Journal of Informatics and Mathematical Sciences
Vol. 9, No. 3, pp. 547–555, 2017
ISSN 0975-5748 (online); 0974-875X (print)
Published by RGN Publications

http://www.rgnpublications.com

Proceedings of the Conference
Current Scenario in Pure and Applied Mathematics
December 22-23, 2016
Kongunadu Arts and Science College (Autonomous)
Coimbatore, Tamil Nadu, India

Research Article

A Hybrid Text Summarization Approach

Shrabanti Mandal[1], Girish Kumar Singh[1,*] and Anita Pal[2]

[1] *Department of Computer Science and Applications, Dr. Harisingh Gour Central University, Sagar, Madya Pradesh, India*
[2] *Department of Mathematics, National Institute of Technology, Durgapur, West Bengal, India*
***Corresponding author:** gkrsingh@gmail.com

Abstract. Today, internet is the storage of huge information. Therefore it is very serious issue to get data fast and efficiently. Text summarization plays an important role in the field of information retrieval. Text summarization is a process of representing a text in concise way with same sense. This hybrid approach mainly based on extractive summarization. The proposed approach combines the concept of statistical measure, sentiment analysis and finally uses the concept of fuzzy logic to select sentence. Based on the level of importance of the sentence, summary is created.

Keywords. Text summarization; Sentiment analysis; Feature extraction; Fuzzy concept

MSC. 03Bxx

Received: January 5, 2017 **Accepted:** March 4, 2017

1. Introduction

Before going to the details of the paper we are trying to get the overview of text summarization. A summary is a concise form of text that is composed from one or more texts that gives

important information in the original text [16]. The purpose of automatic text summarization is to present the source text into a shorter version with semantics [1]. Summary reduces the reading time. There are two types of text summarization methods which are classified into extractive and abstractive summarization. Extractive summarization means selecting the important sentence from the given document. Abstractive summarization means express the meaning of the document in natural language [16].

Two different group of summarization is also available indicative and informative. Indicative summarization gives the main theme of the document in short. The length of this summary is only 5 percent of the given document. Informative summarization gives brief information of the document and size of summary is 20 percent of the given document.

2. Related Work

The work done in [11] by H.P. Luhn was considered as the one of the earliest work in the field of text summarization. In this research a frequency based model has been proposed for single document summarization. The main idea of this research is frequency of words play an important role to calculate the priority of any sentence. Radev et al. in [16] has defined a text summary as "a text that is produced from one or more texts that convey important information in the original texts, and that is no longer than half of the original text and usually significant less than that". Alguliev et al. in [1] has discussed about the two types of text summarization that is extractive and abstractive. As explained, [12, 23] abstractive summarization combines three concepts like Information Fusion, Sentence Compression and Reformation. Abstractive summarization may contain new sentences, phrases, words even which are not present in the source document. Although till now a lot of research has been happened in the last decades in the area of NLP (*Natural Language Processing*), and NLG (*Natural Language Generation*) but still we are not near abstractive summarization. The actual challenge is a generation of new sentences, new phases, along with produced summary must retain the same meaning as the same source document has.

Extractive summarization based on extractive entities, entities may be sentence, sub part of sentence, phrase or a word. In [9, 15, 17, 25] different statistical methods have been used for summarization. In our paper we are trying to use these methods. Aliguliyev in [2] has explained summarization as a three steps process

(i) Analysis of text.

(ii) Transformation- as summary representation, and

(iii) Synthesis-produce an appropriate summary.

In [4, 13] query based summarization has been discussed. The survey of text summarization has been explained in [19, 22]. Babar et al. in [3] has proposed a method for improving the performance of text summarization. The statistical method and Semantic Analysis have been merged to get the better result. Yadav et al. in [24] proposed a "Hybrid approach for single text document summarization using statistical and sentiment features". In this research statistical

measure and sentiment score have combined to calculate the impotence of a sentence. The objective of summarization is to miximize the coverage and minimizes the redundancy. To achieve the gole, a research has been done in a simple way to include a sentence to the summary by modified cosine similarity threshold value by Kamal Sarkar [20]. For better summarization sentiment analysis plays a crucial role. Many researchers worked on emotion detection. Das et al. in [5, 10] has propose a machine learning approach to calculate the emotion in a word and a sentence labeled in Bengali blogs.

3. Fuzzy Logic

Fuzzy logic is an approach to compute based on "degrees of truth" rather than the usual "true or false" (1 or 0) Boolean **logic** on which the modern computer is based. The idea of **fuzzy logic** was first advanced by Dr. Lotfi Zadeh of University of California at Berkeley in the 1960s. In this work we have used the fuzzy concept for selecting the mostly desirable sentences.

4. Steps of Summarization

To summarize the text the following approaches are followed:

(i) Preprocessing,
(ii) Feature extraction,
(iii) Fuzzy logic concept,
(iv) Sentiment scoring,
(v) Sentence selection and assembly.

4.1 Preprocessing

Preprocessing consists of four different steps. Segmentation is the first step, is used to break a documents in sentences. Then second steps is used to remove the stop words. Stop word means frequently occurring words that has less meaning and containing noise. Next is tokenization that will break the input text into separate token. Punctuation marks, spaces and word terminators are the word breaking characters. The final step is word streaming is used to remove the prefix and suffix for converting to the root form of every word.

4.2 Feature Extraction

A text document is represented by set, $D = \{S_1, S_2, \ldots, S_k\}$, where S_i signifies a sentence contained in the document D. The statistical measure includes Title word, Sentence length, Sentence position, Numerical data, Term weight, Sentence similarity, Existence of Thematic words, Proper nouns etc.

(i) Title Word: If a sentence contains the words occurring in title, then highest weightage has been given to this sentence. This feature is computed as follows:

$$F_1 = \frac{N_t}{N_{Total}},$$

where N_t means numbers of title words match with the sentence and N_{Total} means numbers of total words occurring in the title after removing the stop word.

(ii) Sentence Length: Length of the sentence is very important. The sentences whose length is too short are eliminated. For every sentence the normalized length is calculated by

$$F_2 = \frac{TWS}{TWLS},$$

where TWS means number of words belongs to the sentence and $TWLS$ means number of words belongs to the longest sentence.

(iii) Sentence Position: The highest score has been given to the sentence which occurs first in the document. If the document contains n sentences then this feature is computed as

$$F_3(S_1) = \frac{n}{n};\ F_3(S_2) = \frac{n-1}{n};\ F_3(S_3) = \frac{n-2}{n};\ F_3(S_4) = \frac{n-3}{n} \text{ and so on.}$$

(iv) Numerical Data: The sentence carrying numerical data is considered as important so it is selected for summary. This is calculated by $F_4(S_i) = \frac{N_D}{S_L}$. Here N_D means number of numerical data occurring in sentence S_i and S_L means total length of the sentence.

(v) Thematic Words: These are domain specific words with maximum possible relativity. The ratio of the number of thematic words that occurs in a sentence over the maximum number of thematic words in a sentence gives the score of each feature as:

$$F_5(S_i) = \frac{NTW}{MTW}$$

where NTW represents the number of thematic words in sentence S_i and MTW represents the maximum number of thematic words.

(vi) Sentence to Sentence Similarity: To calculate the similarity between sentence to sentence token matching concepts is used here. A Sim[N][N] matrix has been generated where N is number of sentence in document

Sim[N][N] =

Sentence	S_1	S_2	S_3	S_4	S_5
S_1	1				
S_2		1			
S_3			1		
S_4				1	
S_5					1

(vii) Term weight: In **information retrieval**, tf–idf, short for **term** frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a **word** is to a document in a collection or corpus. It is often used as a weighting factor in **information retrieval** and text mining.

$$F_7 = \frac{\sum TF_i}{\max \sum TF_i}$$

(viii) Proper Nouns: Importance of the sentence directly varies with the number of proper nouns.

$F_8 = \frac{PN_i}{L_i}$, where PN_i represent the number of proper noun belongs to S_i and L_i stands for length of S_i.

4.3 Fuzzy Logic Scoring

By extracting the eight features of a sentence we get the eight feature vectors corresponding to each sentence. The feature vectors are derived using the fuzzy logic rules. In our experiment mainly triangular membership function is used for defuzzyfication of feature vector.

4.4 Sentiment Scoring

Sentiment is defined as a tendency to experience certain emotions in relation to a particular object or person [7, 18] and it can be expressed usually in writing, such as products review, websites, blogs, forums, etc. Sometimes, opinions are hidden within long sentences, making them difficult to reads and extract.

There is a technique called, sentiment analysis that relates to natural language processing, text mining and linguistics [21]. The main goal of sentiment analysis is to identify the polarity of natural language text [8], which is not limited to positive and negative [14]. Sentiment analysis can be referred to as opinion mining as it studies people's opinions, appraisals and emotions towards entities, events and their attributes [11] Sentiment can be analysis in different levels like document level, sentence level and word-level. There are some approaches for sentiment analyzed by using SentiWordNet. Here we are going to use the SentiWordNet for sentence level sentiment analysis

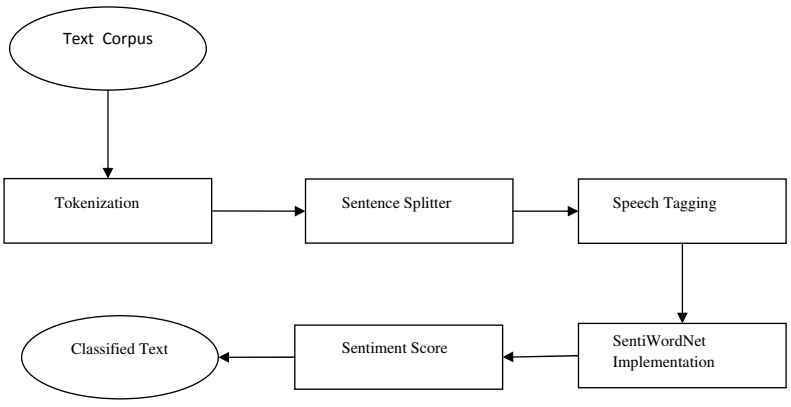

Figure 1. Process of sentiment analysis

5. Proposed Architecture

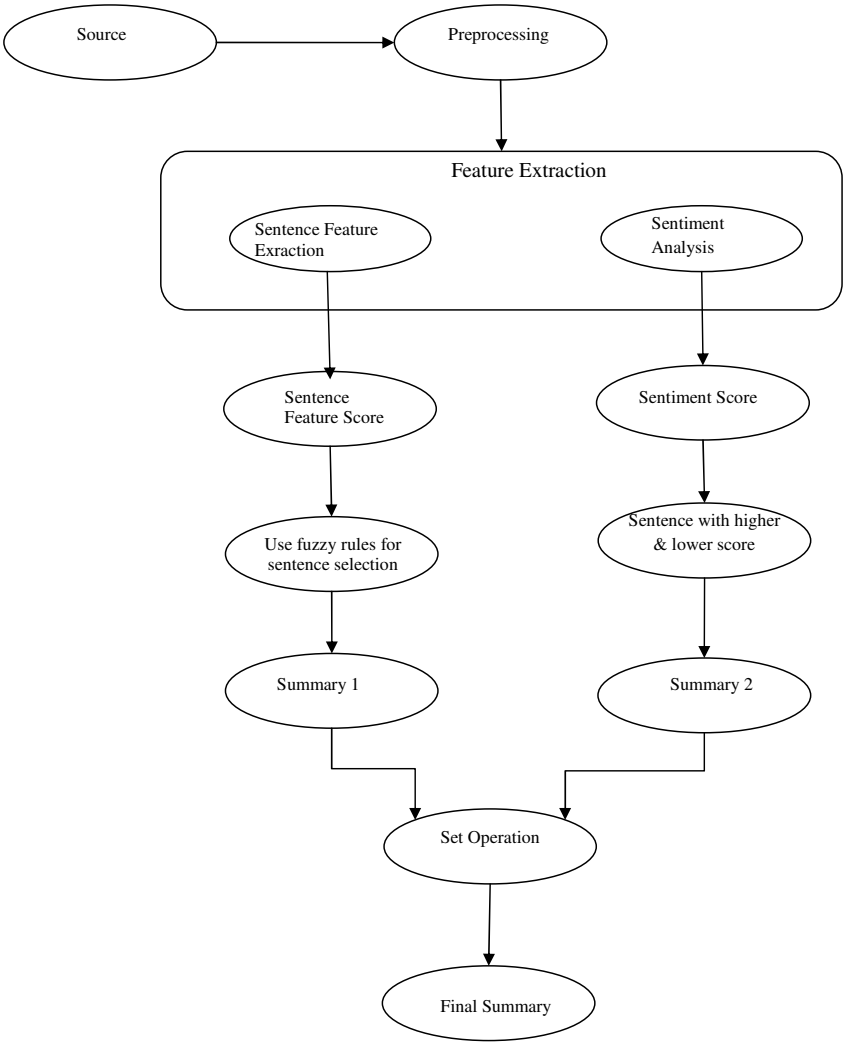

Figure 2. Proposed Architecture

The data flowing of our proposed method has been shown by Figure 2. Take the text document as input and preprocess it. Then extract the feature and apply fuzzy rules for sentence selection. The selected sentences have been considered as summary1. In other hand sentiwordnet has been used to calculate the sentiment score of sentences. The sentences of high and low score are selected at summary2. Finally applied the set operation to generate the final summary.

Mathematical Expression

Consider a document is represented by D and sentence is represented by s. So we have a document in the form like

$$D = \{s_1, s_2, s_3, \ldots, s_n\}.$$

Then apply the fuzzy logic and get the summary1(Sum1) like this

$$\text{Sum1} = \{s_1, s_3, s_6\}$$

Simultaneously calculate the sentiment score of document D. Then select the sentence with high score and low score.

Suppose we get summarry2(Sum2) is given bellow

$$\text{Sum2} = \{s_3, s_7, s_9\}$$

Then calculate the final summary(Fsum)

$$\text{Fsum} = \{\text{Sum1} \wedge \text{Sum2}\} \cup \{\max_{score}(\text{uncommon(Sum1,Sum2)})\}$$

$$\text{Fsum} = \{s_3\} \cup \left\{\max_{score}(s_1, s_6, s_7, s_9)\right\}$$

$$\text{Fsum} = \{s_3\} \cup \{s_1, s_9\}$$

$$\text{Fsum} = \{s_1, s_3, s_9\}$$

Result Analysis

Table 1. Precision, recall, accuracy values of proposed method

Dataset	Precision	Recall	Accuracy
Dataset 1	40	66.67	44.44
Dataset 2	66.67	66.67	81.18
Dataset 3	50	50	60
Dataset 4	50	50	70
Dataset 5	75	75	77.78
Dataset 6	33.33	50	50
Dataset 7	66.67	66.67	66.67
Average	**55.52**	**60.72**	**64.29**

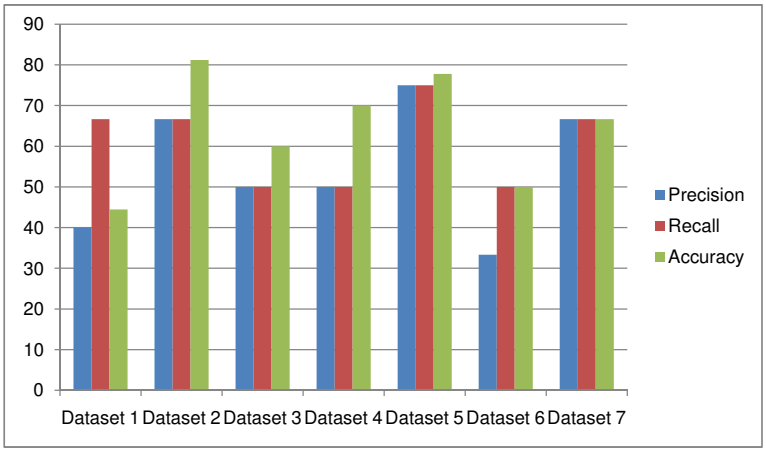

Figure 3. Evaluation graph of proposed method

6. Conclusion

In this paper the proposed method is based on extractive summarization. In extractive summarization to select the relevant sentence is a very complex task. To reach the goal we have use the statistical features of text as well as considered the sentiment analysis. Here sentiment analysis helps to choose the silent which has importance to document. Finally, we use the fuzzy concept to generate the summary. In future we try to extent this concept for multi-document summarization.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] R.M. Alguliev, R.M. Aliguliyev and C.A. Mehdiyev, Sentence selection for generic document summarization using an adaptive differential evolution algorithm, *Swarm and Evolutionary Computation* **1** (4) (2011), 213 – 222.

[2] R.M. Aliguliyev, Automatic document summarization by sentence extraction, *Journal of Computational Technologies* **12** (5) (2007), 5 – 15.

[3] S.A. Babar and P.D. Patil, Improving performance of text summarization, *International Conference on Information and Communication Technologies* (ICICT 2014), *Procedia Computer Science* **46** (2015), 354 – 363.

[4] J. Carbonell and J. Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries, in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335 – 336, ACM (1998).

[5] D. Das and S. Bandyopadhyay, Word to sentence level emotion tagging for bengali blogs, *ACL-IJCNLP 2009*, pp. 149 – 152, Suntec, Singapore (2009).

[6] D. Das and S. Bandyopadhyay, Developing bengali WordNet affect for analyzing emotion, *ICCPOL-2010*, California, USA (2010).

[7] A. Hogenboom, F. Boon and F. Frasincar, A Statistical Approach to Star Rating Classification of Sentiment, in *Management Intelligent Systems*, Vol. **171**, J. Casillas, F.J. Martínez-López and J.M. Corchado Rodríguez (eds.), Springer, Berlin—Heidelberg (2012), pp. 251 – 260.

[8] J. Karlgren, M. Sahlgren, F. Olsson, F. Espinoza and O. Hamfors, Usefulness of Sentiment Analysis, in *Advances in Information Retrieval*, Vol. **7224**, R. Baeza-Yates, A. Vries, H. Zaragoza, B.B. Cambazoglu, V. Murdock, R. Lempel and F. Silvestri (eds.), Springer, Berlin—Heidelberg (2012), pp. 426 – 435.

[9] Y. Ko and J. Seo, An effective sentence-extraction technique using contextual information and statistical approaches for text summarization, *Pattern Recognition Letters* **29** (9) (2008), 1366 – 1371.

[10] C.J. Leuba, *Man: A General Psychology*, Holt, Rinehart and Winston (1961).

[11] H.P. Luhn, The automatic creation of literature abstracts, *IBM Journal of Research and Development* **2** (2) (1958), 159 – 165.

[12] I. Mani and M.T. Maybury (eds.), *Advances in Automatic Text Summarization*, Vol. **293**, IT Press, Cambridge (1999).

[13] Y. Ouyang, W. Li, S. Li and Q. Lu, Applying regression models to query-focused multi-document summarization, *Information Processing and Management* **47** (2) (2011), 227 – 237.

[14] B. Pang and L. Lee, Opinion Mining and Sentiment Analysis, *Found. Trends Inf. Retr.* **2** (2008), 1 – 135.

[15] D.R. Radev, S. Blair-Goldensohn and Z. Zhang, *Experiments in Single and Multi-Document Summarization using MEAD*, Ann Arbor, 1001, 48109 (2001).

[16] D.R. Radev, E. Hovy and K. McKeown, Introduction to the special issue on summarization, *Computational Linguistics* **28** (4) (2002), 399 – 408.

[17] D.R. Radev, H. Jing, M. Stys and D. Tam, Centroid-based summarization of multiple documents, *Information Processing and Management* **40** (6) (2004), 919 – 938.

[18] W.K. Richmond, *Teachers and Machines: An Introduction to the Theory and Practice of Programmed Learning*, Collins (1965).

[19] S. Gholamrezazadeh and M.A. Salehi, *A Comprehensive Survey on Text Summarization Systems*, IEEE (2009).

[20] K. Sarkar, Syntactic trimming of extracted sentences for improving extractive multi document summarization, *Journal of Computing* **2** (7) (2010), 177 – 184.

[21] M.A. Shaikh, H. Prendinger and I. Mitsuru, Assessing sentiment of text by semantic dependency and contextual valence analysis, presented at the *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, Lisbon, Portugal (2007).

[22] V. Gupta and G.S. Lehal, A survey of text summarization techniques, *Journal of Emerging Technologies in Web Intelligence* **2** (3) (2010), 258 – 268.

[23] X. Wan, Wan Using only cross-document relationships for both generic and topic-focused multi-document summarizations, *Information Retrieval* **11** (2008), 25 – 29.

[24] C.S. Yadav and A. Sharan, Hybrid approach for single text document summarization using statistical and sentiment features, *International Journal of Information Retrieval Research* **5** (4) (2015), 46 – 70.

[25] J.Y. Yeh, H.R. Ke, W.P. Yang and I.H. Meng, Text summarization using a trainable summarizer and latent semantic analysis, *Information Processing and Management* **41** (1) (2005), 75 – 95.