International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

Extractive Summarization using EMDS

Mr. Amit. S. Zore, Dept Of Computer,
G.H.Raisoni College Of Engg
University Of Pune,India.

Prof. Aarati Deshpande, Dept Of Computer,
G.H.Raisoni College Of Engg,
University Of Pune,India.

***Abstract*— Summarization is a process of including important stuff only from the original document. Summarization can be used understand overall highlight from huge amount of data. It can be either Generic Summarization or Query Based Summarization. Generic Summarization selects important sentences from source document and puts them inside a summary and in Query Based summarization sentences are extracted base on user query. Summarization can be single document summarization or multiple document summarizations. In this paper, new multi document summarization technique has been proposed called as Extractive Summarizer using GMDS. Here, input document is parsed to extract all the sentences, classes are attached to each sentence, sentences are given rank according to their classes and top ranged sentences are included into a summary. Finally, proposed system will be compared with existing multi document summarization techniques such as RANDOM, LEAD, MEAD technique to show how it is efficient than existing ones.**

***Index Terms*— Summarization, MEAD, Extractive Multi Document Summarization, Query Based Summarizer.**

I. INTRODUCTION

When we search "techniques of summarization" on Google we get results around 2460000 in 0.40 seconds till 2 January 2014.this retrieval is very vast so we need second level of abstraction to reduce this volume of data we required: i.e. summarization. Now a day's Summary is an important part of our day today life. We don't have so much time read whole document or data. Summarization is a process of including only important stuff. From the childhood in school we have question like write a summary from above paragraph. That time we don't know exact meaning of summary. From the base of that technique (knowledge) we implement different type of summarization. Summarization is used to reduce large amount of data. Summarization is a method to retrieve information from source document in form of abstraction.

With help of summarization we visualize large amount of data in short duration. Now a day's internet is very popular for searching data we will receive lot of information from internet but with all these information available, we don't have time to read everything

For example if we want search some content on internet it will give us lots of information with help of summarization we can easily find out meaning of things. Summarizations are useful in many applications. Best example of summarization is news bulletin with help this we easily understand the overall scenario of the day. Also in our reference book, text book at the end of lesson summary is there with help this we easily understand overview of lesson.

Document summarizations [3] having two different dimensions either abstract based or extract based. Extract based summarization depend on sentence extracted from document and abstract summarization may use word and phrase that not appear in original document. In extractive summarization we choose significant paragraph, sentence from source document concatenating them into abstractive form. This is based on linguistic feature and statistical of sentence. Where an abstractive summarization understands original texts and telling it in own word or with new idea.

Text summarization techniques also divided on the basic of volume of document in database, if summarization performed on single document then it's called as single document summarization. If summarization performed on a set of multiple documents then it's called as multiple documents. Single summarizations are useful for cell phone. Where the multi document summarization useful when the data is vast. This paper focus on extractive multi summarization

The rest of paper is organized as follows. Section II consists of Literature Survey and Discussed Present Techniques of Summarization. Section III consists of Proposed System as Extractive Summarization using EMDS. In section IV comparative study between existing systems and proposed system is discussed. Section V consists of Result Evaluation. Section VI contains Conclusion and Future Scope and finally References.

II. LITERATURE SURVEY

A. *Single & multi document summarization technique:*

Single document summarization called when summarization are performed on singe document where multiple summarization called when summarization are performed on set .

B. Present technique of multi summarization:

There are number of multi summarization technique [1] available some are graph based where other are without graph based technique. Following are some generic multi summarization technique.

i. RANDOM based method:

Random based technique [8] is easiest technique in this we select randomly any line from document depending upon the compression rate means summary size. In this random technique randomly we assign value 0 or 1.we provides threshold value for length of sentence is provided. Score of 0 to 1 is given to all sentence that not meet length cutoff. Finally we decided required sentence according highest score for desired summary

ii. Lead based Method:

In this technique [2] we select first or last sentence of paragraph based on comparison rate (CR).this technique is very good for news article as they have main subject which is set in first line of article. So it can be feasible that n% sentence is chosen from beginning of the test e. g .if we select fist sentence of each document, then second sentence of each document till required summary is constructed. This method is called lead based method. In this we assign score 1/n to each sentence, where n assign as sentence number in related document.hat means the first sentence in related document have same scores etc. also provide a threshold value for sentence length.

iii. Mead based Method:

This technique [2] is centroid-based extractive summarization. In this we score sentence which is depends on sentence-level and inter-sentence features they indicate quality of sentence as summary. Afterword chosen top-ranked sentences are included in output summarization. Mead based technique is score sentence depending on certain sentences feature.1.centroid 2.position.3.lengh.this technique use following formula to calculate score of sentence

$$\text{Score (Si)} = \begin{cases} \sum (Wc * Ci + Wp * Pi) & \text{If Length (Si)>Threshold} \\ 0 & \text{If Length (Si) < Threshold} \end{cases}$$

Here,
Wc = The weight for the Centroid feature.
Wp = The weight for the Position feature.
Ci = The calculated Centroid value for ith sentence.
Pi = The calculated Position value for ith sentence.
Si = The i^{th} sentence of the document.
i = Sentence number within the cluster in i<n.

III. EXTRACTIVE SUMMARIZER USING EMDS

EMDS i.e. Extractive Multi document Summarization is Graph based approach for Extractive multi document summarization technique.

The algorithm consists of steps that are mentioned in Figure 1. Input to model set of multiple relevant documents or single document. After taking multiple documents, all documents are assembled together. Then assembled document is passed to summarization model. In next step, sentences are extracted from input document. After parsing each sentence, stop words are removed and stemming is applied to convert each word to their root form. Then we assign class to each sentence. We create undirected graph for each document with sentence denoted as node, similarities denoted as edge..Then sentence are ranked with respect to their score. Topmost ranking sentence are chosen to form the summary for each document. For filter out redundant information we use semantic checking. With help above procedure we form the final Extractive summary.

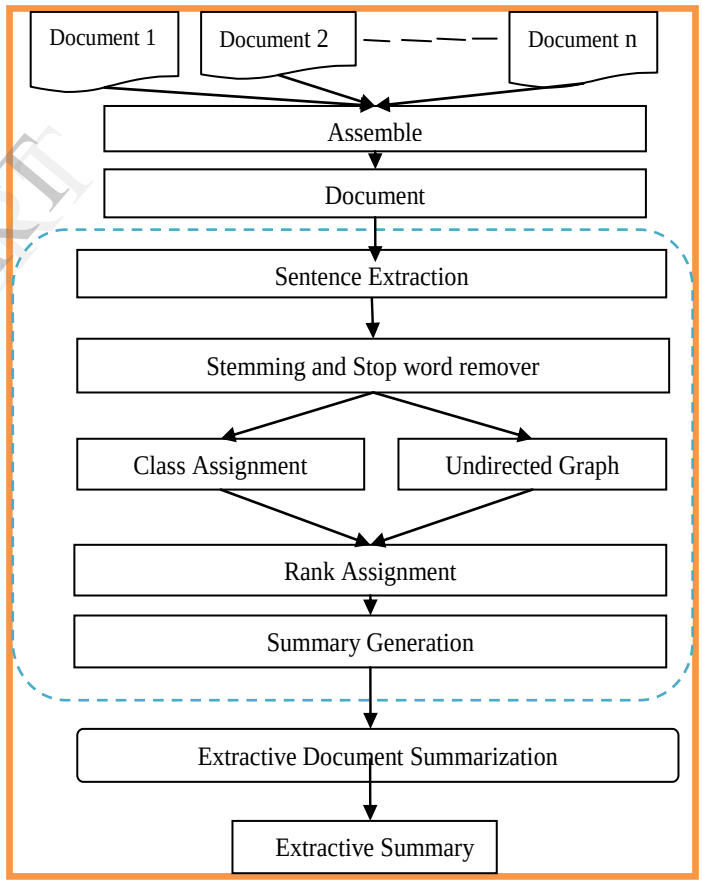

FIG 1. EXTRACTIVE MULTIDOCUMENT SUMMARIZATION

1. Document Assembler:

In this module, all sentences from each document are extracted and assembled into one document. Assembled document is passed to sentence extraction module as a input.

2. .Sentence extraction:

In sentence extraction module, input document is parsed to extract each sentence. After reading each sentence result is passed to next module as an input.

3. Stemming and stop word remover:

Stemming is used for reducing inflected word or derived word from their base term or root. Most time document there different forms of word are present like plural vs. single, past vs. present, noun, phrase these entire world have same meaning unnecessary treat differently. Following are examples of stemmer for English. The word fishing, fisher, fished stemming algorithm reduces unnecessary meaning "fish".

Stop word which can thrown out from sentence prior, its not control or automated by human input. There is no standard list available which used by all tools. There are number word in document that do not contain information but its important for syntactical formation such as "the" "is" do" etc. since these word are not useful or less important they create noise in document .stop word removal with help predefined human. Example of stop word are as follows back, all, the, were, done, do, down, does, about, after.

Before assigning class to a sentence, first removing stemming means word are converted back to their original root and removed the different stop word.

4. Assigning Class:

After removing stop word and process of stemming, classes are assigned to each sentence. Absolute class, Summed class and sentence length is calculated in this module. With help these three parameters ranks are assigned to each sentence.

5. Undirected Graph:

In this module, undirected graph G = (V*E) is constructed in which each sentence in a document is a node and edge of graph denoted as similarity between sentence (nodes).

6. Sentence Rank:

In this module, rank is assigned to each sentence according to classes, sentence similarity and sentence length. The sentence which has highest absolute class will be given highest rank, then sentence length is given priority to give next rank and finally most similar sentence is given next highest rank.

7. Summary Generation:

In this module final extractive summary will be generated through selecting sentences which has highest rank. Also sentenced are ranked according to summed class for appropriate order. Simply sentence with high rank will be selected to include in final summary

IV. COMPARATIVE STUDY:

Here, comparison between existing systems and proposed system is presented. Following are some of the existing summarization techniques.

1. RANDOM
2. LEAD
3. MEAD

RANDOM based in which we choose or select sentence randomly from the source document. Here on the basis of length we decide which should include in summarization.

LEAD based in which we choose or select first and last sentence from paragraph based upon the compression rate (CR) It's useful for news article.

MEAD is a centroid-based Extractive Multi summarization in which highest value of the scored sentences taken in the extract sentence. Here score is decided according to certain features-Centroid, Position and Length. Table 1 shows comparison between systems.

Multi document Summarization Technique/ parameter	Absolute class	Summed class	Sentence length	Similarity sentence
RANDOM BASED	*NO*	*NO*	*YES*	*NO*
LEAD BASED	*NO*	*NO*	*YES*	*NO*
MEAD BASED	*YES*	*NO*	*YES*	*YES*
EXTRACTIVE MULTI GRAPH ALGORITHIM	*YES*	*YES*	*YES*	*YES*

Table 1.Comparison Between Existing System And Proposed

V. RESULT EVALUATION:

Proposed system is under development. So, after final system development it's results will be compared with existing system. ROUGE and Manual evaluation will be done to test results.

VI.CONCLUSION AND FUTURE SCOPE

Summarization is a process of understanding any document in short time. In this paper, new technique for multi document has proposed. In Extractive summarization using EMDS, extractive summary of multiple relevant documents is produced using various sentence features such as word class, sentence length and sentence similarity. In this paper, comparative study between proposed system and existing system is studied. Finally, results of proposed system will be compared with existing systems.

REFERENCES

[1] Jade Goldstein*" Multi-Document Summarization By Sentence Extraction".
[2] *Ercan Canhasi, Igor Kononenko* "Semantic Role Frames Graph-Based Multidocument Summarization" University Of Ljubljana, Faculty Of Computer And Information Science.
[3] Mohsin Ali, Monotosh Kumar Ghosh, "Multi-Document Text Summarization: Simwithfirst Based Features And Sentence Co-Selection Based Evaluation ", Department Of Computer Science And Engineering, Khulna University, Bangladesh 2012
[4] Vishal Gupta, Gurpreet Singh Lehal, "A Survey Of Text Summarization Extractive Techniques", Journal Of Emerging Technologies In Web Intelligence, Vol. 2, No.3, Pp. 258 -268, August (2010).
[5] Rafeeq Al-Hashemi, "Text Summarization Extraction System (Tses) Using Extracted Keywords", International Arab Journal Of E-Technology, Vol. 1, No. 4, June, Pp. 164- 168, (Mobicom), Pp. 255-265, 2000.
[6] Document Summarization Using Multi-Features Combination Method Uacee International Journal Of Computer Science And Its Applications – Volume 2: Issue 2 [Issn 2250 - 3765].
[7] Daan Van Britsom Department Of Telecommunications And Information Processing Ghent Universityghent Automatically Generatingn Multi-Document Summarizations 2011 11th International Conference On Intelligent Systems Design And Applications.
[8] Jayabharathy1, Kanmani2 And Buvana3 "An Analytical Framework For Multi-Document Summarization " IJCSI International Journal Of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011.
[9] Frequent Term And Semantic Similarity Based Single Document Text Summarization Algorithm *International Journal Of Computer Applications (0975 – 8887) Volume 17– No.2, March 2011.*
[10] Xiaoyan Cai And Wenjie Li " Ranking Through Clustering: An Integrated Approach To Multi-Document Summarization" Ieee Transactions On Audio, Speech, And Language Processing, Vol. 21, No. 7, July 2013
[11] Dragomir R. Radev "Centroid-Based Summarization Of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, And User Studies
[12] Satoshi Sekine "A Survey For Multi-Document Summarization" New York University.
[13] Shanmugasundaram Hariharan1," Enhanced Graph Based Approach For Multidocument Summarization "*The International Arab Journal Of Information Technology, Vol. 10, No. 4, July 2013*
[14] Rada Mihalcea And Paul Tarau" A Language Independent Algorithm For Single And Multiple Document Summarization" Department Of Computer Science And Engineering University Of North Texas.
[15] Java, The Programming Language – Http://Www.Oracle.Com/Technetwork/Java/Index.Html