Keyword based Automatic Summarization of HTML Documents

Shivangi Gupta
CSE Dept., M.I.E.T Meerut.

Mukesh Rawat
CSE Dept., M.I.E.T Meerut.

ABSTRACT
Automatic summarization [5] can be defined as the procedure to create a short version of a text by a computer program. Its product still contains the most important points of the existing text. Multi-document summarization [6] can be defined as an automatic procedure which extracts information from multiple texts that is written about the same topic. Resulting summary report allows individual users or professional information consumers, to quickly familiarize themselves with information that is contained in a large cluster of documents. Multi-document summarization creates information reports that are both concise and comprehensive.

General terms
Text summarization

Keywords
Automatic summarization, multi-document summarization, multiple texts, pre- processing of text.

1. INTRODUCTION
Information retrieval (IR) can be stated as finding material i.e. documents of an unstructured nature basically text that satisfies an information needed within large collections. For example, a person is taking out credit card out from his wallet so that he can type the card number is a form of information retrieval. Earlier, information retrieval used to be an activity in which only few people were engaged. They were reference librarians, paralegals, and similar professional searchers. Now, the world has changed and millions of people are engaged in information retrieval via web search engine. Information retrieval tries to retrieve only important information from the documents that are already retrieved. Information Retrieval is the task of finding relevant information from a pool of information. IR is fastly becoming the dominant form of information access. It can also be used to facilitate "semi structured" search i.e. to find a document where title contains Java and body contains threading. IR also covers supporting users in browsing, filtering document collections, further processing a set of retrieved document.

Multi-document summarization can be defined as an automatic procedure [6] which means extraction of information from multiple texts written about the same topic. Summary report we get allows individual users or professional information consumers, to quickly familiarize themselves with information present in large cluster of documents. It produces a single summary from a set of related source documents. Information reports created by multi-document summarization are both concise and comprehensive. It reduces time and effort by pointing at the relevant or main text. It presents the information extracted from multiple sources algorithmically.

An example of summarization technology is search engines i.e. Google.

1.1 Work description
Initially, text is in HTML form and then it is converted in text form using HTML to text parser convertor. Then, text is saved in repository and pre-processing of text is performed. In pre-processing, stop words, cue words and basic dictionary words are removed. Text so generated is in paragraph form and we need to make sentences. Sentences can be generated by separating paragraph with full stop, comma, colon or semi-colon. Then, sentences are saved in repository. Now, score of each sentence is found and this score is used to find the final score. Then, threshold value is calculated to generate summary.

2. LITERATURE REVIEW

2.1 Domain Specific Document Summarization by sentence extraction

In this, sentence extraction and clustering approach are used.

With sentence extraction approach, a small number of sentences which are related to each other are selected from each cluster of the particular category to form a summary. In sentence extraction approach, sentences are extracted from multiple research papers and then ranked accordingly. An extractive multi-document summarizer was described by Radev, whose purpose was to extract a summary from a set of documents and the summary was extracted on the basis of document cluster centroids. Sentences extracted from the documents describe part contents to a certain extent.

In this strategy, clustering is frequently used to eliminate redundant information which results from the multiplicity of the original documents.

2.2 Multi-Document Summarization using Sentence Extraction
In this, we discuss a text extraction approach. Multi-document summarization differs from single document summarization in issues of compression, speed, redundancy and passage selection. Standard Information Retrieval systems firstly find the documents and then rank them based on maximizing relevance to the user query. There are many systems which include relevance assessments of sub-documents and then convey that information to the user Multi-document summarization summarizes either complete documents sets, or single documents in the context of previously summarized ones.

Firstly, we segment the documents into passages, and index them.

Secondly, identify the passages relevant to the query using a threshold.

Thirdly, apply the MMR-MD metric, for that selects a number of passages to compute passage redundancy then, use the

passage similarity scoring as a method of clustering passages depending on length of summary.

Fourthly, reassemble the selected passages to form a summary document using summary cohesion criteria.

2.3 Entropy based Multi-Document Summarization

This technique consists of a set of algorithms and mathematical operations that are being performed on sentences or phrases in a document so that we can identify the relevant sentences. The process of summarization which humans perform is not clearly understood so, we try summarization using tools like decision trees, graphs, word nets and clustering algorithms. Methods that are based on word frequency counts which are obtained after analysing the document summarized performs well in case of multiple documents.

Sentence selection technique- Document we want to summarize is firstly said that it belongs to this particular domain. Then, we use a database of documents and cluster them into domains and topics. After that, an entropy model for various words and collocations is being generated. The entropy values we get are applied to each sentence in the document set and on its basis sentence ranking formula is being computed.

Redundancy removal- Before applying the entropy based ranking formula, we use a graph representation of sentences to detect and remove redundancy.

- We will make a directed graph and in that every sentence will be represented as a node.
- Then, a link will be established from one sentence node to another if at least three non-stop-words are common to them.
- If the parent node represents a longer sentence than what the child node represents, then link weight will be calculated as the ratio of number of words that are common to both the sentences to the length or total number of non-stop words of the child node.
- If not, then the link weight is given by the ratio of common words to the length of the parent node.
- For every parent node, the child nodes which have a link weight greater than a particular threshold and which are shorter than the parent node are excluded from the sentence ranking process.
- Hence, sentences that have been repeated and sentences that are almost similar in word composition are thrown away.

2.4 Single Document Text Summarization Algorithm using semantic similarity

KDT or Knowledge Discovery Text can extract both implicit and explicit concepts and can develop semantic relations between the texts. Extracting concepts and developing relations are the problems of KDT. Text summarization techniques are classified on the basis of summarization they are performing on text.

Two classifications of text summarization techniques are:

- Extractive and Abstractive Text Summarization
- Single Document and Multi Document Text Summarization

Semantic similarity can be defined as a concept in which a set of documents or terms within term lists are assigned a metric based on the likeness of their meaning / semantic content

Classification of semantic similarity

Edge Counting Methods – A method used for measuring the similarity between two terms by the length of the path linking the two terms and by the position of the terms in the taxonomy.

Information Content Methods– a method used to measure the difference in text of the two terms by calculating the probability of occurrence in a text document,

Feature based Methods– a method used to measure the similarity between two terms by examining the properties or on the basis of their relationships with other similar terms in the taxonomy.

Hybrid methods – this method combines the above three methods for calculating the semantic similarity.

3. DESCRIPTION OF THE PROPOSED MODEL

- A HTML document is taken and converted into file document using HTML to text parser.
- By giving the URL of the document text from the document is extracted by using file parser and text is stored in document repository.
- Now, pre-processing of the text is done by removing stop words (I, an, am ,the of), cue words (hence, summary ,conclusion) and basic dictionary words.
- Then, the sentences are extracted from the text document by separating the text with comma, full stop, colon, semi colon and the sentences are stored with indices in sentence repository.
- Then, scoring is done by finding the score of each word on the basis of their frequency of occurrence in the document.
- Sum of word scores gives the sentence score.
- Then, final score is calculated by multiplying the score of each sentence with the ratio “average length / current length” where average length is the ratio of total length of each sentence to the total number of sentences and current length is the length of each sentence.

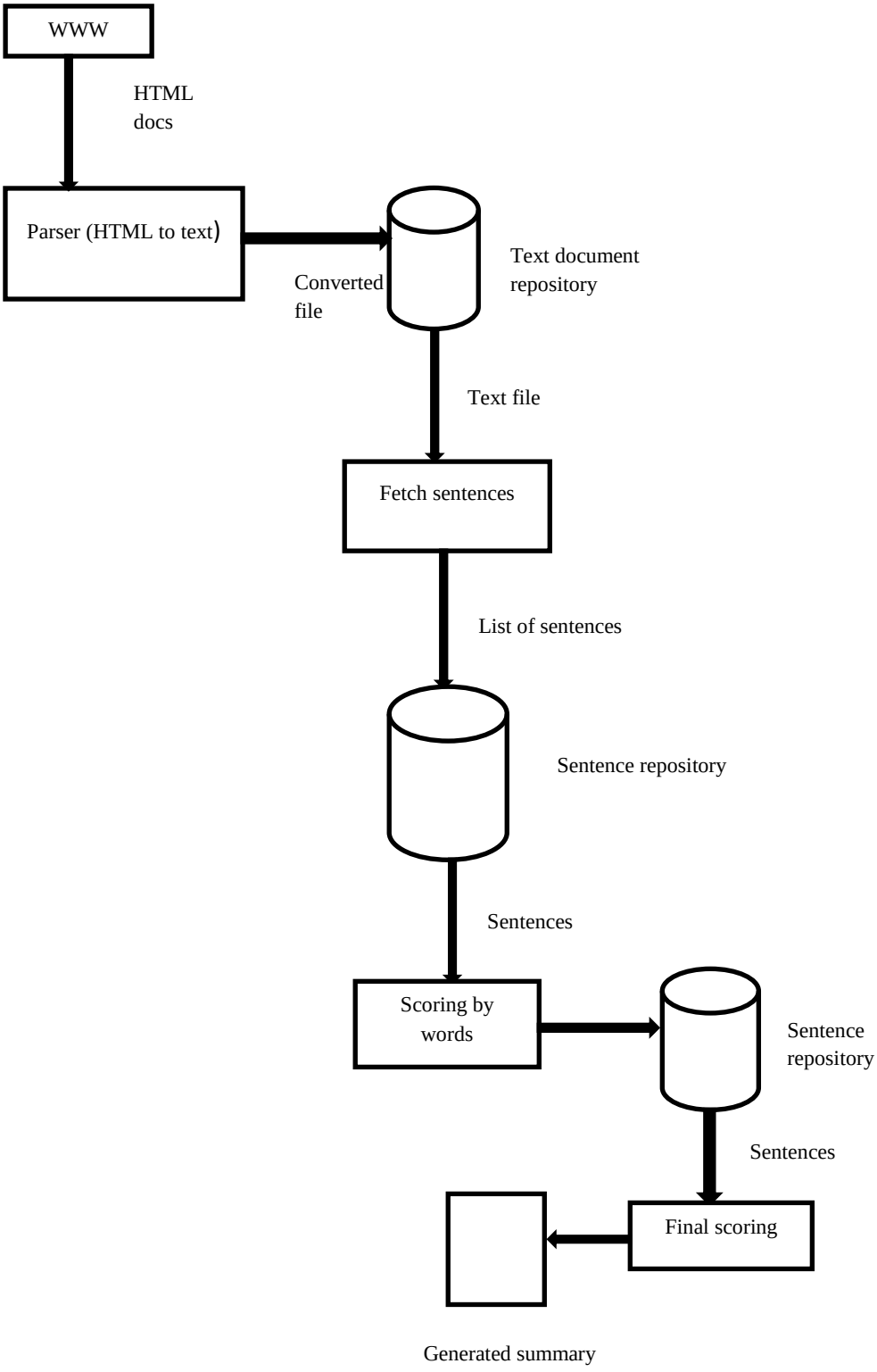

Figure 1: General Architecture for sentence scoring

4. PROPOSED WORKPLAN

I. Conversion of text

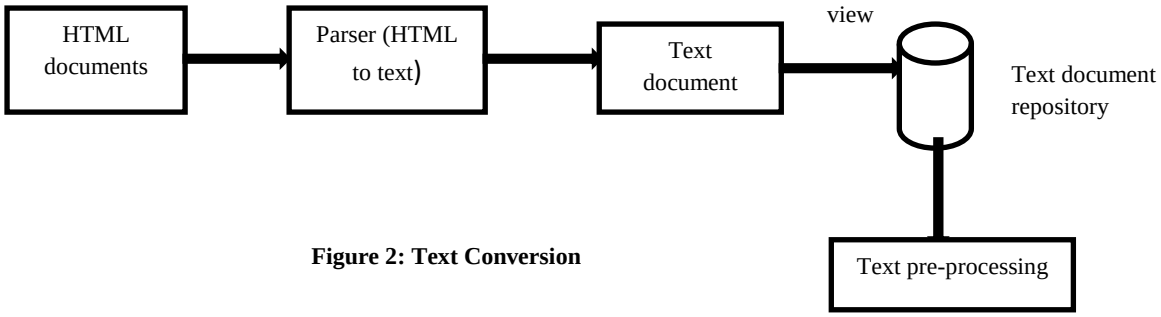

Figure 2: Text Conversion

A HTML document is taken and it is converted into text document using HTML to text parser. View this text document and save it in repository (database).

Perform pre-processing of text by removing stop words, cue words and basic dictionary words.

Stop Words are insignificant words that are commonly used in English language. Ex- I, a, an, of, am, the, etc.

Cue Words are words that are usually used in concluding sentences of a text. Ex:- thus, hence, summary, conclusion, etc.

Basic Dictionary Words are most frequently used words in English language. There are around 850 words in English language.

II. Sentence extraction

Text after pre-processing is in paragraph form so, convert it in sentence form..

Then, generate sentences by separating paragraphs with full stops, comma, semicolon, and colon and save the sentences in repository**.**

III. Scoring to generate summary

Scoring is done by calculating score of each word and score can be defined as the frequency of occurrence in the document. After that score of each sentence can be calculated by summing up the score of each word against each sentence. Now, find the final score by multiplying the score of each sentence with the ratio “average length / current length” where average length is the ratio of total length of each sentence to the total number of sentences and current length is the length of each sentence. Then, calculate the threshold value to generate summary. Threshold value can be calculated as:

TH=n(s) / n

Where, n(s) is the sum of the scores of the sentences of documents and n is the number of sentences of the document.

Output screens of the proposed model are:

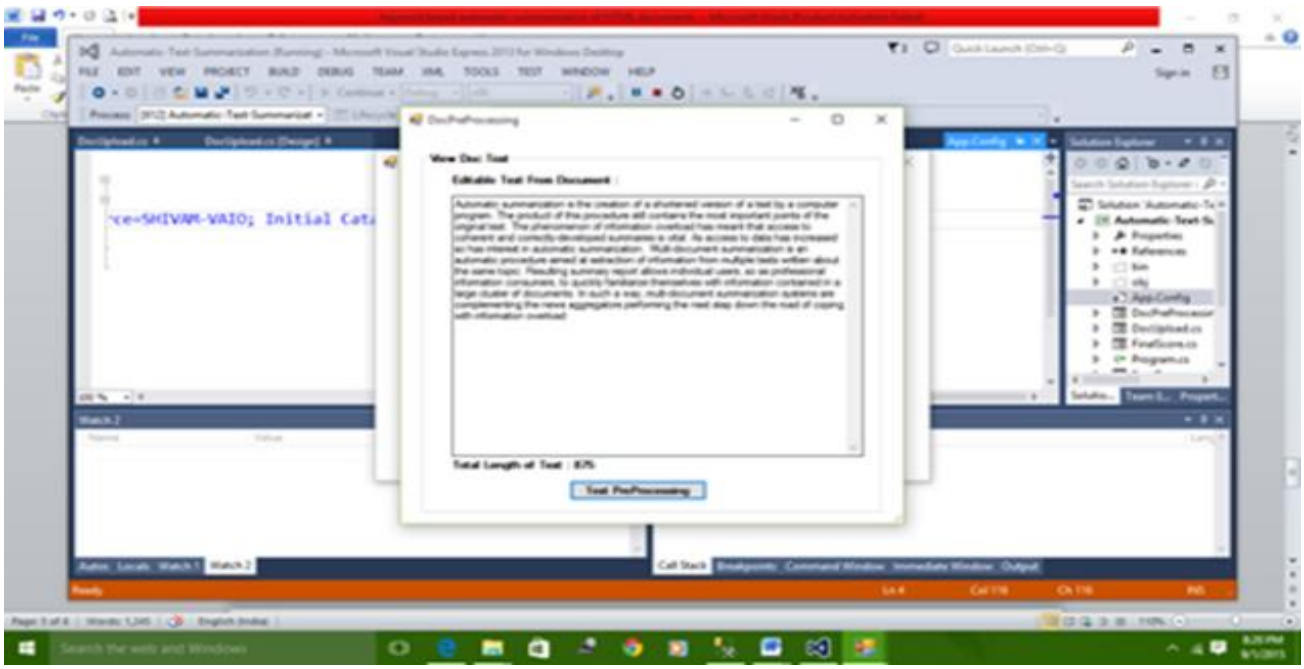

Figure 3: Document pre-processing snapshot

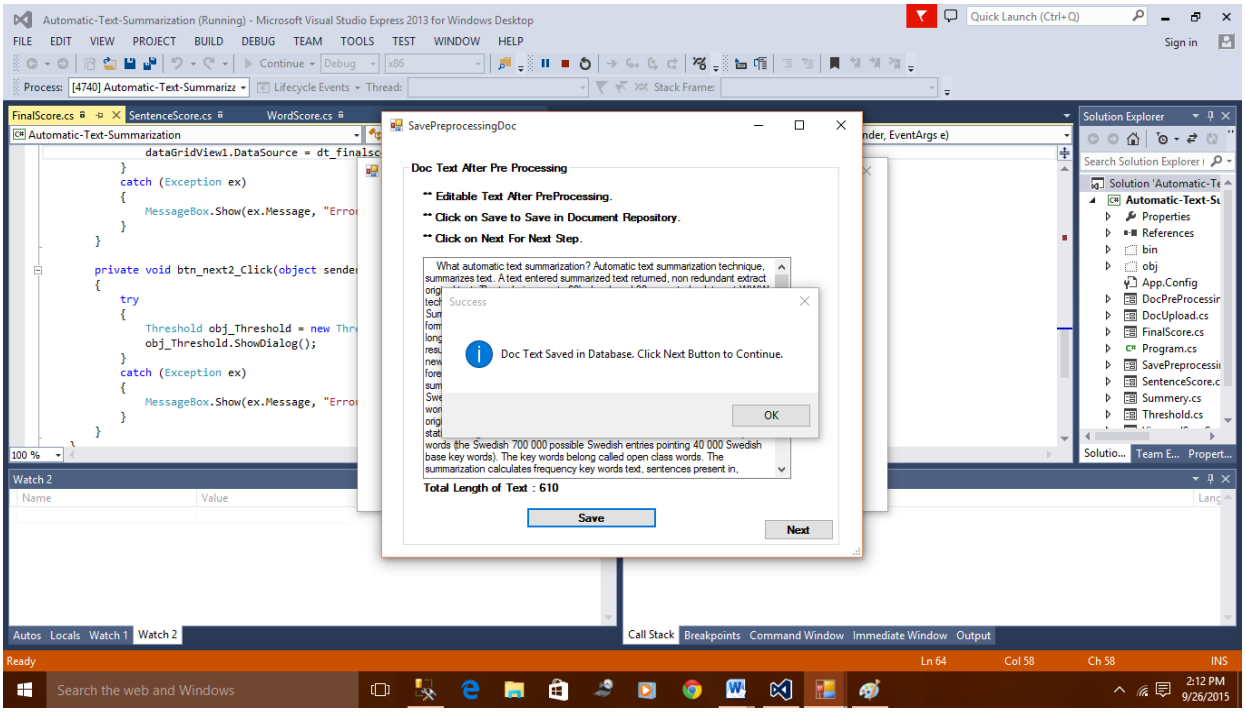

Figure 4: Save pre-processing of document snapshot

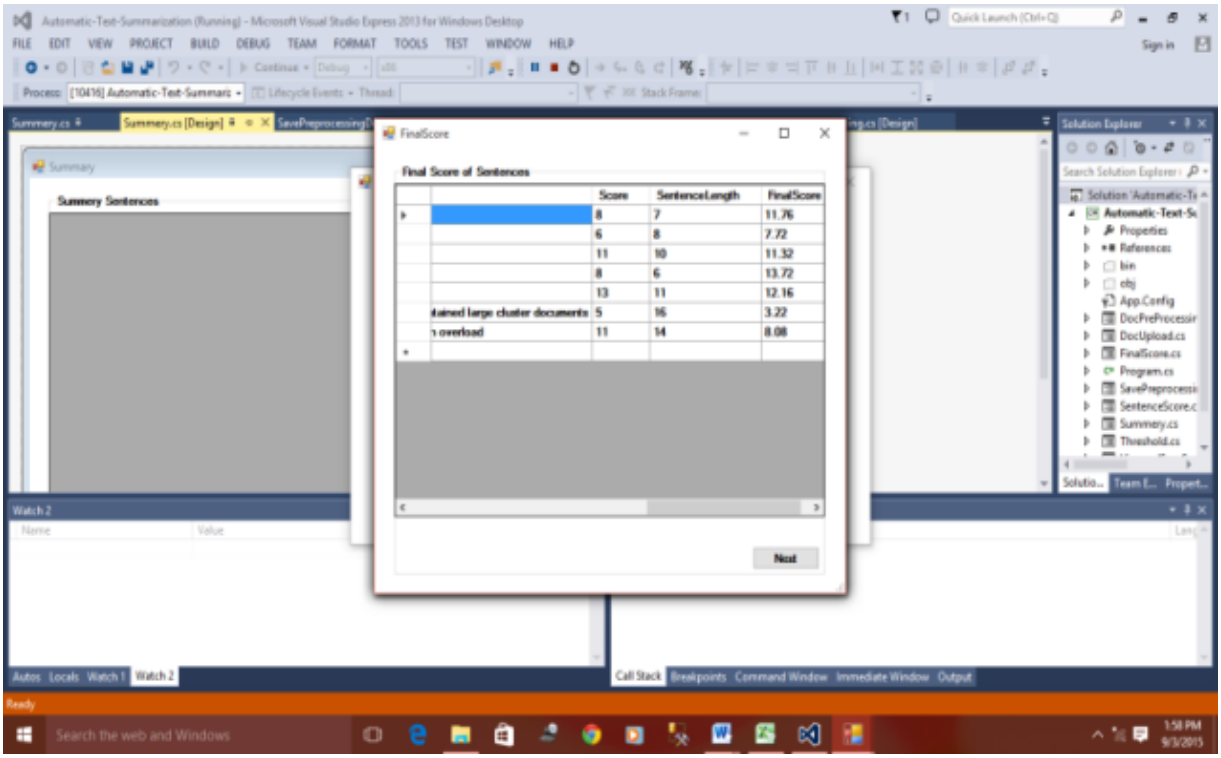

Figure 5: Final score snapshot

5. RESULT ANALYSIS

Table 1: Experimental Result Analysis

Document name	No. of words after pre-processing	No. of words in summary	%age reduction
Doc1	46	40	13.04
Doc2	64	44	31.25
Doc3	97	66	31.95
Doc4	70	55	21.42
Doc5	75	22	70.66
Doc6	88	51	42.04
Doc7	64	38	40.62
Doc8	46	14	69.56
Doc9	66	66	0
Doc10	39	25	35.89
Doc11	87	72	17.24

Doc12	44	31	29.54
Doc13	54	18	66.66
Doc14	56	30	46.42
Doc15	64	39	39.06

avg. = 34.89

For performing experimental result analysis take 15 documents, firstly determine no. of words after pre-processing (removing stop words, cue words, basic dictionary words) then find the no. of words in summary by executing the whole procedure. For each document, calculate %age reduction.

%age reduction= 100- [(No. of words in summary/ no. of words after pre-processing)*100]

Then, find the average of %age reduction by summing up the %age reduction of all the documents and then dividing it by the no. of documents.

Average of %age reduction=34.89 which shows a good result.

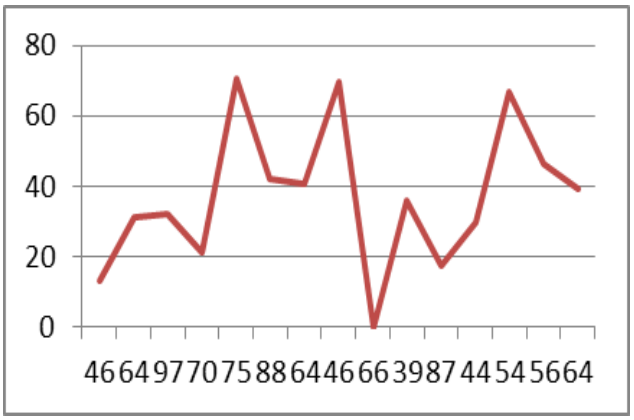

Figure 6: Graphical Representation of result analysed

On X-axis; no. of words after pre-processing is taken and on Y-axis; %age reduction is taken. And graph is plotted as shown in figure 6. This graph shows that as the no. of words are increasing, %age reduction also increases which shows a very good result.

6. CONCLUSION AND FUTURE SCOPE

There are many immediate applications for this system. They include integration into a search engine, so that document summaries are provided instead of documents themselves. The approach discussed above gives us a new idea for creating summaries from text of multiple similar documents by sentence scoring, there is also a new technique for reducing the summary size as the number of documents are increasing by selecting the sentences whose score are more as compared to the other sentences of the summary, but having certain limitations such as without the use of NLP, the generated summaries suffers from lack of cohesion and semantics, it is difficult to relate pronouns to their corresponding nouns in the summary. The possibilities are endless.

With Natural Language Processing:

a. Newspaper headlines can be generated.

b. Forms can be filled up.

c. Bio-data can be generated.

Some can see that some modification could be done to the current system to allow multi-document summarization. Identify sentences to be extracted. Match these sentences across documents, using some form of similarity metric. Filter out repeated sentence, retaining the more salient ones. Reduce sentences using generalization and aggregation. Present the information.

7. ACKNOWLEDGMENTS

The authors are thankful to Director, Meerut Institute of Engineering and Technology, Meerut for his support and guidance.

8. REFERENCES

[1] James Allan, Jaime Carbonell, George Doddington,Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study: Final report. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.

[2] Chinatsu Aone, M. E. Okurowski, J. Gorlinsky, and B. Larsen. 1997. A scalable summarization system using robust NLP. In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pages 66-73, Madrid, Spain.

[3] Breck Baldwin and Thomas S. Morton. 1998. Dynamic coreference-based summarization. In Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3), Granada, Spain, June.

[4] Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In Proceedings of the CL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pages 10-17, Madrid, Spain.

[5] A. Siddharthan, A. Nenkova, and K. McKeown. Syntactic simplification for improving content selection in multi-document summarization. In Proc.of COLING, 2004.

[6] L. Vanderwende, H. Suzuki, and C. Brockett. Microsoft Research at DUC2006: Taskfocused summarization with sentence simplification and lexical expansion. In Proc. of DUC, 2006.

[7] X. Wan and J. Yang. Improved affinity graph based multi-document summarization. In Proceedings of HLT-NAACL, Companion Volume: Short Papers, pages 181–184, 2006.

[8] D.Zajic, B. Dorr, andR. Schwartz.Automatic headline generation for newspaper stories. In Proc. of DUC, 2002.

[9] D. Zajic, B. Dorr, J. Lin, C. Monz, and R. Schwartz. A sentence-trimming approach to multidocument summarization. In Proc. of DUC, 2005.

[10] Satoshi Sekine and C Nobata. Sentence Extraction with Information Extraction Technique. In Workshop on Text Summarization, 2001.

[11] Christopher D. Manning, Prabhakar Raghavan," An Introduction to information retrieval",Cambridge University Press Cambridge, England, Online edition (c) 2009.