A Novel Approach to Summarization based on Centroid Fuzzy

Vetriselvi T, Gopalan N P

Abstract*: Text summarization is a way to create a description of a given document. This is a novel approach to text summarization, which is a combination of fuzzy and centroid methods. In fuzzy method the result is based on the input given to the membership function. In centroid, missing some relevant words may make the summary irrelevant, and then the summary content is not meaningful. Our model overcomes the above two problems by combining the results of those approaches. The rate of summarization determines the size of the summary. Centroid is a group of words which are highly relevant to the document. Fuzzy membership functions helps to categorize the most relevant sentences. Both approaches have their own disadvantages, so we pick the best of the above two and create a novel approach as fuzzy centroids text summarization: this approach performs well in multi document summary when compare with existing

Index Terms: centroid , Frequency, Fuzzy, TF-IDF.

I. INTRODUCTION

Summarization is a process of creating a miniature of the given document or set of documents, where the information or semantic of the original documents should be the same. There are two methods of summarization: one is extractive and the other is abstractive. Summary helps the people to do quick decision making, it reduces human reading time. And multi document summary gives the short overview of all documents at a glance. The general summarization is a way to reduce the size of the content of the original document without destroying it's meaning .Lexicon- based methods, machine-learning methods, and semantic relatedness methods are different approaches to documents. TF-IDF is one of the famous methods to count the terms or keywords in a document, i.e. it shows the importance of the key terms. And it is applied to a single document and multi document, which is under summarization process. Textual Entailment –One sentence is a subset of another one sentence called subsume .In some documents one sentence which contains the meaning of another sentence is also called subsume[1]. In extractive approach the sentences are ranked using the formula (1); the summation of each term weight of a sentence will be the score of the sentence.

$$\text{Score(s)}=\sum_i^n tf_i /n \qquad (1)$$

tf_i = frequency of word i, i.e, number of times that i appears in the source document,
n = length of the sentence without considering stop words.

Revised Manuscript Received on December 22, 2018.

Vetriselvi T, Department of Computer Science and Engineering, K. Ramakrishnan College of Technology, Tiruchirappalli, India
Gopalan N P, Department of Computer Applications, National Institute of Technology, Tiruchirappalli, India

II. LITERATURE SURVEY

A text summarization approach under the influence of textual entailment is a two-step process, such as sentence score and sentence entailment. Word frequency plays a major role in sentence scoring shown in formula(1) .Each and every sentence contains a certain set of words(t1,t2,t3,t4), the sum of weight of all words and Total number of words in that sentence are used to calculate the weight of the sentence or score of the sentence in a single document[1].

Sen1=t1,t2,t3,t4,t5,t6 (2)

S1: Tropical(2) Storm(6) Gilbert(7) formed(1) in(0) the(0) eastern(1) Caribbean(1) and(0) strengthened(1) into(0) a(0) hurricane(7) Saturday(4) night(2):
S2 : There(0) were(0) no(0) reports(1) of(0) casualties(1):

The second step is a textual entailment. It shows how one sentence is closely related to another. The task is to identify the duplicate sentences and remove them. Sentences with same meaning and the sentences having the same words are considered as duplicates. From two entailments remove the small sentence to retain the semantic of the document.it is a way to reduce the size of the document, The duplicate sentences are removed. Recognising Textual Entailment (RTE) helps to identify the TRUE or FALSE relationship between any two given sentences [1]. In the above box there are two sentences pair (s1,s2):TRUE .RTE clearly shows the relationship between the sentences. Sentences are mainly considered for summarization, because it contain much information.[14]

Centroid-based summarization is a mathematical approach, A centroid is the asset of high score keywords as a whole. The high frequency terms are identified and they form a group called centroid. If the terms available in the sentence are used many times throughout the document, then the Cluster Based Sentence Utility(CBSU) will be high. For a specific term the centroid value is calculated by CF*IDF, The terms which have higher centroids are considered for centroid cluster, only if the centroid value is more than threshold value. In Radev's calculate centroid score for the sentence by formula (3).

$$\text{Score(S)}=\sum_{i=1}^{n}(w_c C_i + w_p p_i + w_j F_i) \qquad (3)$$

w_c, w_p, w_j were predefined constant values

c_i C centroid value for the sentence i

p_i Position value for the sentence i

F_i,it is an IDF value, it was used two centroid values namely Pure centroid and Lead centroid. The redundant sentences had high centroid value but it's a fake centroid value, to avoid that redundancy penalty is subtracted from the sentence score, and the formula (3) is rewritten as

Score(S) = $\sum_{i=1}^{n}(w_c C_i + w_p p_i + w_j F_i) - w_R R_S$ (4)

R_s=2*(# of Overlapping words /# of words in S1+ # of words in s2) (5)

The value of redundant score is 1 for similar sentences and 0 for ideal sentences [2]. The above 0.7 similarity score sentences are removed. Radev evaluates this system by newswire dataset; nearly 558 sentences are processed by inter judge agreement and are identified 54 cases as sub-sumption. Further it was proceeded with the number of experiments on the same set of data using MEAD, which is a centroid-based summarizer[3]. Performance of the single document and multi documents are analysed. In single document evaluation MEAD evaluates grammaticality, cohesion and peer organization documents centrality. But there is a vast difference between the evaluation result of the system and the judges result. Just 88 sentences were interpreted by pair judge agreement, but the system produces 4291 sentences. Key phrases instead of keywords for single document summarization is an another approach ,here key phrases and their sentence position play a vital role for summarization .As a first step the important key phrases are extracted and the key phrases whose length exceeds 5 are discarded. The formula (6)

key phrase frequency=PF*IDF (6)

$Score_{pos}=\frac{1}{\sqrt{i}}$ (7)

i=position of the sentence $Score(S)=Score_{pos}+(\frac{score_{PFIDF}}{\max(score_{pfidf})})$ (8)

Kamal maintains two arrays, one is to store the key phrases(KP[5]),and the another one is store the sentences(summary[6][7]).The newly identified key phrase is added to the KP array if it doesn't have any match in KP. Similarly a new sentence's are added to the summary if it doesn't have any match in summary sentences. DUC 2001, DUC 2002 datasets were evaluated using Rouge(Rouge-1.Rouge-2,Rouge-n). Fuzzy logic is smart enough to provide boolean results over any kind of data by fuzzy member function, in weather forecasting, financial application, reviews, recommendation systems. Here in text mining fuzzy plays a major role while grouping the sentences as well as grouping the words.

A single membership function of fuzzy is not enough to work on documents; so parameter based member functions were derived as fuzzifier. Familiar parameters regarding sentences were Location (l), summary type(s), and Word Net(w) measures .Each parameter gave three measures Low, High and Medium as result. The Measures were again evaluated by inference engine using Fuzzy Decision Rule, which return either one or zero, The return of value one showed that the sentence is important.

Kyoomarsi 's system of sentence extraction provides summarization by the combination of key phrase and sentence ranking[6]. Kyroomarsi selected 12 sentences and they were ranked by different ranking methods .The Error Average is the difference between rank of fuzzy based summarization(R_a) and the sum of all other ranks((R_b + R_c + R_d + R_e)/4).Hence the formula (7) depicts the error average calculation

$$E= \sum_{i=1}^{12} R_{a(i)} - \frac{R_{b(i)}+R_{c(i)}+R_{d(i)}+R_{e(i)}}{4} \quad (9)$$

$$E_{av} = \frac{E}{12} \quad (10)$$

The sample is enhanced to large dataset, and the error rate has been reduced much in the final summary. The formulas (8),(9) and (10) applied to calculate the error rate. But the relatedness between the sentences are not considered, so same meaning sentence may repeated. Wikipedia-based semantic relatedness shows the way to overcome the semantic problem. Wikipedia dataset is open source taken for evaluation. This system of explicit semantic relatedness among the sentences is a three step process [9]. The first step is key word extraction from Wiki, secondly creates a weighted vector, the last step is to finds the relationship between the pair of every two terms. The more similar terms have been removed with the rest of the sentences a summary formed. Fuzzy membership function helps to categorize the sentences [4][6][9].This system[4],considers eight parameters to compute fuzzy value about the sentences. The parameters are sentence weight, nouns in the sentences, sentence position and sentence length. Fuzzy categorize the sentences into two groups--, relevant and irrelevant .The relevant sentences are arranged in their rank lower to higher and related to score higher to lower.

Latent Semantic Analysis is a technique to analyse the semantic relatedness among sentences [4].This process starts with key term extraction with their weight, sentence extraction with the weight by the sum of the weight of each term in a sentence. Terms and sentences together form a matrix. The cell of the matrix is filled by word count of that sentence. Singular Value Decomposition(SVD) forms the sentence concept matrix. The highest value sentence in that matrix is considered as an important sentence.

Barbar combined the fuzzy and LSA to improve the summarization. This hybrid approach combined there two outcome summaries to produce a single summary; meanwhile it removes the redundant sentences. Genetic Algorithm is famous for providing optimum solution.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

This famous soft computing algorithm is used in various research areas such as network security algorithms, natural language processing systems, information retrieval and recommendation systems etc. This system [10] tried three-step processes for summarization. The first step was creating member function for producing crisp output. Secondly it is followed by fuzzy inference engine and thirdly by de-fuzzyfier. It's a kind of extractive summarization, extraction happens by removing the unwanted sentences. Ladda adds semantic relatedness with fuzzy genetic approach. This semantic relatedness has been identified by Semantic Role Labelling (SRL). Ladda provides a hybrid approach with a combination of fuzzy, genetic and SRL; still this system is not efficient, because priority of keywords based on sentence length have to be considered. A word embedding quality of summarization was measured by word movers distance is a key word based method for single (and multi document) summary [7]. The key concepts in the document helped to do the single document summary without any human intervention [5].Graph based algorithms also widely used for summarization [11]. Major issue with that model is it consume additional time to construct graph. Different document summarization techniques are explored[13].Fuzzy along with rough est theory in extractive summary produce good result in this context[12].chapter 3 contains proposed approach. Chapter four is an experimental setup. Conclusion is an last and five section of this article.

III. PROPOSED METHOD

The proposed model produces a competitive summarization. It's a positive combination of best features of fuzzy model and Centroid -based re-ranking approaches. It is implemented as two parallel methods for summary. Sentences' scores are calculated by fuzzy member function with decision rule from the result of the member function. Decision rule has been framed in the combination of various sentence parameters such as sentence Length (F1), sentence Score(F2),sentence Position(F3),sentence similarity (F4). The value for F1 was High, medium, low, for computation it is set as 1,0,-1.
if the sentence length is from 3 to 5

then the F1 score =-1

if the sentence length is from 6 to 9,

then the F1 score =0.

if the sentence length is above 9,

then the F1 score =1.

Sentence weight is the sum of the term weight by TF*IDF, sentence score has to be normalised to 10 ,hence all scores fall between 1-10

If sentence score <=5 then F2 is low

if 5<sentence score >=10 then F2 is High. The value of F3 is for sentence position where the sentence is exactly available in the original document. Sentence similarity (F4) value is calculated by identifying similar words in both the sentences and the maximum similarity will be considered as the same sentences. Sentence matrix helps to calculate the similarity value between sentences.

All diagonal values of the sentence matrix are replaced with 0.The sentences which have high similar values are considered as redundant sentences, and the least similar value sentences are consider for summary. Decision rule was framed from the result of four parameters. Thus the following rule is applied to every sentence and valuates their importance of it.

if(F1=High and F2=High

and F3=High and F4=Low)

then that sentence is most relevant

if(F1=High and F2=Medium

and F3=High and F4=Low)

then that sentence is irrelevant

Important sentences are arranged in descending order of weight, Summarization rate determines the number of sentences selected as summary. In our parallel approach fuzzy –rule- based summarization produces one set of summary, and the other is Centroid based summary. The centroid identification is the first step in this method.

Term Frequency Inverse Document Frequency is a way to identify the important score of every term in a document. The descending order of term score helps to rank the terms. Cosine similarity between terms is useful to identify the threshold.

Certain terms are selected as a centroid with the help of threshold value if the terms of the sentence are in the centroid then that sentence is useful for summary. To know the sentence centroid adds all the centroids of the terms of that specific sentence. Shows in Fig 1.

Position of the sentence and frequency of the terms of the sentences are taken into account for calculating the final score of the sentence.

$$(S_i) = \sum_{i=1}^{n} (w_c C_i + w_p p_i + w_j F_i) \quad (3)$$

By changing the value of W's, the score can be calculated. As in formula (3) the sentence score is calculated.

Formula (11) is used to calculate the redundancy penalty based on the common terms among the sentences. Similarity between the sentences is identified by using common sentence information utility.

Here utility between two sentences of two different documents are captured in the same equivalence class. A common utility graph that shows the connectivity among sentences where ever commonality occurs.

Re-ranking Redundancy Penalty for Sentences (RPS) is a way to reduce the score of the similar sentences. Redundancy Penalty is calculated by the formula given below, for every pair of sentences RPS is calculated as follows,

$$RPS=2*C/TT \qquad (11)$$

$C_{si,si+1}$=common terms in both the sentences

$TT_{Si,Si+1}$=Number of terms in Si+ Number of terms in S_{i+1}

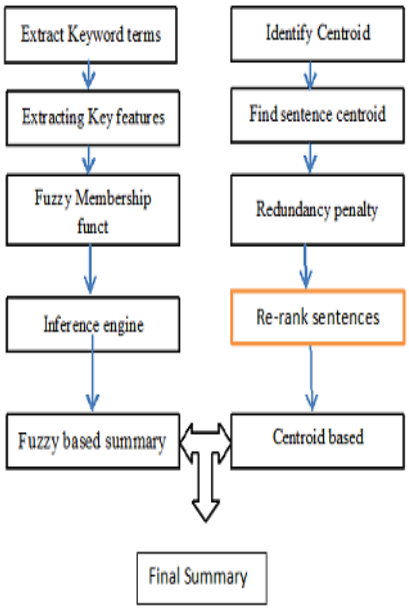

Fig1 Fuzzy based text Summary process

In centroid approach after calculating the redundancy penalty, it is applied to remove certain repeated sentences from a given document is identified. repeated sentences are numbered in ascending order as per they are in the original document. The first occurrence alone is considered when two or more sentences are same meaning and rest of them are removed. It gives the small number of sentences. Again the sentences are ranked by their actual score.

After removing the redundancy from the existing score, the score is calculated again for the entire set of sentences. By re-ranking the top ranked scored sentences are arranged in the descending order and set of statements are chosen based on summarization rate. Fig 1.shows our novel approach and it combines both summaries and takes union among them to get the final summary. The joining process of two different summaries also has some conflict i.e. difference. That are captured by function Conflict (s1, s2).It is a sentence wise comparison which helps to identified the dissimilar sentences. Similar sentences alone consider for final summary.

Conflict(s1,s2) is one when both the sentences are same otherwise zero.

IV. EXPERIMENTAL RESULTS

Table 1: Term Frequency form pre-processed

S.No	Term	Frequency
1	hurricane	23
2	Heavy	8
3	wind	16
4	Rain	9
5	Weather	8
6	Warning	5
7	storm	16

The Wikipedia dataset and newswire dataset are mostly use familiar datasets. Document Understanding Conference Series data set is having gold standards for summarization. For evaluation , Recall Oriented Understudy for Gisting Evaluation (ROUGE), helps us to evaluate the system generated summary with user generated summary. We take dataset from Duc-2002 and sample shown in paragraph 1 and 2. Paragraph 1-“south coast to prepare for high winds, heavy rains and high sea broad area of cloudiness and heavy weather” rotating around the center of the storm”.
Paragraph-2” Heavy rain and stiff winds downed power lines and caused flooding in the Dominican Republic on Sunday night as the hurricane’s center passed just south of the Barahona peninsula, then less than 100 miles from neighbouring Haiti.” As a first process of summarization, we remove the stop words such as “the”, ”and”, ”of”, ”for,” ”in”, ”on”, ”as”, from the above paragraphs. The second step of pre-processing is stemming i.e removing the suffix’s of the terms such as,”ily”,”ed”,”ing”.Weight of the frequent terms are identified by TF*IDF. Centroid formed with high value terms. Sentence score calculated from fuzzy method is used to produce fuzzy based summarization and centroid based sentence score is used to produce another set of summarization. Both of them are put together to get a new form of summary. The key terms and their weights are shown in the Table 1. This system produces a summary that is evaluated by Rouge Tool analyse and the performance measures such as Recall, Precision, F-Measure values are computed. Rouge has a Set of Tools namely Rouge-1(Unigram),Rouge-2(bigram),Rouge-L(Longest common sequence).Recall and Precision values given in Table-2 ,It shows that this system performs better than the existing systems .Especially in Skip-bigram plus unigram-based co-occurrence statistics (ROUGE-SU)our approach produces better performance.

V. CONCLUSION

In this paper, we present a novel approach to text summarization, which is a combination of fuzzy and centroid methods. In sentence categorization fuzzy works better. In this centroids-based model depends on term count and sentence score. Duplication of sentences are removed by redundancy measure. The combination of fuzzy and centroid performs well when compared with excising text summarization approaches.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

TABLE 2. COMPARISON WITH EXSISTING METHODS

		ROUGE1	ROUGE2	ROUGE-L	ROUGE-SU4
Word Frequency	Recall	42.483	17.912	38.247	20.014
	Precision	40.567	17.024	36.529	19.035
	F-Measure	41.502	17.442	37.337	19.495
BASELINE	Recall	43.741	17.522	39.575	20.195
Word	Precision	43.398	17.362	39.248	20.016
Pre-processed texts	F-Measure	43.538	17.435	39.388	20.094
Textual Entailment	Recall	45.428	20	41.264	21.779
	Precision	45.004	19.234	40.86	21.553
	F-Measure	45.184	19.421	41.038	21.654
FuzzyBased Centroid	Recall	46.783	20.531	43.62	23.735
	Precision	45.904	19.435	43.654	22.643
	F-Measure	46.339	19.967	43.635	23.704

REFERENCES

1. Elena Lloret,Oscar Ferrandez,Rafael Munaz,and Manuel Palomar,"A Summarization Approach under the Influence of textual Entailment"Published in MLPCS,PP 22-31,2008
2. Dragomir R. Radev , Hongyan Jing,and Malgorzata Budzikowska "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies" Proceedings of the 2000 NAACL-anlpworkshop on Automatic summarization - Vol 4,PP 21-30,2000
3. Dragomir R. Radev "Experiments in Single and Multi-Document Summarization Using MEAD" In First Document Understanding Conference, Ann Arbor,2001
4. Mr.S.A.Babar and Prof.S.A.Thorat Improving "Text Summarization using Fuzzy Logic & Latent Semantic Analysis" International Journal of Innovative Research in Advanced Engineering (IJIRAE) ,Vo l 1 Issue 4 ,PP 171-179 , 2014
5. kamal Sarkar"Automatic Single Document Text Summarization Using Key concepts in Documents",Journal of Internation Process Systems,Vol 9,Issue 4,PP 602-620,December 2013
6. f. Kyoomarsi, h. Khosravi, e. Eslami and m. Davoudi "Text extraction-based summarization using fuzzy analysis" Iranian Journal of Fuzzy Systems Vol. 7, Issue. 3, pp. 15-3, 2010
7. Liqun Shao, Hao Zhang, Ming Jia, and Jie Wang "Efficient and Effective Single-Document Summarizations and a word-Embedding Measurement of Quality:" sumarxiv:1710.00284 vol 1 ,PP 1-10 ,1 Oct 2017.
8. Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). "Efficient estimation of word representations in vector Space". Arxiv preprint arxiv:1301.3781 ,Vol. 03,PP 1-12,2013
9. Evgeniy Gabrilovich and Shaul Markovitch "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis" Proceedings of the Twentieth International Joint Conference on Artificial Intelligence PP 1606-1611,2017
10. Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan, "Fuzzy Genetic Semantic Based Text Summarization" Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing,PP 1185-1195,2011.
11. Gune¸s Erkan, Dragomir R. Radev"LexRank: Graph-based Lexical Centrality as Salience in Text Summarization" Journal of Artificial Intelligence Research 22 ,PP 457-479,2004
12. Hsun-Hui Huang ,Yau-Hwang Kuo ,Horng-Chang Yang ,"Fuzzy-Rough Set Aided Sentence Extraction Summarization",Proceedings of the FirstInternational Conference on Innovative Computing, Information and Control (ICICIC'06),0-7695- 2616.
13. Archana AB, Sunitha. C ,"An Overview on Document Summarization Techniques" ,International Journal on Advanced Computer Theory and Engineering(IJACTE) ,ISSN (Print) : 2319 ″U 2526, Volume-1, Issue-2, 2013 .
14. Rafael Ferreira ,Luciano de Souza Cabral ,Rafael Dueire Lins ,Gabriel Pereira e Silva ,Fred Freitas ,George D.C. Cavalcanti ,Luciano Favaro , "Assessingsentence scoring techniques for extractive text summarization ",Expert Systems with Applications 40 (2013) 5755-5764 ,2013 Elsevier .