ISSN:0975-9646

Extractive Multi Document Summarizer Alogorithim

Amit.S.Zore[1], Aarati Deshpande [2],

[1] Research Scholar,Department of Computer ,Pune University,
Marathwada Mitra Mandal Institute Of Tech ,Pune

2.Assisatance Professor,Department of Computer ,Pune University,
G.H.Raisoni College Of Engg & Management ,Pune .

Abstract— Extraction Text summarization systems are among the most attractive research areas nowadays. Summarization systems offers the possibility of finding the main points of texts and so the user will spend less time on reading the whole document. Different types of summary might be useful in various applications and summarization systems can be categorized based on these types. This paper presents existing summarization techniques, we compare emds algorithm result with existing technique .this result compare with help of rough algorithm, in this for measurement we use three parameters .ie recall ,precision & f-Score ,with we show how our system is better than existing systems

Keywords— Keywords: Summarization, Multi Document Summarization, MEAD, Query based

.

I. INTRODUCTION

The world of texts a vast and widespred world. Most of data networks, like internet, the great deals of important information were still in the text format. Nowadays, with the massive increase in the text information that we receive every day, text summarization system are helpful in finding the most important contents of the text in a short time. Text summarization system can be used in different situations; for instance, as a summarizer in a search engine to give a summarized information of each page to a user [1], and for summarizing the letters and other document in offices. Moreover, newsgroups are using multi-documents summarization system to use the most important information of documents which are discussing one topic. Summarization systems were also popular in areas that we want to decrease the amount of transferred information. For example, users who check their emails by cell phones prefer to less transferred data while connecting to the internet.

Achieving this goal, web sites may use these systems to decrease the amount of transferred data which Results in to access to the information more quickly

Algorithms for extractive summarization are typically based on techniques for sentence extraction, and attempt to identify the set of sentences that are most important for the overall understanding of a given document. Some of the most successful approaches consist of supervised algorithms that attempt to learn what makes a good summary by training on collections of summaries built for a relatively large number of training documents

The rest of the paper is organized as follows: Section II consist of Literature survey. It describes existing multi document summarization techniques. Section III describes the graph theory. IV.EMDS system Section V result evaluation VI describes conclusion and section VII consist of references

.

II.LITRATURE SURVEY

Text summarization or rather automatic text summarization Is the process in which a computer creates a shorter version of the original text (or a collection of texts) still preserving most of the information present in the original text? This process can be used compression and it necessarily suffers from information loss. Thus an ETS system must identify important parts and preserve them. What is important can Depend upon the user needs or the purpose of the summary.

1.2.1 Application of Text Summarization

Text Summarization is helpful for save time. Text Summarization can speed up other information retrieval and text mining processes.

Text Summarization can also be helpful for text display on Hand-held devices, such as PDA. For instance a Summarized version of an email can be sent to a hand-held device instead of a full email.

1.2.2 Classification of Text Summarization Techniques

Text Summarization is condensing the source text into a shorter version preserving its information content and overall meaning. The text summarization techniques can be classified by using the way by which the summarization is going to be performed over the text data. Following are the two broad level

classifications of text summarization techniques.

1.2.2.1 Extractive and Abstractive Text Summarization

Text Summarization methods are classified into extractive and abstractive summarization. An extractive summarization method is selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences decided based on statistical and linguistic features of sentences. An abstractive summarization method depend on understanding the original text and re-telling it in fewer words. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text conveys the most important information from the original text document. [2]

1.2.2.2 Single Document and Multi Document Text Summarization

Text summarization techniques can also be classified on the Based on volume of text documents available in the text database.

If summarization performed for a single text document then it is called as the single document text summarization. If the summary to be created for multiple text documents then it is called as the multi document text summarization technique

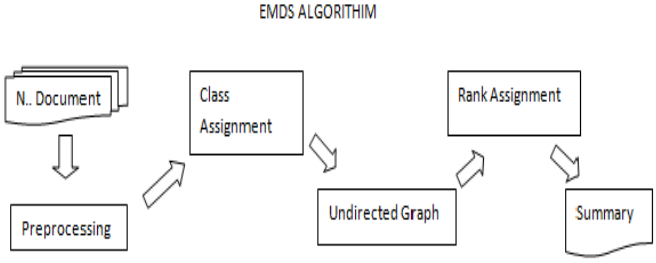
III. GRAPH THEORY

Graph used for representing the structure of text as well as the relationship between sentences of the document. Sentences in documents are presented as nodes. The Edges between nodes illustrates connections between sentences. These connections are introduced by similarity relation and this relation is measured as a function of likelihood between contents. By deploying different similarity criteria, the similarities between two sentences are calculated and each sentence is scored. All the scores for one sentence are combined to form a final score for each sentence. When the graph processed, the sentence will be categorized by their scores and sentences in higher orders are chosen for final summary [4].

In the authors introduced a stochastic graph based method for computing relative importance of textual units for natural language processing. They also test this technique on the problem of text summarization. In this method, a connectivity matrix based on intra-sentences cosine similarity is used as the adjacency matrix of the graph representation of sentences.

IV. EMDS SUMMARIZATION MODEL

Extractive Multi Document Summarization based is graph based multi document summarization algorithm Algorithm consists of following steps as shown in. The input to the model is a set of related documents. Firstly, the set of documents is pre-processed. The undirected acyclic graph is constructed for each document with sentences as nodes and similarities as edges. Thereafter, weighted ranking algorithm MEAD [3] is performed on the graph to generate salient score for each sentence in the document. The sentences are ranked according to their salient scores. The top-ranking sentences are selected to form the summary for each document. Secondly, all the single summary of each document assembled into one document. Finally, the described above process is applied to this combining document to form the final extractive summary



A.Preprocessing

Before constructing graph, the input set of related documents required to be preprocessed. In the first step, input documents were parsed to extract all sentences. Those sentences, whichtoo short or almost contain no information, are eliminated

B. Undirected Graph Construction

The directed acyclic graph G = (V x E) represent each document were constructed as follow. Each sentence appearing in the document becomes a node in the graph

C. Rank Assignment

Once document graph is built, the sentences in a document will be ranked through random walk on G. We compute a salient score for each node using the MEAD algorithm.

D. High Scoring Sentence Selection

In this step, high scoring sentences are selected by calculating absolute class, summed class, sentence length

E. Summary Generation

In this step, final summary is generated using sentences selected simply, sentences with high ranking scores may be chosen as the final ones in the summary. However, there may be much redundancy among the top ranking sentences, since similar sentences tend to get similar ranking scores during the ranking process. The modified version of Maximal Marginal Relevance (MMR) is applied to re-rank and select sentences to add into summary. A sentence is added if it is high ranked and not too similar to any sentence existing in the summary

V RESULTS EVALUATION:

The two generic summarization methods that we used in comparison have already been established. To evaluate Extreactive Multi Document Summarization , it has compared it with two summarizers: Random and LEAD.

1. Random Summarizer: A baseline summarization system that randomly selects sentences with no repetition till it reaches the desired length of 40 words. RANDOM based technique randomly selects sentences and put them inside summary.

TABLE 1.1

Random Summary Eventhough the cost is high the product is fine for its quality making. It is well built and the size is smaller than the previous one. Product is good as per reviews but not suitable for business purpose. Fabulous.
LEAD Summary The product is considered best in the market. Fabulous. According to reviews of the newspaper, the product is on top. Considering its bad color people will not buy this product. I hope the product will do well in the market.
EMDS Summary Fabulous. The product is considered best in the market. This type of products is good in today's world. There are some nice points which are needed to be mentioned. This is not an excellent product.

2. LEAD Summarizer: In LEAD based technique first sentence is included in the summary depending upon sentence length. LEAD based summarizer selects first sentence of each document, then the second sentence of each, etc. until the desired summary size is met. A LEAD summarization system that selects sentences with no repetition till it reaches the desired length of 40 words.

The resulting summaries are assessed using the automatic metric ROUGE and manual evaluation. Example summaries are shown in Table 1.1.

1.1 MANUAL EVALUATION**:**

In the manual evaluation, it has asked for three people to evaluate the readability of the generated summaries. Without showing the reference summary, we asked each participant to rate the following linguistic qualities with a rating scale ranging from a maximum of 5 (very good) to a minimum of 1 (very poor).

1. Grammaticality: grammatically correct.
2. Redundancy: absence of unnecessary repetitions.
3. Clarity: easy to read.
4. Coverage: coverage of overall aspects.
5. Coherence: well structured and organized.

The average scores for each criterion are shown in Table 1.2.

From Table 1.2 it is clear that the scores for Grammaticality, Redundancy, Clarity and Coherence are in all systems very close to each other. The only gap can be observed in the Coverage metric. This metric expresses how many opinions and aspects are actually covered in the review/summary. The scores indicate that EMDS based on graph is able to generate summaries with a wider range of aspects than the other two systems.

1.2 ROUGE EVALUATION:

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is a well known evaluation method for summarization, which is based on the common number of n-grams between a peer, and one or several model summaries. It is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation.

The metrics taken into consideration for this evaluation are ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-3 (R-3), ROUGE-4 (R-4), ROUGE-L (R-L), and ROUGE-SU4 (RSU4). R-1 and R-2 compute the number of unigrams and bigrams, respectively, that coincides in the

Automatic and model summaries. R-SU4 measures the overlap of skip bigrams between them allowing a skip distance of 4. ROUGE-L is a Longest Common Subsequence (LCS) based statistics. Longest common subsequence problem takes into account sentence level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically.

From Table 1.3 we can see that EMDS outperforms the other two systems in all ROUGE metrics. This means that, according to ROUGE, our summarizer generates summaries whose lexical content is closer to human ones and thus is more likely to capture the summaries than the other two systems.

TABLE 1.2 MANUAL EVALUATIONS

	Random	LEAD	EMDS
Grammatically	3.54	3.68	3.71
Redundancy	2.84	2.90	3.10
Clarity	2.80	2.97	3.05
Coverage	2.69	2.33	3.36
Coherence	2.05	2.60	2.62

TABLE 1.3 ROUGE EVALUATIONS

Sr. No.	Metric (Run ID)	LEAD			RANDOM			EMDS		
		Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
1	R-1	0.51667	0.43851	0.47439	0.42888	0.53026	0.51043	0.54850	0.57328	0.56062
2	R-2	0.34112	0.28911	0.31297	0.27258	0.40158	0.32474	0.41553	0.43446	0.42478
3	R-3	0.28868	0.24431	0.26465	0.2427	0.35848	0.28944	0.38168	0.39920	0.39024
4	R-4	0.24524	0.20724	0.22464	0.22147	0.32797	0.2644	0.35769	0.37425	0.36578
5	R-L	0.51019	0.43301	0.46844	0.42727	0.56790	0.50851	0.54737	0.57210	0.55946
6	R-SU	0.23853	0.16911	0.19791	0.16080	0.24049	0.21844	0.26959	0.28937	0.27913

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. That shows how EMDS algorithm is better than existing system.

REFERENCES

[1] Jade Goldstein*" Multi-Document Summarization By Sentence Extraction" [13] Shanmugasundaram Hariharan1," Enhanced Graph Based Approach For Multidocument Summarization "The International Arab Journal Of Information Technology, Vol. 10, No. 4, July 2013.

[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.[4] 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.

[13] Java, The Programming Language - Http://Www.Oracle.Com/Technetwork/Java/Index.Html