the inference engine will be converted into the final crisp values by the defuzzfier using membership function for representing the final sentence score. In this step, Defuzzification utilizes the output membership function which can be classified as: Output (Low, Medium,

High) that converts the Fuzzy results from the inference engine into a crisp output to derive the final evaluation of each sentence. The suitable words which are represented as High and they must be considered as the principal keywords.

F. Synonym, Hyponymy, Homonym and Polysemy problems (SHHP) method

The output of the extracted keywords will be sent to the input of this method .A Synonym means different words having same meaning are grouped . Similar words are grouped and their topic inference is made. The topics of the similar words are extracted. Each similar group of words contain single topic and that topic is extracted as the output of this problem. By extracting topics the synonym problem has solved. Hyponymy means one word denoting subclass is considered and super class keyword is extracted . A Hyponym is a lower class, specific term whose referent is included in the referent of higher class term. Here subclass of the word is considered . By extracting super class of each word is considered as the output. Here Synonym , Hyponym, Homonym and Polysemy problem was solved by using Word net as training dataset. This method focused on solving Synonym and Hyponym problems by reducing the similar keywords and provides better results.

Homonym means a word can have two or more different meanings. For example, Left might appear in two different contexts: Car left (past tense of leave) and Left side (Opposite of right). Here , keywords are grouped under hidden topics. These topics are labeled with generalized context. Homonym keywords are identified by comparing with hidden topic keywords. The corresponding topics name gives context of keywords and then calculated the frequency used for extraction. The outputs of this problem are keywords and different meaning words. A polysemy means word with different, but related senses. For example, Count has different related meanings: to say number in right order, to calculate. Different keywords are presented in the output. Related meaning keywords are identified by comparing with hidden keywords. Homonym and Polysemy problem gives more accurate meaning to the keyword. Finally the reduced keywords are extracted from this method [19].

IV. RESULTS AND DISCUSSION

A. Data set

In this paper, the domain meeting transcripts has been focused. Meeting speech is significantly different from written text and other speech data. For example, in meeting transcripts, many people can participate, even the discussions are not organized well, and the speech is unplanned one and it contains disfluencies and also the sentences are not constructed well .The people who are involved in the meetings speak different pronunciations and they use different types of words. People can play different roles and speak various topics in the transcripts. So extracting keywords from meeting transcripts is difficult one compared to the documents[14][15].

The proposed system validated using data set from Supreme Court Dialogs Corpus . The text file is similar to meeting transcripts format. This corpus contains a collection of conversations from the U.S. Supreme Court Oral Arguments (http://www.supremecourt.gov/oral_arguments/) with metadata: It contains 51,498 utterances making up 50,389 conversational exchanges from 204 cases involving 11 Justices and 311 other participants (lawyers or amici curiae) .The metadata includes case outcome, vote of the Justice ,section in which the conversation took place,gender annotation. Case outcome and vote data were extracted from the the Spaeth Supreme Court database (http://scdb.wustl.edu/). Here, randomly conversations have been chosen from the Corpus. By applying this proposed framework essential keywords are extracted from these Corpus. These inputs are tested by using both Existing systems , Classifiers and Proposed systems. The performance of the Proposed system is compared with the Existing systems and Classifiers. The results are shown that the proposed system is an better one. In the same way this proposed method will be implement for other datasets also.

B. Existing systems and Classifiers

1) Keyword Extraction through LDA model

LDA is a generative probabilistic model of a corpus which is widely used in document analysis. It models the semantic relationships between words based on their co-occurrences in documents[26]. Here the documents are represented as random mixtures over latent topics and each topic is characterized by a distribution over words. In this model, the words are grouped based on probability frequency . LDA takes many iterations to find more relevant words and hidden words as Output . The Output is generated randomly using Gibb sampling algorithm. This method will return no of keywords are high but compared with proposed approach the accuracy is less .

2) Keyword Extraction through MaxEnt Classifier

The MaxEnt classifier is a Maximum Entropy classifier and it is mainly used in extracting the keywords from the given input file based on frequency. The input file format is given as .csv. Also with the frequency, some of the features are included and these features are produced by the environmental layers.

It works based on the formula

$P(x) = \exp(c1 * f1(x) + c2 * f2(x) + c3 ** f3(x) ...) / Z$

Where C1, C2, ….. ->constant. f1, f2, ….. ->features. Z -> scaling constant .

In Linear features, for each species, the output distribution has the same expectation in quadratic feature same expectation and variance of the environmental variables are the samples. Product of two continuous environmental variables produces mean value are called product feature. Environment variable is resultant from a continuous environmental variable. It produces binary values 0 and 1. When variable value >1, it produces 1. Hinge feature is same as the linear feature which has been derived from a continuous environmental variable. Many files are produced by the MaxEnt for every species. For a species called *file1*, it produces files *file1.csv*. *file1.asc*(or*mySpecies.grd*). *file1.lambdas* containing the computed values of the constants *c1, c2, …. file1.png* is a picture of the prediction of each of the continuous environmental given by a directory containing the layers[21] .But in this method the output of the keywords are very less. It will not cover the important keywords in the input.

3) Keyword Extraction through SVM Classifier

A Support Vector Machine is used for classification purpose. It takes a set of input data and predicts for each given input to which category does the input fall. The training examples of one or two categories are taken. SVM training algorithm builds a model using those categories. It assigns new examples into one or the other categories. SVM classifier takes a set of input data and predicts for each given input, to which category does the input fall. The SVM classifier can be trained by taking some sample documents per each category. The sample documents can be prepared by browsing the web pages which contain topic and category. The new input documents can be classified based on predefined category. As the number that stands for each category the result from the SVM classifier is extracted [21]. This method will return no of keywords are high but compared with proposed approach the accuracy is less

4) Graph based Methods for Keyword Extraction

In this method keywords are extracted by applying graph based and the Ranking algorithm method. In a given input the accuracy of keyword extraction will be affected by the low similarities between unrelated words. To reduce the influence of low weights and to enhance keyword extracting performance, the data preprocessing function is implemented. The similarity relationship is calculated in every single word by using Disco tool. By applying similarity measures the graph is drawn. In this method, the words are represented as vertices and edges that can be sketched between two words where the similarity relation between these words are existed and the edges are measured by weight .Once the similarity measures among all the words are recovered, the similarity words are greater than zero and it will be represented in graph. The words are ranked by using Diffusion rank algorithm method. Highly ranked words are considered as important keywords[1]. When the similarity is less some important keywords are not coverd using this method.

5) Improved Keyword and Keyphrase Extraction Framework(IKKEF)

Our proposed framework aims to extract low frequency keywords and keyphrases from meeting transcripts . This framework extracted keywords by both MaxEnt and SVM classifiers and also Bigram and Trigram keywords are retrieved by using N-gram based approach .It identify the low frequency keywords using LDA method .Additionally it also includes the features of Double quoted keywords and Capital keywords. This method is performed by both human transcripts and ASR transcripts. Finally, the quality of the extracted keywords are improved by using sequential pattern mining method[21].

6) Synonym, Homonym, Hyponymy & Polysemy problems Using Keyword Extraction Framework (SHHP-KEF)

Our another proposed framework focuses on solving Synonym, Homonym, Hyponymy and Polysemy problems from the extracted keywords. Synonym means different words having similar meaning are grouped and single keyword is extracted. Hyponymy means one word denoting subclass is considered and super class keyword is extracted. Homonym means a word can have two or more different meanings. For example, Left might appear in two different contexts: Car left (past tense of leave) and Left side (Opposite of right). A Polysemy means word with different, but related senses. For example, count has different related meanings: to say number in right order, to calculate. Finally MaxEnt classifier is used for extracting keywords and topics in the transcripts[23].

7) Fuzzy Logic based Keyword Extraction Framework(FL-KEF)

Fuzzy logic can handle the problems with imprecise and incomplete data ,and it can model nonlinear functions of arbitrary complexity[27].Fuzzy logic technique is introduced to improve the accuracy of the keyword list and also it includes the benefits of simplicity and flexibility. This method has been extracted for all the important keywords using Fuzzy Feature Extractions like Frequency Calculation, Noun Extraction. Applying Fuzzy C-means clustering algorithm for extracted keywords ,the similar keywords clustered .Finally ,it provides the reduced the keywords list. Here Fuzzy logic simplifies the job of the user and the computer. The results are more accurate representations of the way systems behave in the real world.

C. Proposed System

1) Fuzzy logic based for solving Synonym, Hyponymy, Homonym and Polysemy (SHHP) problems from the extracted keywords using proposed Keyword Extraction Framework.(FL-SHHP-KEF)

To extract the accuracy of keywords list and also to increase the efficiency of the existing methods, some features are added to this method. By applying FL –KEF method the keywords are extracted .The extracted keywords are solved by Synonym, Hyponymy ,Homonym and Polysemy problems. The similar keywords are grouped by Synonym and Hyponymy method. To apply the Homonym and Polysemy method the keywords are reduced. By evaluating both the existing and proposed systems, hence it is proved that Fuzzy logic method enhances the process of keyword extraction which ultimately improves the accuracy and cover the low frequency keywords of the proposed system.

2) Proposed Algorithm for Improved Keyword Extraction

Input : Supreme Court Dialogs Corpus

Step 1.Read Input Data

Step 2.Collect Dataset D={D1,D2,…..Dn}

Step 3. Remove Stop words and Stem words from the inputs(Preprocessing).

//Frequency calculation

Step 4.Find Frequency calculation from the preprocessed data

Step 5.Divide the words into three category based on condition

Step 6. Repeat step 4 to 5 for all the inputs.

// Noun Extraction

Step 7.Extract Nouns from the inputs using Qtag tool

Step 8.Compare Extracted Nouns with trained dataset

Step 9.Divide the Nouns into three category based on condition

Step 10.Repeat steps 7 to 9 for all the inputs.

//Fuzzy C-Means clustering

Step 11.Select the initial value of k for n clusters

Step 12.Calculate the dissimilarity between an object and the mean of a cluster

Step 13.Allocate an object of the cluster that mean is nearest to the object

Step 14.Recalculate the mean of a cluster of the objects allocated to it, in such a way that the intra cluster dissimilarity is minimized

Step 15.Repeat steps 12 to 14 until the algorithm converges.

Step 16.Compute the centroid for each cluster and co-efficients.

Step 17.Calculate the Euclidean distance and find the centroid for each cluster

Step 18.Divide the cluster into three categories based on similarity values

Step 19.Translate linguistic values using membership function

Step 20.Apply fuzzy rules convert three category values into single group.

Step 21.Convert linguistic values into final crisp values.

Step 22.Repeat steps 19 to 21 for all values of three categories.

//SHHP Method

Step 23 .Find the meaning of each extracted keywords

Step 24.Similar words are compared with Predefined dataset .

Step 25. After comparison similar words are replaced into single word.

Step 26. Similar group of words are grouped as Synonym word and reduces redundancy.

Step 27. Next step is reducing keywords by finding hyponym words.

Step 28. WordNet is used as dataset for comparing hyponym words

Step 29. Words in the input file is compared with words in WordNet using gethypernym method.

Step 30. Superclass of similar subclass words are retrived using WordNet.
Those retrieved words are Hyponym keywords.

Step 31. Homonym and polysemy words are identified which contains relevant words under some topic

Step 32. Identify the different meanings of the word which are relevant to the conversation

Step 33. Extract those words and substitute to homonym words.

Step 34. Similarly, Different sense of the words is identified and extract from dataset.

Step 35. Replaced with polysemy words.

*Output :*Reduced and accurate keywords.

In this algorithm the process is done step by step for the best result in extracting keywords . Here, supreme court dialogs corpus have been taken as input. Steps 1 to 3 data preprocessing has been implemented for this given input. In Step 4 to 5 the frequency value of all the words are found. The words are divided into three categories based on condition. Step 7 to 9 Nouns are extracted from the input using Qtag tool. Extracted Nouns has compared with trained dataset. Nouns are divided into three categories based on condition. In Step 11 to 21 the initial value of v for n clusters is selected. Dissimilarity between object and the mean of a cluster is calculated. An object of the cluster is allocated and cluster of the objects is recalculated. The centroid for each cluster and co-efficients is computed and the Euclidean distance is calculated. Clusters are divided into three categories based on similarity values. Linguistic values are translated using membership function. Fuzzy rules are applied to convert three category values into single group and also linguistic values converted into final crisp values.

Step 23 to 26 the meaning of each extracted keywords and similar keywords are compared with dataset . After compared, similar keywords are replaced into single word .In these way Synonym problem was solved. In Step 28 WordNet is used as dataset for comparing hyponym words. Here Words in the input file is compared with words in WordNet using gethypernym method and Superclass of similar subclass words are retrived using WordNet. Those retrieved words are Hyponym keywords. In Step 31 to 35 Homonym and polysemy keywords are identified which contains relevant words under some topic and Identify the different meanings of the word which are relevant to the conversation. Extract those words and substitute to homonym words. Similarly, Different sense of the words are identified, extracted from dataset and replaced with polysemy words. Finally this method will extracted the accurate keywords.

3) Metrics Considered for Evaluation

The performance of the proposed framework is measured in terms of the quality measures namely Precision, Recall and F-Measure.

Precision

Precision is the fraction of retrieved keywords that are relevant

$$\text{Precision} = \frac{\{\text{Number of Relevant Keywords}\} \cap \{\text{Number of Retrieved Keywords}\}}{\{\text{Number of Retrieved Keywords}\}}$$

Recall

Recall is the fraction of the keywords that are relevant to the query that are successfully retrieved.

$$\text{Recall} = \frac{\{\text{Number of Relevant Keywords}\} \cap \{\text{Number of Retrieved Keywords}\}}{\{\text{Number of Relevant Keywords}\}}$$

F-Measure

F-Measure computes both precision and recall as the test to compute the score. Here precision is the number of correct keywords divided by number of all returned keywords. Recall is the number of correct keywords divided by the number of keywords.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

The results are shown in the Table 1. This table shows the performance comparison of the various techniques with proposed Fuzzy logic .The experiments have been repeated for randomly shuffled conversation and the results are obtained using Supreme Court Dialogs Corpus is shown in the Table 1. Table 1 tabulates the Precision and Recall achieved for various Techniques. In Table 2 shows the F-Measure value.

TABLE 1: PRECISION AND RECALL ACHIEVED

EXISTING METHODS			PROPOSED METHODS		
Techniques Used	Precision	Recall	Techniques Used	Precision	Recall
LDA	0.7522	0.6725	IKKEF	0.9454	0.7878
MAX Ent Classifier	0.833	0.7575	SHHP-KEF	0.9563	0.7899
SVM Classifier	0.933	0.707	FL-KEF	0.9743	0.7954
GRAPH Based Method	0.9387	0.7254	FL-SHHP-KEF	0.9878	0.8167

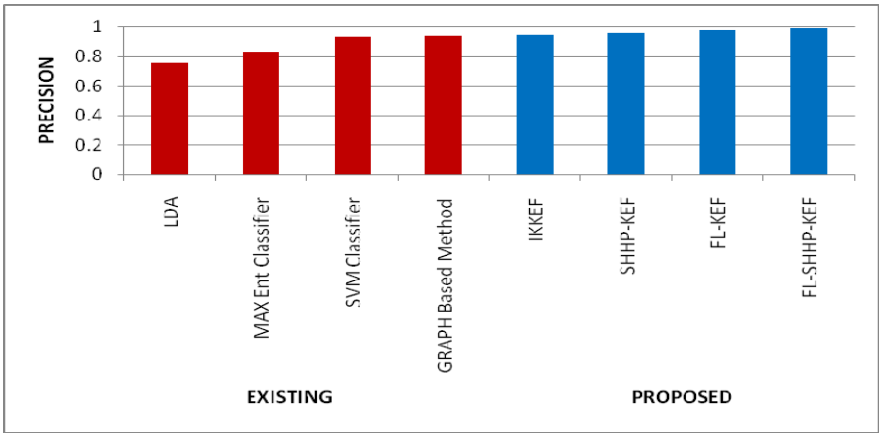

Figure 2. Precision achieved

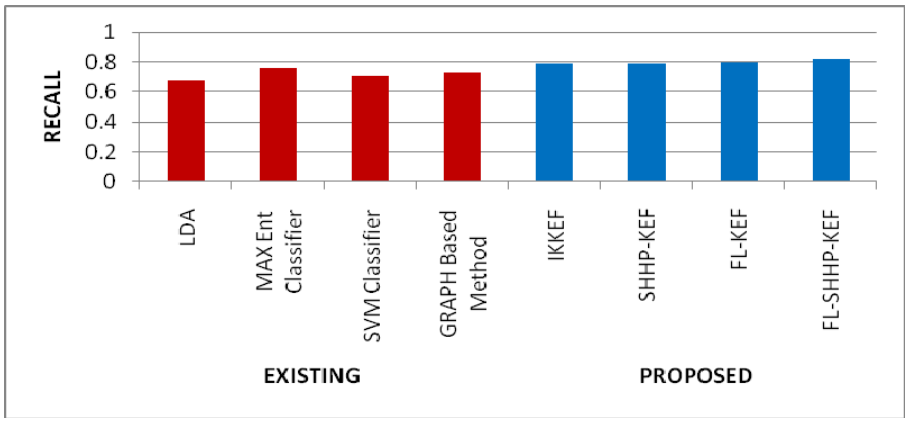

Figure 3. Recall achieved

It is observed from Figure 2 that the FL-SHHP-KEF achieves the best Precision and from Figure 3 the best Recall .It is also seen that the use of Fuzzy logic improves the performance of all the techniques with SHHP performing better than SHHP-KEF. For Precision FL-SHHP-KEF performs better by 3.15% then SHHP-KEF .Similarly FL –SHHP-KEF performs better by 1.35% then FL-KEF. For Recall FL-SHHP-KEF performs better by 2.68% then SHHP-KEF. Similarly FL-SHHP-KEF performs better by 2.13% then FL-KEF.

TABLE 2. F-MEASURE

EXISTING METHODS		PROPOSED METHODS	
Techniques Used	F-Measure	Techniques Used	F-Measure
LDA	0.7100	IKKEF	0.8593
MaxEnt Classifier	0.7931	SHHP-KEF	0.8650
SVM Classifier	0.8036	FL-KEF	0.8757
GRAPH Based Method	0.8183	FL-SHHP- KEF	0.8941

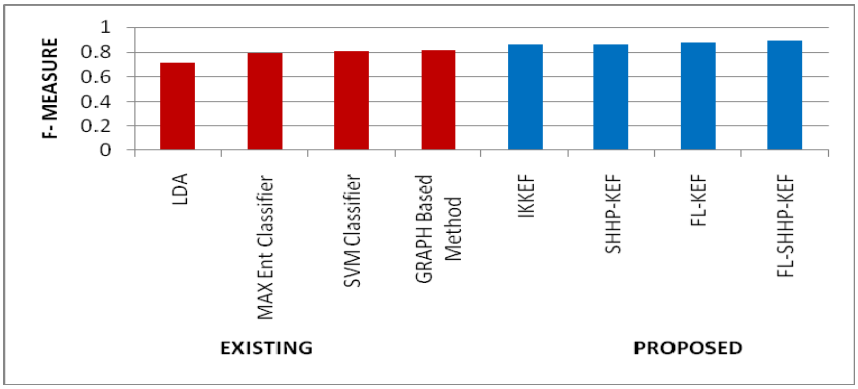

Figure 4. F-Measure

TABLE 3.CLASSIFICATION ACCURACY OBTAINED FOR VARIOUS TECHNIQUES

EXISTING METHODS		PROPOSED METHODS	
Techniques Used	Classification Accuracy	Techniques Used	Classification Accuracy
LDA	67	IKKEF	79
MAX Ent Classifier	76	SHHP-KEF	80
SVM Classifier	71	FL-KEF	84
GRAPH Based Method	75	FL-SHHP- KEF	87

It is observed from Figure 4 that the FL-SHHP-KEF achieves the best F-Measure. It is also seen that the use of Fuzzy Logic improves performance of all the techniques with SHHP performing better than SHHP-KEF .FL-SHHP-KEF performs better by 2.91% then SHHP-KEF .Similarly FL –SHHP-KEF performs better by 1.84% then FL-KEF.

The result obtained from Classification Accuracy is shows in Table 3. Root Mean Squared Error(RMSE) is shows in Table 4.Figure 5 shows the Classification Accuracy obtained from the various Techniques and Figure 6 shows RMSE.

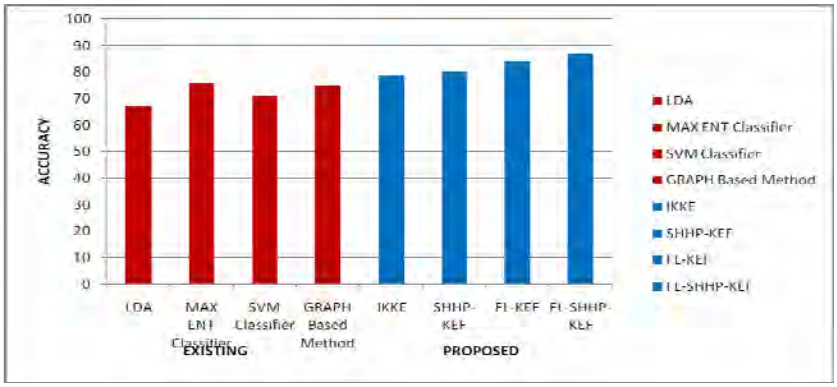

Figure 5:Classification Accuracy for various techniques

TABLE 4.RMSE OBTAINED FOR VARIOUS TECHNIQUES

EXISTING METHODS		PROPOSED METHODS	
Techniques Used	RMSE	Techniques Used	RMSE
LDA	0.4325	IKKEF	0.2868
MAX Ent Classifier	0.3927	SHHP-KEF	0.2437
SVM Classifier	0.4157	FL-KEF	0.2208
GRAPH Based Method	0.3201	FL-SHHP- KEF	0.2013

It is observed from Figure 5 that the FL–SHHP-KEF achieves the best Classification Accuracy .It is also seen that the Fuzzy Logic based SHHP method improves the performance compare with SHHP based KEF.FL-SHHP-KEF performs better by 7% then SHHP-KEF .Similarly FL –SHHP-KEF performs better by 3% then FL-KEF. Figure 6 shows the RMSE of different Techniques.

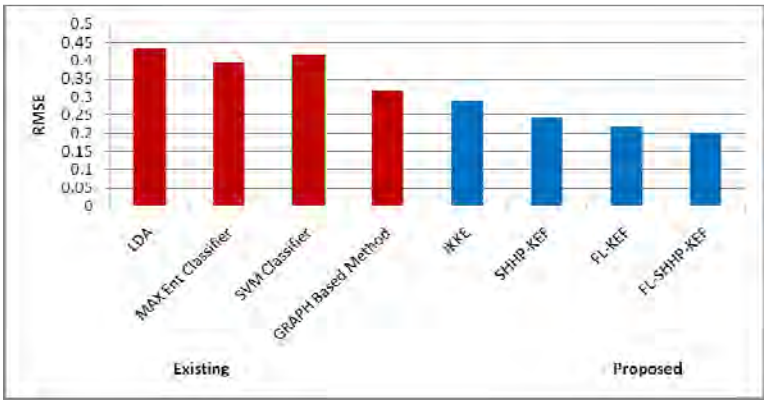

Figure 6.Root Mean Squared Error for various Techniques

It is observed from Figure 6 that the FL-SHHP-KEF achieves the least RMSE. FL-SHHP- KEF decreases by 4.24% then SHHP-KEF. Similarly FL-SHHP-KEF decreases by 1.95% then FL-KEF.

In this work a new extraction algorithm called FL-SHHP-KEF has been proposed for enhancing the accuracy of keyword extraction for the conversation corpus. The results obtained from this work obviously shows that the proposed algorithms achieving better accuracy when trained with fuzzy logic and SHHP method .

V. CONCLUSION

This research paper attempt to represent improved keywords extraction by applying Fuzzy logic method. The three features namely as Frequency Calculation, Noun Extraction and Clustering and they are extracted each feature will be categorized into three sets. Eventually, Fuzzy rules are applied on these sets to extract the keywords. The extracted keywords are reduced by SHHP method. To adopt more accuracy this framework has been implemented. It also enhances the process of extracting keywords from the Supreme court Dialogue corpus which are compared to the existing systems like LDA, Graph based method ,IKKEF,SHHP-KEF,FL-KEF and the keywords are extracted through MaxEnt and SVM classifiers . Here Fuzzy logic balances the importance of all the features and also returns the accuracy of the all important keywords from the inputs compared to existing methods.

REFERENCES

[1] Zhiyuan LIU, Maosong SUN, "Can prior knowledge help graph-based methods for keyword extraction?",ResearchArticle Higher Education press and springer - verlag Be ofrlin Heidelberg 2012, pp .242-253.

[2] Zhiyuan Liu, Xinxiong Chen, "Keyphrase Extraction by Bridging Vocabulary Gap", Proceedings of the Fifteenth Conference on Computational Natural Language Learning, Portland, Oregon, USA, 23–24 June 2011,Association for Computational Linguistics. pp. 135–144.

[3] Rucha S. Dixit, Prof. Dr.S.S.Apte, "Improvement of Text Summarization using Fuzzy Logic Based Method" , IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727 ,Volume 5, Issue 6 (Sep-Oct. 2012), pp .05-10.

[4] Weinan zhang and Dingquan wang , "Advertising Keywords Recommendation for Short-Text Web Pages Using Wikipedia"**,** ACM Transactions on Intelligent Systems and Technology, Vol. 3, No. 2, Article 36, Publication date: February 2012.pp. 36:25.

[5] SoheilaKarbasi and Mehdi Yaghoubi, "TheEffect of Term Importance Degree on Text Retrieva"**l,** International Journal of Computer Applications (0975 – 8887), Volume 38– No.1, January 2012, pp .27 -31.

[6] Shady Shehata and FakhriKarray, "An efficient concept-based retrieval model for enhancing text retrieval quality" ,Springer-Verlag London Limited 2012 .

[7] Long Thanh Ngo and Dinh Sinh Mai, "GPU-based Acceleration of Interval Type-2 Fuzzy C-Means Clustering for Satellite Imagery Land-Cover Classification", 12th International Conference on Intelligent Systems Design and Applications (ISDA),2012 IEEE, pp. 992-997.

[8] Nayana Mariya Varghese and Jomina John, "Cluster Optimization for Enhanced Web Usage Mining using Fuzzy Logic", World Congress on Information and Communication Technologies,2012 ,IEEE, pp.948-952.

[9] M. Mir and G. Tadayon Tabrizi , "Improving Data Clustering Using Fuzzy Logic and PSO Algorithm", 20th Iranian Conference on Electrical Engineering, (ICEE2012), May 15-17,2012, Tehran, Iran. IEEE, pp .784-788.

[10] Tushar, Dilip Kumar Pratihar , "Design of cluster-wise optimal fuzzy logic controllers to model input-output relationships of some manufacturing processes", Int. J. of Data Mining, Modelling and Management, 2009 Vol.1, No.2, pp.178 – 205,DOI:10.1504/IJDMMM.2009.026075.

[11] S. Selva Kumar; H. Hannah Inbarani , "Analysis of mixed C-means clustering approach for brain tumour gene expression data", Int. J. of Data Analysis Techniques and Strategies, 2013 Vol.5, No.2, pp.214 – 228, DOI: 10.1504/IJDATS.2013.053682.

[12] Maciej Piasecki; Michał Marcińczuk; Radosław Ramocki; Marek Maziarz, "WordNetLoom: a WordNet development system integrating form-based and graph-based perspectives", Int. J. of Data Mining, Modelling and Management, 2013 Vol.5, No.3, pp.210 – 232, DOI: 10.1504/IJDMMM.2013.055861.

[13] Fei Liu, Feifan Liu, Yang Liu, "Automatic Keyword Extraction for the Meeting corpus using Supervised Approach and Bigram Expansion", ACM ,2008.

[14] Feifan Liu, Deana Pennell, Fei Liu, "Unsupervised Approaches for Automatic Keyword Extraction Using Meeting Transcripts", ACM ,2009.

[15] Fei Liu, Feifan Liu, "A Supervised Framework for Keyword Extraction From Meeting Transcripts", IEEE Transactions On Audio, Speech, And Language Processing, VOL. 19, NO. 3, March 2011,pp.538-548.

[16] Xuan-Hieu Phan, Cam-Tu Nguyen, "A Hidden Topic-Based Framework toward Building Applications with Short Web Documents" , IEEE Transactions On Knowledge And Data Engineering, Vol. 23,2011.

[17] D. Metzler, S. Dumais, and C. Meek , "Similarity Measures for Short Segments of Text", Proc. 29th European Conference IR Research (ECIR), ACM 2007.

[18] W.Yih and C. Meek , "Improving Similarity Measures for Short Segments of Text", Proc. 22nd National Conference on Artificial Intelligence (AAAI) 2007 .

[19] M. Sahami and T. Heilman, "A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets",Proc.15 th International conference on World Wide Web ,ACM 2006.

[20] E. Gabrilovich and S. Markovitch, "Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis", Proc. 20th Int'l Joint Conference. Artificial Intelligence 2007.

[21] J. I. Sheeba , K. Vivekanandan , "Improved Keyword and Keyphrase Extraction from Meeting Transcripts" **,** International Journal of Computer Applications (0975 – 8887), Volume 52– No.13, August 2012,pp. 11-15.

[22] J. I. Sheeba , K. Vivekanandan, "Low Frequency Keyword and Keyphrase Extraction from Meeting Transcripts with Sentiment Classification using Unsupervised Framework"**,** CCSEIT-2012, October 26~28, Coimbatore, Tamilnadu, India.ACM ,pp.212-216, ISBN 978-1-4503-1310-0.

[23] J. I. Sheeba , K. Vivekanandan , "Unsupervised Hidden Topic Framework for Extracting Keywords (Synonym, Homonym, Hyponymy & Polysemy) and Topics in Meeting Transcripts", Jul 13-15,pp.299-307,Chennai,springerlink.com ©Springer-Veraag Berlin Heidelberg 2012. ISBN no 978-3-642-31552-7 (Online)

[24] J. I. Sheeba , K. Vivekanandan , **"**Improved Unsupervised Framework for solving Synonym, Homonym, Hyponymy & Polysemy Problems from Extracted Keywords and Identify topics in Meeting Transcripts", International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.5, October 2012, DOI : 10.5121/ijcsea.2012.2508,pp.85-92.

[25] Earl cox, "Fuzzy modeling and Genetic algorithms for data mining and exploration", Published by Elsevier, Morgan Kaufmann publishers. ISBN No :0-12-194275-9,2005.

[26] Blei ,D.M.,NG, A.Y.,and Jordan ,M.I, "Latent Drichlet Allocation" .J.Math.Learn.Res.3, 2003. pp.993-1022.

[27] Innocent P.R and John ,R.I. , "Computer Aided Fuzzy Medical Diagnosis , Information sciences" ,Vol .162,No.2,pp.81-104,2004.

AUTHORS PROFILE

J.I.Sheeba received her B.E in Computer Science and Engineering from Bharathidasan University and M.E in Computer Science and Engineering from Anna University. She is currently pursuing her Ph.D in Computer Science and Engineering, from Pondicherry Engineering College affiliated to Pondicherry University. Presently she is working as a Assistant Professor in Department of Computer Science and Engineering, Pondicherry Engineering College. Her research interest includes Data mining and Fuzzy Logic.

Dr.K.Vivekanandan received his B.E from Bharathiyar University, M.Tech from Indian Institute of Technology, Bombay and Ph.D from Pondicherry University. He has been the faculty of Department of Computer Science and Engineering, Pondicherry Engineering College from 1992. Presently he is working as Professor in the Department of Computer Science and Engineering. His research interest includes Software Engineering, Object Oriented Systems, Information Security and Web Services. He has coordinated two AICTE sponsored RPS projects on "Developing Product Line Architecture and Components for e-Governance Applications of Indian Context" and "Development of a framework for designing WDM Optical Network".