ISSN:0975-9646

Enriching Text Summarization using Fuzzy Logic

S. Santhana Megala
Research Scholar,
PRIST University
Thanjavur, Tamil Nadu.

Dr. A. Kavitha
Dept. of Computer Science
Kongunadu Arts & Science College
Coimbatore, Tamil Nadu.

Dr. A. Marimuthu
Dept. of Computer Science,
Govt. Arts College
Coimbatore , Tamil Nadu.

***Abstract*— Automatic Text Summarization is a process of generating Summary/Head note for the text document. Text Summarization is carried out by two main methods, namely, Extraction and Abstraction. This paper utilizes the extraction process for sentence selection. Here some Feature based sentence scoring techniques also used, which played an important role in text summarization. Finally an analysis is done by comparing the Fuzzy Logic and Neural Networks techniques based upon the Precision, Recall & F-Measure. Fuzzy Logic rules were used to balance the weights between important and unimportant features based on the feature Extraction. The Experimental result shows that fuzzy Logics give an improving result than the Neural Networks.**

***Keywords*— Fuzzy Logic, Neural Network, Sentence Scoring, Feature Extraction, Text Summarization**.

I. INTRODUCTION

In the digitized world, increasing amount of digital data results in a major problem called Information Overload, Which needs a technological tool, to create a head note summary, which interprets the large amount of text data in the digital documents. Text Summarization plays a vital role in creating the summary of a document by extracting the important information. An Automatic Text Summarization is a process of generating the summary of a source text by a machine without any human intervention, which represents the most important information of the source text in a shorter view / abstract view. Text Summarization helps the people to quickly understand huge amount of information within a finite time, which attracts the technological people to do more research in this field to upgrade this novel techniques to the next level.

Basically the text summarization process is divided into two type's abstraction and extraction, based upon how they are composed. Abstraction creates the summary which acts as a substitute for the original document i.e it employs the words and phrases which is not present in the original document, it replaces the word with same meaning. But Extraction process creates the summary by means of extract the sentences from the original document i.e it selects the important sentences in the documents to frame the summary. But most of the research work is done by using the extraction techniques only because abstraction techniques need lexical analysis and a data dictionary which is used to find the related meaning words in the document. On the other hand, there are more number of research was done using extraction techniques, which is a simple and easy one.

Automatic summarization using extraction techniques uses the extraction of features like Indicators / Cue Phrases, Named Entity Recognition, Local Feature & Layout Feature, Legal Vocabulary, State Transition Feature, Paragraph Structure, Similarity to Neighbouring Sentences, Citation, Absolute Location, Term Weight, Proper Noun, and Sentence Position. In this approach 10 feature Extraction methods were used to utilize the feature fusion technique to find out more accuracy in the summary generation. This paper set in motion with a brief analysis of Text Summarization Approaches in section (II), and Feature Extraction Techniques were discussed in the section (III). In Section (IV) the most popular Extraction Based Text Summarization Methods were discussed. The Experiments and Results were covered in section (V). Finally the conclusions and plan for future work is discussed in section (VI).

II. FEATURE EXTRACTION TECHNIQUES

A. Pre- Processing

The pre- processing step involves cleaning the noisy text containing grammatical and typographical errors. The major problems in text summarization are that the size of the document is not well known. Thus each word in the documents were represented by the terms in the vector space model, which cause the number of dimensions as too high for the text summarization algorithms. This pre-processing method plays a vital role in reducing the number of dimensions passed to the text summarization process. In this paper, the followed pre-processing methods were applied, namely, Case Folding, Removal of Stop Words, Punctuation Removal, Removal of Extra White Spaces, Word Stemming, Key Phrase Identification, Sentence Segmentation and Tokenization.

Case Folding is the process of converting all the capital letters to small letters, which is used to say that "ACT", "Act", "act" these all words are same. Removal of Stop Words is the process of removing the unimportant words which appears frequently in the document and provides less meaning in the text processing. Punctuation Removal is the process of removing the unwanted punctuation except dot (.) operator, which act as a sentence separator. Removal of Extra White spaces is the process of removing the additional white spaces, which reduces the size of the document in some rare cases. Word Stemming is the process of producing the root word by removing the suffixes and prefixes of each word in the document. Key Phrase Identification is the process of identifying the important phrases by finding the occurrence of the word pairs by using relative frequency approach. Sentence segmentation is the process of detecting the separating the paragraph into sentence. Tokenization is a process of separating the text document into individual words.

B. Sentence Features

Next to the Pre- processing step, by using sentence feature each sentence in the document is represented by a vector attribute. Each attribute represents the data used for their task. In this paper 10 features were used, which gives a value between "0" to "1". The 10 features were as follows:

- **Indicators/Cue Phrases**
  Cue Phrase means frequently used key phrases, which acts as the indicators of frequent used rhetorical roles in the sentence.
- **Legal vocabulary**
  Legal Vocabulary means words or phrases that include the Legal words of basic vocabularies from a training data.
- **Paragraph Structure**
  Paragraph Structure show the internal structure of the Paragraph, which have the high level sumup in the starting or summary towards the end.
- **Citation**
  Citation means referring some one, which is a needed one in the legal field, where in the arguments involve citing some other cases.
- **Term Weight**
  Term Weight is used to calculate the importance of sentence by finding the frequency of occurrences of the term within a document. The frequency of term occurrences within a document has often been used for calculating the importance of sentence. By summing the score of the words in the sentence, the score of the sentence is calculated.
- **Named Entity Recognition**
  Named Entity Recognition finds the presence or absence of the entities in the sentence by generating a binary value 0 or 1, Default Legal entities were stored in the legal dictionary.
- **Similarity to Neighbouring Sentences**
  Similarity to Neighbouring Sentence is used to find the similarity between sentences in the document. The sentence similarity between each sentence is calculated by using the cosine similarity measure, which compares all the sentences in the document.
- **Absolute Location**
  Absolute Location finds the location of a sentence, which has most important feature for sentence selection. In Legal field this feature indicates the original location of the sentence by ignoring the duplication.
- **Proper Noun**
  Proper Noun finds the sentence which contains more named entity called as proper noun, which is the important sentence that should be included in the document summary.
- **Sentence Position**
  Sentence Position is used to identify the importance of the sentence by identifying the position in the document. if the position is in the first few lines or paragraph or section then the rank is high.

III. EXTRACTION BASED TEXT SUMMARIZATION METHODS

A. Text Summarization Based on Fuzzy Logics

A Design of a Fuzzy logic system usually involves selecting membership function and fuzzy rules. The performance of the fuzzy logic system will directly affect by the selection of fuzzy rules and membership functions. The four main components of the Fuzzy Logic were: fuzzifier, inference engine, defuzzifier, and the fuzzy knowledge base. In the fuzzifier section, snappy inputs are translated into linguistic values, using a membership function, which is to be used to the input linguistic variables. After fuzzification, to derive the linguistic values, the inference engine refers to the rule base containing fuzzy IF THEN rules. Finally, the defuzzifier converts the output linguistic variables from the inference to the final crisp values using membership function which represents the final sentence score. The output membership function in the defuzzfication step is divided into three membership functions: Unimportant, Average, Important, Which is used to convert the inference engine result into a crisp output to obtain a final score for each sentence.

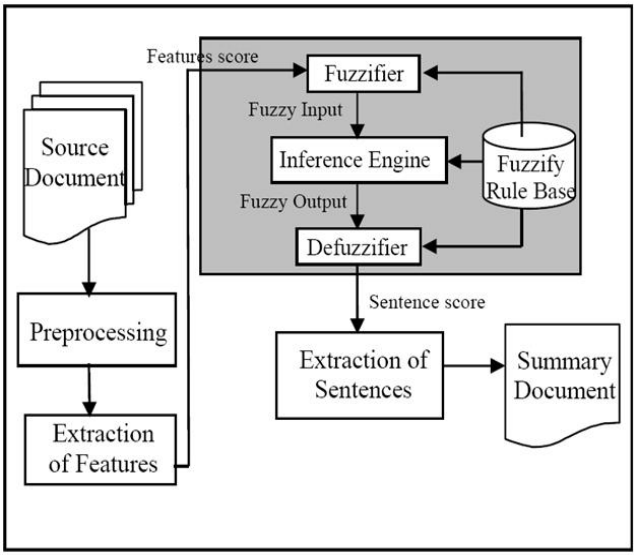

Fig. 1 shows text summarization based on fuzzy logic system architecture.

Here fuzzy centroid method is used to calculate the score for each sentence in a document, which is obtained by using generalized triangular membership function which depends on the three parameters a, b, and c. where the parameter a and c are left and right most feet of a triangle and b is the peak of a triangle. Based upon the sentence features and knowledge base the output is obtained as a value from zero to one for each sentence. Such obtained value shows the degree of importance of the sentences in the final summary. The formula to calculate the fuzzy centroid (1) is given below.

$$f(x,a,b,c) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \quad \text{----(1)}$$

The standard values of Low Medium and High were denoted by a, b, c respectively.

The most important part in the Inference Engine is the definition of Fuzzy If Then rules. The most important sentences were taken out from the rules mentioned below, which is based on our features selection.

IF (Indicatorcuephrase is VH) and (Legalvocabulary is H) and (Paragstructure is VH) and (Citation is H) and (Termwght is VH) and (Nameentityrecog is H) and (Absolutelocation is H) and (SentenceSimilarity is VH) and (NoProperNoun is M) and (SentencePosition is H) THEN (Sentence is important)

B. Text Summarization Based on Neural Networks

On the other hand, A Machine Learning Techniques i.e. an artificial neural network is implemented to generate text summarization system. The neural network is trained on a corpus of Legal Judgments available in the website, and then the trained one is modified through feature fusion to produce a summary by selecting the highly ranked sentences in the document. Totally the system is divided into three phases they are Training, Feature fusion and sentence selection.

1) Neural Network Training

In the first stage of the process, a training is given to the neural networks by which it learns the kind of sentences to include in the summary. Such thing is accomplished by training the network by using several test paragraphs which identifies each sentence whether to included in the summary or not. These test paragraphs were framed by the human readers. By using this training the neural networks, inherits the pattern of the sentences that should be included in the summary. Among the other Neural Network methods, the universal function approximation technique called three layered feed forward network is used. This neural network consists of ten input layer neurons, seven hidden layer neurons, and one output layer neuron. The purpose of training the neural networks is to find for the global minima of the energy function. Thus penalty function drives the unnecessary connection to a very small value to strengthen the remaining connections. Finally the unnecessary neurons and connections were pruned without influencing the performance of the network.

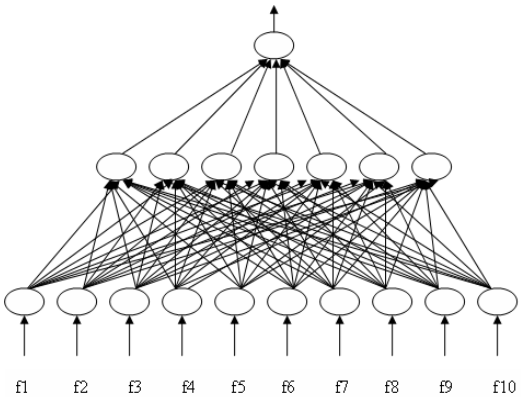

Fig. 2 shows the Neural Network after Training.

2) Feature Fusion

After the training, the Neural Network learnt which sentence features should exist in summary. With the help of feature fusion phase, the current trends and relationships among the features, which inherent the majority of the sentences were discovered. The above process consists of two steps i) Elimination the uncommon features ii) Collapsing the cause of the common features.

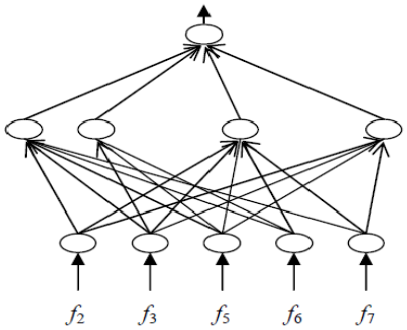

Fig.3 shows the Neural Network after Pruning.

The trained networks having miniature weights will be pruned without affecting the network and its performance. Thus any hidden layer neuron having no connection will be removed from the network, which leads to the elimination the uncommon features from the neural network. The hidden layer activation values in each hidden layer neuron were clustered using an adaptive clustering technique. Based upon the centroid and frequency, each cluster was identified and the activation value is replaced by the centroid of the cluster in each hidden layer neuron, which fusion the effects of common features. These two steps collectively generalize the effects of features and it provides the control parameters for sentence ranking.

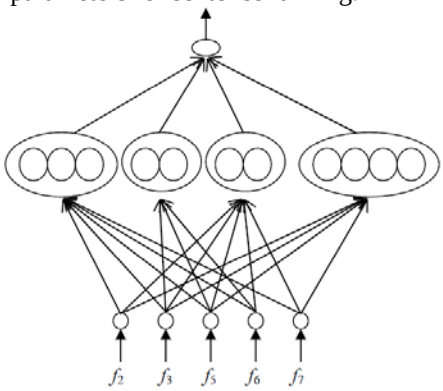

Fig. 4 shows the Neural Network after Feature Fusion.

3) Sentence Selection:

The Network can be act as a tool to filter the sentences in any paragraph of the document and it may determine whether the sentence may be include in the sentence or not, which can be done after the training, pruning and generalization steps. In this phase it provides control parameters for radius and frequency layer activation to select the highly ranked sentences in the document. The sentences which satisfy the cluster boundary and frequency were selected for high ranked summary. The hidden layer having less frequency connections were removed, which eliminates the uncommon features from the network.

IV. EXPERIMENTS AND RESULTS

In this paper, the system is implemented and tested in the MATLAB, which is a standard toolkit for the research analysis, and also it is suitable for the text data. It compares the summary generated by the program using Fuzzy Logic Method and Neural Network Method.

A) Experiments:

The experimentation is done by taking 50 Legal Judgement Documents from the legal website. And then by applying the pre-processing and feature extractions techniques on the data set taken to obtain the feature vectors.

TABLE 1
SAMPLE FEATURE SCORE FOR THE LEGAL TEXT DOCUMENT

S. ID	Feature Score									
	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
S1	1	1	0.6	1	0.9	0.5	1	1	0.2	0.5
S2	1	0	0.9	1	0.6	0.3	1	1	0.3	0.3
S3	1	0	0.8	1	0.3	0.6	1	0	0.6	0.2
S4	1	0.4	0.7	1	0.4	0.1	1	1	0.5	0.7
S5	0.6	0.6	1	1	0	0.5	1	0	0.9	0.3
S6	0.7	0	1	1	0	0.9	1	0.6	0.1	0.6
S7	0.9	1	1	1	0.5	0	1	0.9	0.2	0.4
S8	0.5	0.3	1	1	0.9	0	1	0.2	0.3	0.9
S9	0.7	0.2	0.9	1	0.3	1	1	0.3	0.4	0.2
S10	0.6	0.8	0.8	1	0.2	0.6	0.9	0.4	0.5	0.3

B) Evaluation Measure

The performance of the text summarization system is evaluated using the evaluation measures like Precision, Recall, F-Measure. Precision evaluates the correctness of the sentences in the summary. Recall is used to evaluate the relevant sentences included in the summary. F-Measure is the weighted harmonic mean of the measure Precision and Recall.

$$Precision = \frac{|\{\text{Retrieved sentences}\} \cap \{\text{Relevant sentences}\}|}{|\{\text{Retrieved sentences}\}|}$$

------- (2)

$$Recall = \frac{|\{\text{Retrived sentences}\} \cap \{\text{Relevant sentences}\}|}{|\{\text{Relevant sentences}\}|}$$

------- (3)

$$\text{F-measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

---- (4)

C) Results

The Resultant table, which found that the Fuzzy Logic Method have the highest reasoning capacity when compared with the Neural Network Method. Our evaluation shows that the Average precision, recall and F- Measure was high in the Fuzzy Logic Method when compared with Neural Network method.

TABLE 2
THE COMPARISON OF AVERAGE PRECISION, RECALL AND F-MEASURE SCORE AMONG THE TWO METHODS

Text Summarizer	Average value		
	Precision	Recall	F-Measure
Fuzzy Logic	0.48629	0.41269	0.46823
Neural Network	0.47598	0.40964	0.42763

Table 2 shows the comparison for the average precision, recall and f-measure score between Fuzzy Logic and Neural Networks for the same 50 Legal Judgement Documents, which is collected from the legal website.

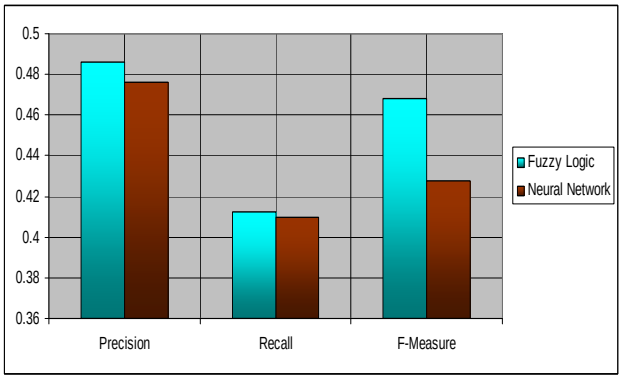

Fig. 5 shows graphical representation for the comparison of average Precision, Recall & F-Measure Score for the two methods.

The results are shown in Table 2, The Fuzzy Logic method shows the average Precision of 0.48629, Recall of 0.41269 and F-Measure of 0.46823. The Neural Network method shows the average Precision of 0.47598, Recall of 0.40964 and F-Measure of 0.42763.

V. CONCLUSIONS AND FUTURE WORK

In this paper, An Automatic Text Summarization system is implemented and investigated using two most popular methods, Neural Network and Fuzzy Logic Method, which involves feature based extraction also. Here 13 Feature Extraction methods were used to improve the quality of summary and the current system is generated by a single document summarization technique. The system is tested with the input of 50 Legal Judgement documents, which is collected from the legal website. The result shows that finest average precision, recall and f-measure were produced by the Fuzzy Logic method when compared to Neural Network method. Certainly, the experimental results shows the improvements in the quality of summary by using Fuzzy Logic Method. In future, the current system is proposed by combining the Fuzzy Logic and Machine Learning Techniques.

REFERENCES

[1] A. Archana, C. Sunitha, "An Overview of Document Summarization Techniques", International Journal on Advanced Computer Theory and Engineering, P: 113-118, 2013.

[2] F. Kyoomarsi, H. Khosravi, P.K. Dehkordy, "Optimizing Text Summarization Based on Fuzzy Logic", IEEE- Computer and Information Science, P: 347-352, 2008.

[3] M. Esther Hannah, T.V.Geetha, Saswati Mukherjee, "Automatic Extractive Text Summarization Based on Fuzzy Logic: A Sentence Oriented Approach, Springer-Lecture Notes in Computer Sceince, P: 530-538, 2011.
[4] S. Rucha, S.S.Apte, "Improvement of Text Summarization using Fuzzy Logic Based method", IOSR Journal of Computer Engineering, P:5-10, 2012.
[5] Ladda Suanamali, Naomic Salim, Mohammed Salem Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization", International Journal of Computer Science and Information Security, P: 150 -156, 2009.
[6] Kyoomarsi F,Khosravi H, Eslami E, and Davoudi M, "Extraction Based Text Summarization using Fuzzy Analysis", Iranian Journal of Fuzzy Systems, Vol. 7, No. 3,P: 15-32, 2010.
[7] Sayantani Gosh, Sudipta Roy, Samir K. Bandyopadhyay,"A Tutorial Review on Text Mining Algorithms", International journal of Advanced Research in Computer and Communication Engineering, 2012.
[8] D.Y.Sakhare, Dr.Raj Kumar, "Syntactic and Sentence Feature Based Hybrid Approach for Text Summarization"I.J. Information Technology and Computer Science, P: 38-46, 2014.
[9] Dipti, Sakhare, Raj Kumar, "Neural Network Based Approach to study the effect of Feature Selection on Document Summarization", International Journal of Engineering and Technology, P:2585 – 2593, 2013.
[10] L.Yulia, G.Alexander, Rene Amulfo, Gracia-Hemandez, "Terms Derived from Frequent Sequences for Extractive Text Summarization", CICLing, LNCS, Springer, pp.593-604, 2008.
[11] Rajesh Prasad, Uday Kulkarni, " Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization", Journal of Computer Science, P: 1366 – 1376, 2010.
[12] Rafeeq Al- Hashemi, "Text Summarization Extraction System (TSES) using Extracted Keyword", International Arab Journal of E-Tehnology", P: 164-168, 2010.
[13] S. Santhana Megala, A. Marimuthu, "A Study on Text Summarization Techniques and its Applications", National Conference on Recent Trends and Advances in Information Technology, 2012, P: 5.
[14] S. Santhana Megala, A. Marimuthu, "A Comparative Analysis of Legal Text Summarization", International Conference on Design and Applications on Structures, Drives, Communicational and Computing Devices, 2012.
[15] S. Santhana Megala, A. Marimuthu, A. Kavitha," Improvised Stemming Algorithm: TWIG", International Journal of Advanced Research in Computer Science and Software Engineering, P: 168-171, 2013.

AUTHORS' PROFILES

- S.Santhana Megala is currently pursuing Ph.D in Computer Science in PRIST University, Thanjavur, Tamil Nadu and working as an Assistant Professor in SNMV College of Arts & Science, Coimbatore, Tamil Nadu, India.
- Dr. A.Kavitha is currently working as an Assistant Professor in Computer Science, Kongunadu Artsand Science College, Coimbatore, Tamil Nadu, India.
- Dr. A. Marimuthu is currently working as an Associate Professor in Government College of Arts & Science, Coimbatore, Tamil Nadu, India.

.