IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Architecture for Hybrid genetic fuzzy system for text summarization

S. Saiyed[1*], P. Sajja[2]

[1] S.S. Agrawal Institute of Computer science, Navsari, Gujarat, India
[2]G.H.Patel PG Department of computer Science, Sardar Patel University, Vallabh vidhyanagar, Gujarat, India

[*]Corresponding Author: *saziya.saiyed2013@gmail.com*, Tel.: 09924049200

Available online at: www.ijcseonline.org

Accepted: 18/Jun/2018, Published: 30/Jun/2018

***Abstract*—**Massive amount of information in form of text is available on internet. To get useful or important information from this massive amount of information is tough and tedious task. One can get important information by creating summary. Manual creation of summary is again a tough task. Hence research community is developing new approaches for creating automatic summaries; these approaches are called automatic text summarization. There are number of text summarization techniques available, some of them lack with quality of summary and some of them lacks in user specific needs of summary. This paper discusses the architecture for extractive type of text summarization that uses hybrid genetic fuzzy system. The goal of this paper is to give idea about effectiveness of Genetic algorithm and fuzzy logic system together to create good summary.

***Keywords*—** Text Summarization, Extractive summarization, Hybrid Genetic algorithm & Fuzzy system for text summarization

I. INTRODUCTION

In today's fast intensifying world of information, text summarization [1] is very essential and useful tool for understanding text information. There is a lot of text material and documents available on the internet which provides information in large volume than requirement, and creates the situation [2] called "infobesity". To select important information from large amount of information and different sources is difficult for human beings. Due to the volume of information, to manually summarize information is really challenging and complicated task.

The aim of automatic text summarization is to reduce the source text into a compact form which will defend contents and general meaning of source information, and also minimizes reading time and efforts. [2]

Text summarization can be categorized into two categories: Extractive and Abstractive. Extractive text summarization [3] uses statistical and linguistic features to select important paragraphs, lines and words from source text for creation of summary where as abstractive text summarization [4][5] examines and understands the main concept of source text by using linguistic methods and then finds some other notions that can best describe the underlying concept. It then uses the new notions to create summary.

This paper introduces a hybrid approach for text summarization which uses genetic algorithm and fuzzy logic system [6] for text summarization.

This paper is divided in five sections, section I describes introduction, section II describes linguistic features used for text summarization. Section III describes related work. Section IV shows that how the source text is summarized using hybrid genetic fuzzy system and in Section V we concluded paper.

II. LINGUISTIC FEATURES USED FOR TEXT SUMMARIZATION

Some features to be considered for summary are **[3][7][8]**

A. Title word feature:

The source texts Sentences which contains words that are similar to the title word are indicating the concept of the document. Such sentences are having higher chances to get included into summary.

B. *Content word (Keyword) feature:*

Content words or Keywords are generally nouns. They can be determined using term frequency - inverse document frequency (tf-idf) method[9]. Sentences which contain keywords are of greater chances to get included into summary.

C. Sentence Length feature:

Very large and very short sentences are not considered in summary. So they are having lesser chances to get selected for summary.

D. Sentence position feature:

Sentence position feature is very important from the point of view of abstractive text summarization. Usually first and/or

last sentence of first and/or last paragraph of a source text document are additional important and are having higher chances to get included into summary.

E. Proper Noun feature:

Proper noun can be name of an entity, name of place and name of any concept etc. so the Sentences containing proper nouns are having more chances to get included into summary.

F. Biased Word Feature:

Biased list is list of predefined words & it may contain words based on the domain. If a word appearing in a sentence is from biased list of words, then that sentence is important.

G. Sentence-to-Sentence Cohesion:

For each sentence of document, similarity between any two sentences "s1" and each other sentence "s2" of the document is calculated. By summation of all those similarity values, raw value of this feature can be obtained for a specific sentence. The process is repeated for all the sentences.

H. Thematic word:

The terms that occur more frequently in source text are more related to the concept of source document

III. RELATED WORK

A. Term frequency- inverse document frequency(TF-IDF) based method:

In TF-IDF a Bag-of-words model at sentence level, with the usual weighted term-frequency and inverse sentence frequency [9], the sentence-frequency is considered as the number of sentences in the document that contain the term present in query. Scores to the sentence vectors are assigned by similarity to the query and the sentences with highest scores are considered as part of the summary.

B. Clustering based method:

This method is consists of two steps [10]. First sentences are clustered and then important sentences are defined and extracted based on each cluster.

C. Graph theoretic method:

This method considers each sentence as a node of undirected graph [11]. Any one node is connected to the other node with edge if they share some common words. So different topics can be identified by looking different sub graphs of document which are covered in the document. Identification of important sentences to be included in summary can be determined by cardinality of the node. The node with high cardinality represents important statement. Cardinality of node is number of edges connected to the node.

D. Neural network based method:

A neural network is given training to study the features of sentences that can be selected in the summary of the source document for news article [12]. Then the neural network is modified to simplify and join the related features noticeable in summary sentences. Finally, the modified neural network is used as a filter for creating summaries of source news articles.

E. Lexical chain based method:

Lexical chain based method[13] creates summary does not requiring complete semantic understanding , it uses linguistic analysis and semantically related terms are accepted & created the lexical chains of them, from those chains strong chains are identified and then important sentences are extracted from them.

F. Fuzzy logic based method:

This method considers each features of a text such as sentence length, similarity to key word and others as the input of fuzzy system[3][14]. Fuzzy Rules required for summarization are created and stored in the knowledge base of system. Then, a value between zero to one is obtained for each sentence in the output based on sentence characteristics and the available rules in the knowledge base. The obtained value as an output decides the importance of the sentence for the final summary. The input membership function for each feature is divided into three membership functions which are composed of insignificant values (low L), very low (VL), medium (M), significant values (High h) and very high (VH). The important sentences are extracted using IF-THEN rules according to the feature criteria. Text summarization based on fuzzy logic system architecture [6] design usually implicates selecting fuzzy rules and membership function. The fuzzy logic system consists of four components: fuzzifier, inference engine, defuzzifier, and the fuzzy knowledge base. In the fuzzifier, crisp inputs are translated into linguistic values using a membership function to be used to the input linguistic variables. After fuzzification, the inference engine uses rules of rule base containing fuzzy IFTHEN rules to obtain the linguistic values. Finally the output linguistic variables from the inference are converted to the final crisp values by the defuzzifier using membership function for representing the final sentence score.

G. MEAD:

MEAD [15] is platform for multi-lingual summarization and evaluation. Summarization algorithms such as position-based, centroid-based, largest common subsequence and keywords are used in MEAD. It first converts documents in to MEAD's internal format which is XML. Then numbers of features are extracted for each sentence of the grouped to create a combine score for each sentence. Then these scores can be further distinguished after considering possible cross-sentence dependencies like recurring sentences, sequential ordering, source preferences etc.

IV. ARCHITECTURE FOR HYBRID GENETIC FUZZY SYSTEM FOR TEXT SUMMARIZATION

This Architecture is made up of two steps.

A. *Pre processing of text:*

Preprocessing of text includes stop word removal, tokenization and stemming for each sentence of the source document. Stop word removal removes most commonly used words not relevant to document like 'a', 'the' etc from the text. In tokenization each word of text is divided into tokens. Stemming finds the root or base word for each token by removing prefixes and suffixes.

B. *Processing of text:*

In processing of text, score of each feature mentioned in section I are obtained for each sentence of the text. Each feature has a value ranging from 0 to 1. Processing uses hybrid system of Genetic algorithm and Fuzzy System.

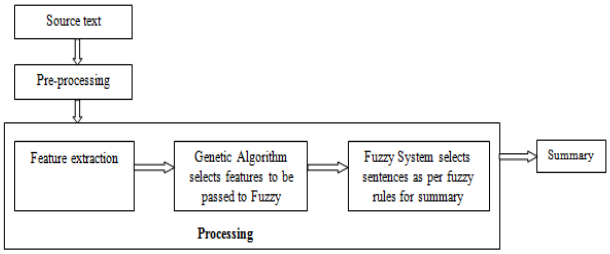

Fig.1.Architecture of Hybrid genetic Fuzzy system for text summarization

Genetic algorithm (GA)**:** Feature score (described in section II) of each sentence are used as an input to Genetic algorithm which optimizes input features. Number of input features can be reduced by using GA. Reduction of input features can be helpful in minimizing processing time and efforts of Fuzzy system. GA mutates and alters the candidate solution and try to provide better solution. Each candidate solution is represented in form of chromosome.

	Title word	Content word	Sentence Length	Sentence position	Proper Noun	Biased Word	Sentence-to-Sentence Cohesion	Thematic word
Feature weight	0.6	0.5	0.6	0.4	0.5	0.8	0.4	0.2
	1	2	3	4	5	6	7	8

Fig.2. sample chromosome for representing sentence as candidate solution

Evolution process starts with initial population. Population generated in each evaluation is called generation. For each evaluation the following steps are performed until some termination criteria τ met. τ can be defined as per domain specific needs. i). crossover and mutation operation is performed on parent chromosomes results in two child chromosomes to be added into population. ii). Based on selection criteria some chromosomes from population are passed to fitness function. iii). Fitness function selects some of the good features to be taken into next generation. Fitness function can be defined as per user specific needs, for example features with values greater than 0.8 are fit and are passed to fuzzy system.

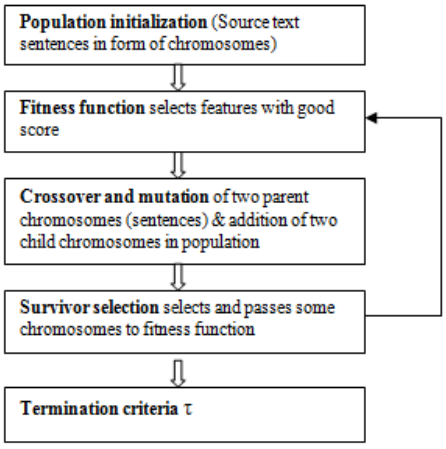

Fig.3. process of Genetic algorithm for selecting feature with good scores

Fuzzy logic system**:** The features along with their scores that are obtained as output of genetic algorithm are passed to fuzzy system. Fuzzy system is made up of fuzzifier, inference engine, Rule base and Defuzzifier[6]. Fuzzifier converts the feature scores into fuzzy values, for this purpose it uses some membership function. Fuzzy sets can be created by considering minimum and maximum values of features. We can use three (HIGH,MEDIUM,LOW) or five(VERY HIGH,HIGH,MEDIUM,LOW,VERY LOW) fuzzy sets. Each feature along with fuzzy set is given as an input to inference engine. Inference engine gives output which is based on rules written in rule base. Rule base consist of set if IF…THEN rules. IF…then rules are written such that they compare scores of features with fuzzy sets. The sample rule looks like "IF (TitleWord is VH) and (Content Word is H) and (Sentence Length is H) and (SentencePosition is H) and (ProperNoun is H) and (BiasedWord is H) and (ThematicWord is VH) and (Sentence-to-Sentence Cohesion is H) " THEN (Sentence is important to be included in summary.) These rules are used by Inference engine to decide whether the statement is important or not. **Dufuzzifier** generates summary after rules of rule base are executed and aggregates the output of each rule to generate summary.

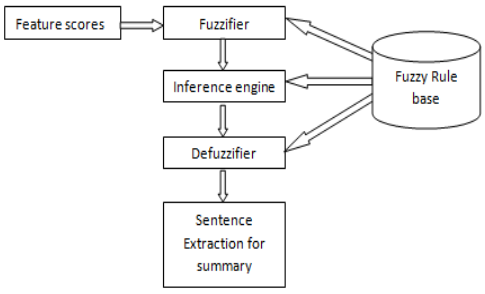

Fig.4. Process of Fuzzy system for selecting Sentences for summary

V. CONCLUSION

In this architecture, use of genetic algorithm is helpful in filtering sentences with good feature score and only selected sentences are passes to fitness function. The list of features given by GA as output, are passed Fuzzy Logic System to create summary.

REFERENCES

[1] Jezek, K., & Steinberger, J. (2008). Automatic text summarization. In*Znalosti* (pp. 1-12).

[2] Saziyabegum, S., & Sajja, P. S. (2016). Literature Review on Extractive Text Summarization Approaches. *International Journal of Computer applications*, *156*(12).

[3] Kyoomarsi, F., Khosravi, H., Eslami, E., Dehkordy, P. K., & Tajoddin, A. (2008, May). Optimizing Text Summarization Based on Fuzzy Logic. In*Computer and Information Science, 2008. ICIS 08. Seventh IEEE/ACIS International Conference on* (pp. 347-352). IEEE.

[4] Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*,*22*, 457-479.

[5] Hahn, U., & Romacker, M. (2001, March). The SYNDIKATE text Knowledge base generator. In *Proceedings of the first international conference on Human language technology research* (pp. 1-6). Association for Computational Linguistics.

[6] Suanmali, L., Salim, N., & Binwahlan, M. S. (2009). Fuzzy logic based method for improving text summarization. *arXiv preprint arXiv:0906.4690*

[7] Chen, F., Han, K., & Chen, G. (2002, October). An approach to sentence-selection-based text summarization. In *TENCON'02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering* (Vol. 1, pp. 489-493). IEEE

[8] Osborne, M. (2002, July). Using maximum entropy for sentence extraction. In *Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4* (pp. 1-8). Association for Computational Linguistics

[9] García-Hernández, R. A., & Ledeneva, Y. (2009, February). Word sequence models for single text summarization. In *Advances in Computer-Human Interactions, 2009. ACHI'09. Second International Conferences on* (pp. 44-48). IEEE.

[10] Alguliev, R., & Aliguliyev, R. (2009). Evolutionary algorithm for extractive text summarization. *Intelligent Information Management*, *1*(02), 128.

[11] Kruengkrai, C., & Jaruskulchai, C. (2003, October). Generic text summarization using local and global properties of sentences. In *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on* (pp. 201-206). IEEE.

[12] Kaikhah, K. (2004). Text summarization using neural networks.

[13] Barzilay, R., & Elhadad, M. Using Lexical Chains for Text Summarization

[14] Suanmali, L., Binwahlan, M. S., & Salim, N. (2009, August). Sentence features fusion for text summarization using fuzzy logic. In *Hybrid Intelligent Systems, 2009. HIS'09. Ninth International Conference on* (Vol. 1, pp. 142-146). IEEE.

[15] Radev, D. R., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., . & Otterbacher, J. (2004, May). MEAD-A Platform for Multidocument Multilingual Text Summarization. In *LREC*.

Authors Profile

Mrs Saiyed Saziyabegum pursed Master of Computer Application from Narmada college of computer application affliated to Veer Narmad South Gujarat University in 2007. She is currently pursuing Ph.D. from G.H.Patel PG Department of computer Science, Sardar Patel University and currently working as Assistant Professor in S.S.Agrawal Institute of Computer Science,Navsari Since 2013. Her main research work focuses on Text Summarization. She has 8 years of teaching experience and 3 years of Induary Experience.

Dr. Priti Srinivas Sajja has been working at the Post Graduate Department of Computer Science, Sardar Patel University, India since 1994 and presently holds the post of Professor. She specializes in Artificial Intelligence and Systems Analysis & Design especially in knowledge-based systems, soft computing and multi-agent systems. She is author of Essence of Systems Analysis and Design (Springer, 2017) published at Singapore and co-author of Intelligent Techniques for Data Science (Springer, 2016); Intelligent Technologies for Web Applications (CRC, 2012) and Knowledge-Based Systems (J&B, 2009) published at Switzerland and USA, and four books published in India. She is supervising work of a few doctoral research scholars while eight candidates have completed their Ph.D research under her guidance. She has served as Principal Investigator of a major research project funded by the University Grants Commission,India.She has produced 184 publications in books, book chapters, journals, and in the proceedings of national and international conferences out of which five publications have won best research paper awards.

