Enforcing Text Summarization using Fuzzy Logic

Rajesh D. Shinde[#1], Suraj H. Routela[#2], Savita S. Jadhav[#3] , Smita R.Sagare[#4]

[#1234]*Fellow Internship*
[#] *Innovatus Technologies, Pune, Maharashtra, India*

***Abstract*— In today's modern era of information revolution, e-business expands rapidly with a large volume of document. In order to analyze the importance data from the document "Text Summarization" will be useful in serving the need of user. There are two main methodologies to perform text summarization -Extraction and Abstraction. Summarization by Extraction involves selection of most frequent words from the original text, whereas abstraction method uses linguistic method to interpret the text. A rouge and Pyramid method is well utilized to extract text for summarization. Previously adopted text summarization method uses information fragment that has been weighted as an important fragment by human summaries for the text. The main drawback of abstractive method is that, it understands the original text and re-telling it into shorter version. Our proposed system uses extraction method i.e. semantic matching instead of lexical matching. The fuzzy logic is use to classify the sentences based on feature score to get the summary of a whole document.**

***Keywords*—** Pre-processing, Sentence Features, Fuzzy Logic, tf_idf, crisp values,

I. Introduction

With the tons of information being uploaded in day today life on the internet and the growth in the quantity and complexity of information source available on the World Wide Web. It has become necessary to provide a system for the user, which helps to extract information from the available document. Text summarization plays vital role in interpreting large volume of information from the available document.

The role of text summarization is to present the most important information from the text in the shorter version without changing meaning of the original text. Summarization can be classified in two types Extraction and Abstraction. Extraction of the document is the selection of sentence that has highest score among other document. Wherein abstraction involve use of linguistic method and extract the sentence together to constitute something new, that is not present in the source, and substitute them in the summary with new concept. [1] Proposed, a technique for summarization of document using neural network that combines the relevant characteristics of sentence that should be included in the summary of article. Later it generalizes the characteristics and filter to summarize the article. [2] Narrates, text summarization based on Naïve Bayes, that uses vocabulary and WordNet to increase natural language processing in summary system and also used to be automatic text summarization.

The proposed system describes mainly three parts A] Pre-Processing B] Feature Extraction C] Fuzzifier. Pre-Processing including four major procedures namely:

- Sentence Segmentation
- Removing Stop Words
- Removing Stemming Words
- Removing Special Characters

The second step in summarization is extraction of features of document based on

- sentence length
- sentence location [3]
- term frequency [4]
- number of words occurring in title [5]
- number of proper noun [6]
- sentence-to-sentence similarity
- thematic word
- number of numerical data[7]

The feature score are then feed to Fuzzy logic for decision support which extract important sentence with powerful reasoning.

The rest of the paper is organized as follows. Section II, presents literature survey. Section III describes proposed methodology. Section IV, presents result and discussion and finally section V, describes the future scope and conclusion.

II. Literature Servey

Previously,[8] Proposed Theory to perform text summarization based on the structural aspect of text known as Rhetorical Structure Theory (RST) Tree, for all the segment in the text. According to RST, a logical text may be structured as a dissertation tree for each document in the form of segments. Summarizations consider the fact that text segment in the tree are classified according to their importance. However, there were some deficiencies, that whenever a document does not contain many Rhetorical relations there is a little telling which segment is more important than the other is.

[9] Narrates, a Dist. Al algorithm based on inter-pattern which construct a single hidden layer of hyper spherical threshold neurons. The weights and threshold of neurons are determined directly by comparing the intern-pattern distance of training patterns. These neurons are trained by an applicable weight-training rule and time-consuming perceptron training procedure. Structure gives non-deterministic result because their behaviour is not identical in different runs for a given training set. Must Link and Cannot Link(MLCL) is grounded on clustering algorithm [10], that develop a meaningful cluster which maintain the relationship between the key term of document.

[11]Discuss, Pyramid method for automation of summary, which weight on the task of determining if a summary express the same content as set of manual model. The drawback that lies within pyramid method is, that it

implement the SCUs which is set of contributor (text fragments) that has been weighted as an important fragment by human summaries for the text. [12] Carried out text summarization, based on extractive multi-document summarization algorithm, which uses graph for signifying the structure of text as well as relationship sentence of the document. [13] Describes, grouping of syntactic structure and feature based technique for summarization of text. Here both neural network are trained based on features score and syntactic structure of sentence. Later, both neural network are combine with weighted average to find the sentence score of the sentence.

III. Proposed Methodology

In this section, we describe the approach of elevating the text summarization. In text Summarization there are three core parts are A) Pre-Processing B) Feature Extraction C) Fuzzifier

A) *Pre-Processing*

There are four core activities executed in this step:

Step 1: Sentence segmentation is boundary detection and separating source text into sentence.

Step 2: Removing Special Character is replacing special symbols with empty character in input document.

Step 3: Afterward, Removing Stop Words, stop words are the words which seem repeatedly in document but provide fewer meaning in recognizing the important content of the document such as 'a', 'an', 'the', etc..

Step 4: Word stemming is the process of removing prefixes and suffixes of each word to bring to its base form. For Example goes remain go, going remain go, etc.

B) *Feature Extraction*

In this process, each sentence of the document is characterized by an attribute vector of features. These features are attributes that attempt to signify the data used for their task. We are emphasis on eight features for each sentence. Each feature is given a value between '0' and '1'. There are eight features as follows:

1) Title feature

The word in sentence that also occurs in title gives high score. This is resolute by calculating the number of matches between the content words in a sentence and the words in the title. We calculate the score for this title feature, which is the ratio of the number of words in the sentence that arise in the title over the number of words in title.

$$S_{F1}(S) = \frac{\text{No.of Title word in S}}{\text{No. of Word in Title}} \quad (1)$$

2) Sentence Length

This feature is beneficial to filter out short sentences such as datelines and journalist names, venues, time commonly found in news articles. Such sentences are not predictable to belong to the summary. We are using the length of the sentence, which is the proportion of the number of words arising in the sentence over the number of words arising in the longest sentence of the document.

$$S_{F2}(S) = \frac{\text{No.of Words occuring in S}}{\text{No.of Word occurring in longest sentence}} \quad (2)$$

3) Term Weight

The frequency of term incidences within a document has frequently been used for calculating the rank of sentence. The score of a sentence can be intended as the sum of the score of words in the sentence. The score of significant score *wi* of word *i* can be intended by the traditional *tf_idf* method as follows [14]. We applied this method to *tf_isf* (Term frequency, Inverse sentence frequency).

$$W_i = tf_i x isf_i = tf_i x \log\frac{N}{n_i} \quad (3)$$

Where tf_i is the term frequency of word *i* in the document, *N* is the total number of sentences, and n_i is number of sentences in which word *i* arises. This feature can be intended as follows.

$$S_{F3}(S) = \frac{\sum_{i=1}^{k} W_i(S)}{Max(\sum_{i=1}^{k} W_i(S_i^N))} \quad (4)$$

k is number of words in sentence.

4) Sentence Position

Whether it is the first 5 sentences in the article, sentence location in text gives the rank of the sentences. This feature can contain several stuffs such as the location of a sentence in the document, section, and paragraph, etc., suggested the first sentence is highest ranking. The score for this feature: we consider the first 5 sentences in the paragraph. This feature score is intended as the succeeding equation (5).

$$S_{F4}(S) = \frac{5}{5}\text{for 1st}, \frac{4}{5}\text{for 2nd}, \frac{3}{5}\text{for 3rd}, \frac{2}{5}\text{for 4th}, \frac{1}{5}\text{for 5th}, \frac{0}{5}\text{for other sentences} \quad (5)$$

5) Sentence to Sentence Similarity

This feature is a similarity among sentences. For each sentence *S*, the similarity between *S* and each other sentence is computed by the cosine similarity measure with a resulting value between 0 and 1 [15]. The term weight *wi and wj* of term t to n term in sentence *Si* and *Sj* are denoted as the vectors. The similarity of each sentence couple is intended based on similarity formula (6).

$$sim(S_i, S_j) = \frac{\sum_{t=1}^{n} W_{it} x W_{jt}}{\sqrt{\sum_{t=1}^{n} W_{it}^2} x \sqrt{\sum_{t=1}^{n} W_{jt}^2}} \quad (6)$$

Now, the score of this feature for a sentence *S* is acquired by computing the proportion of the summary of sentence similarity of sentence *S* with each other sentence over the maximum of summary

$$S_{F5}(S) = \frac{\sum Sim(S_i, S_j)}{Max(\sum Sim(S_i, S_j))} \quad (7)$$

6) Proper Noun

The sentence that holds more proper nouns (name entity) is an essential and it is most probably included in the document summary. The score for this feature is intended as the proportion of the number of proper nouns that arise in sentence over the sentence length.

$$S_{F6}(S) = \frac{\text{No.of Proper nouns in S}}{\text{Sentence Length(S)}} \quad (8)$$

7) Thematic Word

The number of thematic word in sentence, this feature is essential because terms that arise frequently in a document are perhaps related to matter. The number of thematic words specifies the words with maximum probable relativity. We used the top 5 most repeated content word for consideration as thematic. The score for this feature is calculated as the proportion of the number of thematic words that arise in the sentence over the maximum summary of thematic words in the sentence.

$$S_{F7}(S) = \frac{\text{No.of Thematic word in } S}{\text{Max(No.of Thematic Word)}} \qquad \textbf{(9)}$$

8) Numerical Data

The number of numerical data in sentence, sentence that holds numerical data is important and it is most probably included in the document summary. The score for this feature is intended as the proportion of the number of numerical data that arise in sentence over the sentence length.

$$S_{F8}(S) = \frac{\text{No.of Numerical data in } S}{\text{Sentence Length}(S)} \qquad \textbf{(10)}$$

C) Fuzzifier

The aim of text summarization is based on extraction method of sentence selection. One of the methods to get the appropriate sentences is to consign some numerical measure of a sentence for the Summary known as sentence weighting and then select the best ones. Therefore, the features score of each sentence that we termed in the prior section are used to acquire the significant sentences. In this section, we use method to extract the essential sentences: text summarization based on fuzzy logic method. The system involves of the following core

Steps:

Step 1: In the fuzzifier, crisp inputs are taken, which are result of the feature extraction.

Step 2: After fuzzification, the inference engine refers to the rule base containing fuzzy IFTHEN rules.

Step 3: In the last step, we get the final sentence score. In inference engine, the most important part is the definition of fuzzy IF-THEN rules. The essential sentences are extracted from these rules according to our features criteria. Sample of IF-THEN rules are described below.

IF (NoWordInTitle > 0.81) and (SentenceLength > 0.81) and (TermFreq > 0.81) and (SentencePosition > 0.81) and (SentenceSimilarity > 0.81) and (NoProperNoun > 0.81) and (NoThematicWord > 0.81) and (NumbericalData > 0.81) THEN (Sentence is important).

After this process all the document sentences are ranked in a descending order according to their scores. A set of uppermost score sentences are extracted as a document summary. The whole process of text summarization is briefed as shown below in figure 1.

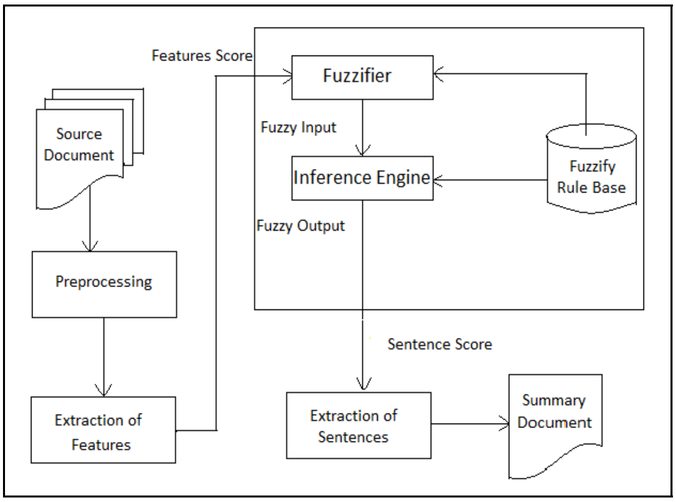

Figure 1. Text summarization based on fuzzy logic system architecture

IV. Result And Discussion

The performance and scalability of Text summarization evaluated by experiments on window7 with Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz and 4GB RAM. These experiments used Text Font of Entire Document .txt, .doc, .pdf files as an input. Text summarization implemented in JAVA using Net Beans IDE 6.9.1.

Table 1 evaluates the performance of proposed system with the MS-WORD Summarizer for following parameter in our experiment.

Number Of Sentences in a document	Ms-Word Summary	Fuzzifier Summary
10	5	4
20	6	4
30	8	6
40	11	7
50	13	7

Table 1. Evaluation of summary factor

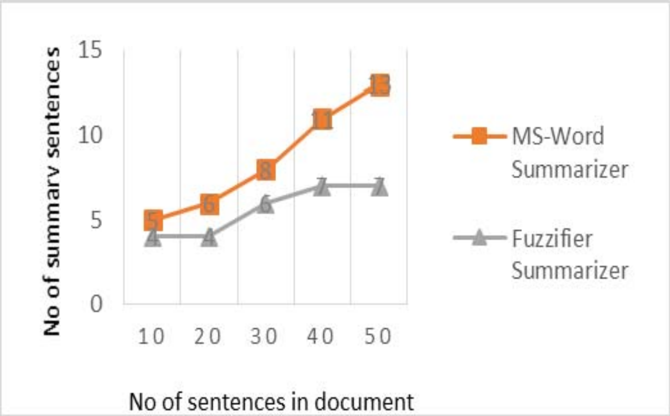

Figure 2. Fuzzy Summarizer Performance

The plot in figure 2 clearly indicates the proposed system of fuzzy summarizer yields lesser number of sentences which is actually the abstract of that document and it is clearly over performed than the MS-Word summarizer.

V. CONCLUSION AND FUTURE WORK

In this paper, we have extracted important features for each sentence of the document. The experimental result is based on fuzzy logic to improve the quality of summary. We extracted the important features for each sentence of the document represented as the vector of features containing the following elements: title feature, sentence length, term weight, sentence position, sentence to sentence similarity, proper noun, thematic word and numerical data.

Existing method uses machine translation for document translation or summary translation. However, machine translation services are far from acceptable which result in the quality of cross language.

The system can enhance to perform summarization on cross language platform where a summary for a document in one language can yield summary in many different language by using other language dictionaries.

REFERENCES

1) Khosrow Kaikhah, "Text Summarization Using Neural Networks",Deapartment of Computer Science,Texas State University,San Macros,Texas 78666, 1988.
2) Ha Nguyen Thi Thu, "An Optimization Text Summarization Method Based on Naïve Bayes and Topic Word for Single Syllable Language",Applied Mathematical Sciences, Vol. 8, 2014no. 3, 99 – 115, HIKARI Ltd, www.m-ikari.com http://dx.doi.org/10.12988/ams.2014.36319
3) M.A. Fattah and Fuji Ren, "Automatic Text Summarization" In proceedings of World Academy of Science, Engineering and Technology Volume 27. pp 192-195. February 2008.
4) G. Salton, "Automatic Text Processing: The Transfor-mation, Analysis, and Retrieval of Information by Computer" Addison-Wesley Publishing Company, 1989.
5) G. Salton, C. Buckley, "Term-weighting approaches in automatic text retrieval" Information Processing and Management 24, 1988. 513-523. Reprinted in: Sparck-Jones, K.; Willet, P. (eds.) Readings in I. Retrieval. Morgan Kaufmann. Pp.323-328, 1997.
6) J. Kupiec. , J. Pedersen, and F. Chen, "A Trainable Document Summarizer" In Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Seattle, WA, pp.68-73, 1995.
7) C.Y. Lin, "Training a selection function for extraction" In Proceedings of the eighth international conference on Information and knowledge management, Kansas City, Missouri, United States. pp.55–62, 1999.
8) Vinícius Rodrigues Uzêda, Thiago Alexandre Salgueiro Pardo, Maria das Graças Volpe Nunes, "Evaluation of Automatic Text Summarization Methods Based on Rhetorical Structure Theory",Núcleo Interinstitucional de Lingüística Computacional (NILC) Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo CP 668 – ICMC-USP, 13.560-970 São Carlos-SP, Brazil http://www.nilc.icmc.usp.brvruzeda@gmail.com,{taspardo,gracan}@icmc.usp.br, 2008.
9) Jihoon Yang *, Rajesh Parekh, Vasant Honavar , "DistAl: An inter-pattern distance-based constructive learning algorithm" ,Intelligent Data Analysis 3(1999) 55-73, Received 3 May 1998; received in revised form 26 May 1998; accepted 2 November 1998.
10) 1 J.Dafni Rose, 2 Divya D. Dev, 3 C.R.Rene Robin, "A Novel Approach For Text Clustering Using Must Link And Cannot Link Algorithm", Journal of Theoretical and Applied Information Technology, 10th Feb, 2014.
11) Aaron Harnly, Ani Nenkova, Rebecca Passonneau, Owen Rambow, Center for Computational Learning Systems "Automation of Summary Evaluation by the Pyramid Method", Columbia University New York, NY, USA, 2005.
12) Amit.S.Zore, Aarati Deshpande, "Extractive Multi Document Summarizer Algorithm" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, 5245-5248
13) D.Y. Sakhare, Raj Kumar ,"Syntactic and Sentence Feature Based Hybrid Approach for Text Summarization", International Journal of Information Technology and Computer Science (IJITCS), 3 Feb, 2014.
14) M. Wasson, "Using leading text for news summaries: Evaluation results and implications for commercial summarization applications" In Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the ACL. Pp.1364-1368, 1998.
15) Amy J.C. Trappey, Charles V. Trappey, "An R&D knowledge management method for patent document summarization", Industrial Management & Data Systems, vol.108. Pp.245-257, 2008.