IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. II (May-Jun. 2014), PP 42-49
www.iosrjournals.org

An Approach to Single Documnent Text Summarization & Simplification

Shubhangi C. Tirpude[1]
Assistant Professor Shri Ramdeobaba College of Engineering & Managemnet, Nagpur, Maharashtra

Abstract: *The amount of information available on the internet is increasing day by day which is leading to information overload. For more information than can realistically be digested is available on the World-Wide Web and in other electronic forms. News information, biographical information, minutes of meetings missed, it isn't possible to read everything one would want to read and so some form of information condensation is needed. Summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Somplification os done to append meanings of the nouns so as to facilitate the easy reading and understanding of the text by the user.*

In this summarization. the input text document is divided into two parts informative & non informative and summarizing and simplifying them individually. It uses the NLTK tagger to tag the words and get their parts of speech. The nouns in the informative sentences were simplified using WorldNet. In order to summarize and simplify non-informative sentences keyword selection approach was used. Finally, the two files obtained after summarizing and simplifying were merged together to obtain the output file.

Keywords: *keyword selection,, Weighted approach, NLTK Tagger, Text Features, keyword selection, informative & non informative sentence.*

I. Introduction

Text summarization has become an important tool for analyzing and interpreting text documents in a fast growing information world. According to Lin and Hovy [1] "Summary can be defined as a text that is produced from one or more texts, that contain a significant portion of the information in the original text(s), and that is no longer than half of the original text(s)"

1.1 Types of Summary

Summaries of text can be divided into different categories, some of them harder to automate than others.

Origin Based Summary

One division is based on the origin of the text in the summary: Extractive & Abstractive Extractive summary is where the summary consists of sentences that have already appeared in the text. Methods for extractive Summarization are: Term Frequency Inverse Document Frequency Method. Cluster Based Method, Graph Theoretic Approach, Machine Learning Approach,LSA Method, Text Summarization with Neural Networks ,Automatic Text Summarization based on Fuzzy Logic, Text Summarization using regression for estimating feature weights, Multi – document Extractive Summarization, Query Based Extractive Text Summarization, Multilingual Extractive Text Summarization[5]

Many variations of the extractive approach have been tried in the last ten years. However, it is hard to say how much greater interpretive sophistication, at sentence or text level, contributes to performance. Without the use of NLP, the generated summary may suffer from lack of cohesion and semantics. If texts containing multiple topics are summarized, the generated summary might not be balanced .Deciding proper weights of individual features are very important as quality of final summary is depending on it.

Abstractive summarization techniques are broadly classified into two categories: Structure Based approach, & Semantic Based Approach

Different methods that use structured based approach are Tree Base Method, Template Based Method, Ontology Based Method, Lead and Body Phrase Method, Rule Based Method.

Methods that use semantic based approach are Multimodal Semantic model, Information Item Based Method, Semantic Graph Based Method [8].

Purpose Based Summary

Summaries can also be categorized by their purpose:

Indicative: These summaries are meant to give the reader an idea as to whether it would be worthwhile reading the entire document. The topic and scope of the text should be expressed but not necessarily all of the factual content.

Informative: This type of summary expresses the important factual content of the text.
Critical: This sort of summary criticises the document. It expresses an opinion on | in the case of a scientific paper, say the methods employed and the validity of the results.
Indicative Summaries are the most feasible to automate, out of the three, and critical summaries probably the least. Informative summaries are a little harder than indicative ones, since fuller coverage of the information in the text is required.

II. Summarization Approach

It is to indicate which sentences to be added to the summary. This is divided in to four categories: Statistical, Linguistic, Hybrid & Rhetorical approach.

2.1 Statistical Method

This extracts sentences that occurred in the source text, without taking into consideration the meaning of the words. It uses the techniques from information Retrieval. Some of the parameters used are Word frequency , Position of Sentence or words in the sentence[1].

2.2 Linguistic Method

In this, method needs to be aware of and know deeply the linguistic knowledge, so that the computer will be able to analyze the sentences and then decide which sentence to be selected. It identifies term relationship in the document through part-of-speech tagging, grammar analysis, thesaurus usage, and the like, and extract meaningful sentences. Parameters can be cue words, title feature or Noun and verbs in the sentences [2]. Statistical approaches may be efficient in computation but linguistic approaches look into term semantics, which may yield better summary results. In practice, linguistic approaches also adopt simple statistical computation (term-frequency-inverse-document-frequency (TF-IDF) weighting scheme) to filter terms. It is however, relatively few researches in literature to discuss the linguistic approaches that adopt a term weighting scheme derived from a formal mathematical (probabilistic) model to make more sense in weight determination.

2.3 Hybrid Method

It exploits best of both the previous method for meaningful and short summary [3].

2.4 Rhetorical Method

Rhetorical structure theory (RST) is based on the Rhetorical connections between different parts of the text. In this theory the Rhetoric behind the decomposed text is extracted. In summarization systems, Rhetorical structure (RS) presents the logical connections between different parts of the text and interprets these connections. These information represent the discourse structure and features of the main document .After identifying text units and rhetorical connections between them, the RS tree is formed based on these information.[9]

III. Summarization Features

Text summarizer identifies and extracts key sentences from the source text and concatenates them to form a concise summary. In order to identify key sentences for summary, a list features as discussed below that can be used to for selection of key sentences.[4]
Keyword Selection: Depending on the domain for which the text summariser is going to work, important words relating to that domain are selected. The keywords are assigned a weight/priority number depending on which the sentences are either selected or rejected for the summary.
Term Frequency: Statistics provide salient terms based on term frequency, thus salient sentences are the ones that contain the words that occur frequently. The score of sentences increases for each frequent word. The most common measure widely used to calculate the word frequency is TFIDF.
Location: It relies on the intuition that important sentences are located at certain position in text or in paragraph, such as beginning or end of a paragraph. Based on this location the sentences are either selected or rejected and then the summary is generated.
Cue Method: Words that would have positive or negative effect on the respective sentence weight to indicate significance or key idea such as cues: 'in summary', 'in conclusion', 'called as', 'significantly'.
Title/Headline word: It assumes that words in title and heading of a document that occur in sentences are positively relevant to summarization.
Sentence length: Short sentences express less information and therefore excluded from summary. Keeping in view the size of summary, very long sentences are not appropriate for summary.

Similarity: This feature determines similarity between the sentence and the rest of the document sentences and similarity between the sentence and title of the document. Similarity can be calculated with linguistic knowledge or by character string overlap.

Proper noun: Sentences having proper nouns are considered important for document summary. Examples of proper nouns are: name of a person, place or organization.

IV. Litrature Review:

Vishal Gupta & Gurpreet Singh Lehal (2012): Extractive summaries are formulated by extracting key text segments (sentences or passages) from the text, based on statistical analysis of individual or mixed surface level features such as word/phrase frequency, location or cue words to locate the sentences to be extracted. The "most important" content is treated as the "most frequent" or the "most favorably positioned" content. 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 that conveys the most important information from the original text document. In this paper, a Survey of Text Summarization Extractive technique has been presented.*[5]

Atif Khan, Naomie Salim (2014): focused on extractive summarization, which forms summary by selection of important sentences from the documents. Statistical methods are often used to find key words and phrases.[6]

Md. Majharul Haque, Suraiya Pervin, and Zerina Begum (2013***):*** In this paper, automatic multiple documents text summarization task is addressed and different procedure of various researchers are discussed. Various techniques are compared here that have done for multi-document summarization. In multi-document summarization, several key points are involved, such as reducing each document, incorporating all document's significant idea, compare the ideas found from each, ordering sentences come from different sources keeping the logical and grammatical structure right. Some promising approaches are indicated here and particular concentration is dedicated to describe different methods from raw level to similar like human experts, so that in future one can get significant instruction for further analysis. [7]

V. Proposed Approach

Fig. 1.illustrates proposed approach. The input file contains information on the medicine penicillin. The Input File is Split into informative and non-informative files. The Simplification Process is performed on the informative file. The nouns are identified, their meaning is searched in the WordNet, and the definition for the same is appended in the bracket that follows the noun.

Each word is tagged with its respective Part of Speech tag. It can be seen that the file becomes bulkier due to this application of NLTK Tagger. Grammar Rules specified before are applied and changes like 'can be reduced' is replaced by 'reduces' & 'should be discarded' is replaced with 'discards' etc.
The development & implementation of the proposed work will be carried out using different steps. The steps are as follows:

Input File: The input file considered is from the medical related data

Informative Sentences: The sentences consisting of words such as 'called' , 'defined', 'known', 'referred' , 'date' , 'time' i.e. which give some useful information are declared as informative sentences. I have used the 'cue method' text summarization feature for the implementation of this part

NLTK Tagger: The nltk.tag module defines functions and classes for manipulating *tagged tokens*, which combine a basic token value with a tag. *Tags* are case-sensitive strings that identify some property of a token, such as its part of speech.

WordNet: WordNet helps in grouping English words into sets of synonyms called synsets. It provides short, general definitions, and records the various semantic relations between these synonym sets.

Non-Informative: All the remaining sentences of the input text that do not belong to the informative category are declared as non-informative. This part of the input file was chosen for the purpose of file summarization as the file size attained here was larger. Informative sentences need to be present in the summary; and hence cannot be summarized further; this basic logic was the deciding factor.

Grammar Rules: The grammar rules were used to reduce the length of the non-informative sentences. The rules that we have used in our project are as follows:
MD+VB+VBN (VBN ends with ed) → VBN+es
MD+VB+VBN(VBN ends with es) → VBN+sses
MD+VB+VBN(VBN ends with y) → VBN+ies

Keyword Selection via Weighted Approach:
The weights are assigned to the keywords based on their priority or importance in relation to the omain text. The input text is them compared in a sentence wise manner with these keywords i.e the keywords that are above the preset threshold value, and corresponding sentences are selected to be displayed in the summarized text file.

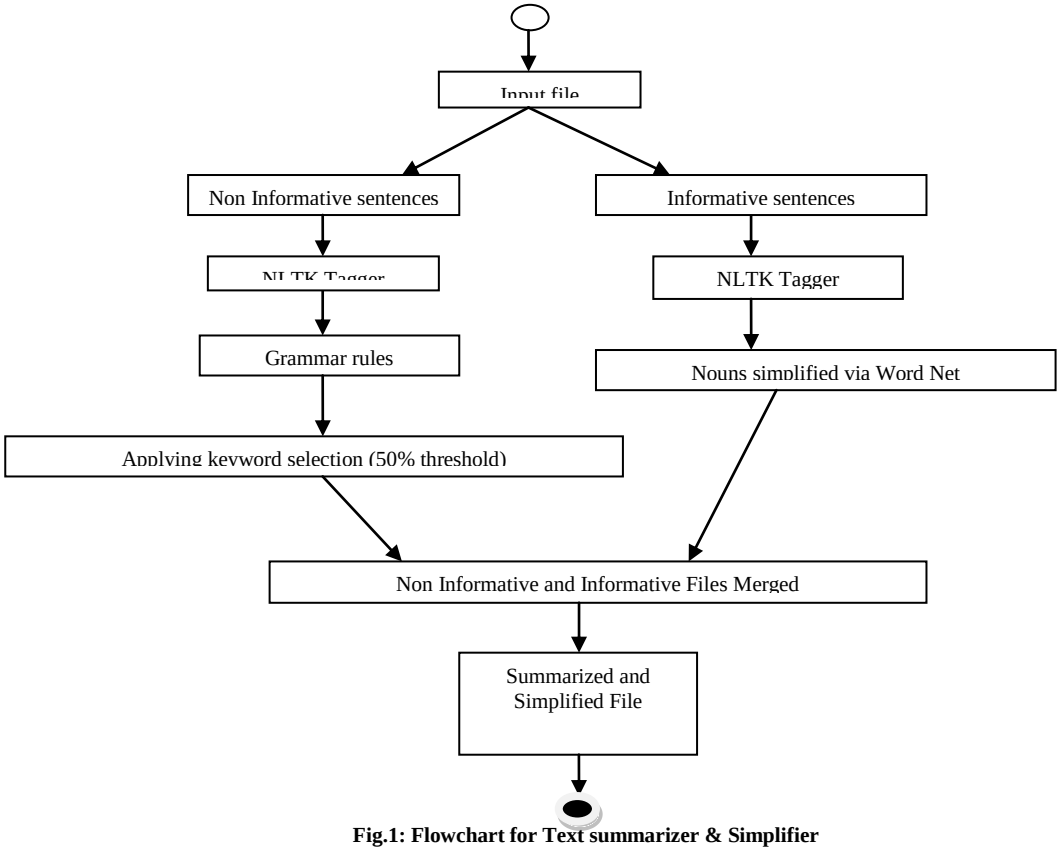


Fig.1: Flowchart for Text summarizer & Simplifier

VI. Implementation

The Input File (Fig 2) is Split into informative (Fig 3) and non-informative files (Fig 4). The Simplification both files are tagged using NLTK Tagger (Fig 4 & Fig 5). Simplification process is performed on Informative file.(Fig 6). The Summarization Process is performed on the non- informative file using sentence length reduction. A threshold value of 50% is applied to the list of keywords and their respective weights. Only one sentence contains the keyword. The summarized and the simplified files(fig 7) are merged in logical ordering and a final output file is generated(fig. 8).

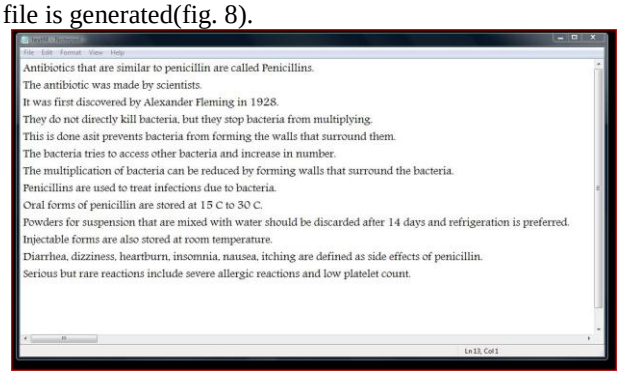

Antibiotics that are similar to penicillin are called Penicillins.
The antibiotic was made by scientists.
It was first discovered by Alexander Fleming in 1928.
They do not directly kill bacteria, but they stop bacteria from multiplying.
This is done asit prevents bacteria from forming the walls that surround them.
The bacteria tries to access other bacteria and increase in number.
The multiplication of bacteria can be reduced by forming walls that surround the bacteria.
Penicillins are used to treat infections due to bacteria.
Oral forms of penicillin are stored at 15 C to 30 C.
Powders for suspension that are mixed with water should be discarded after 14 days and refrigeration is preferred.
Injectable forms are also stored at room temperature.
Diarrhea, dizziness, heartburn, insomnia, nausea, itching are defined as side effects of penicillin.
Serious but rare reactions include severe allergic reactions and low platelet count.

Fig. 2 Input File

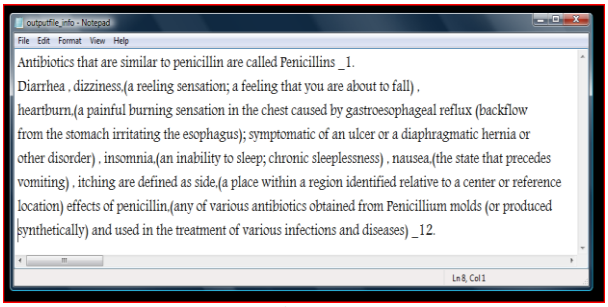


Fig.3: Informative File

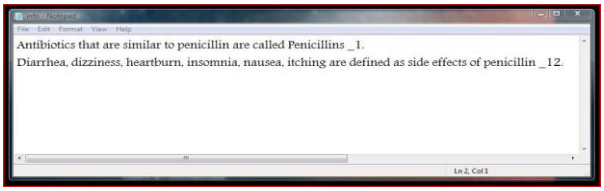


Fig.4: Non Informative File

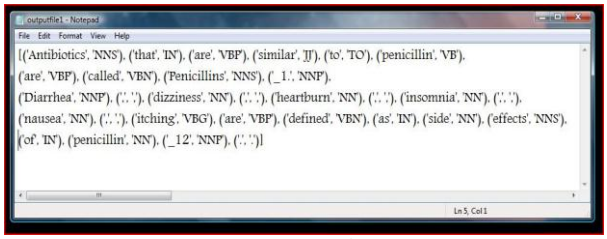


Fig.4: NLTK Tagged Informative File

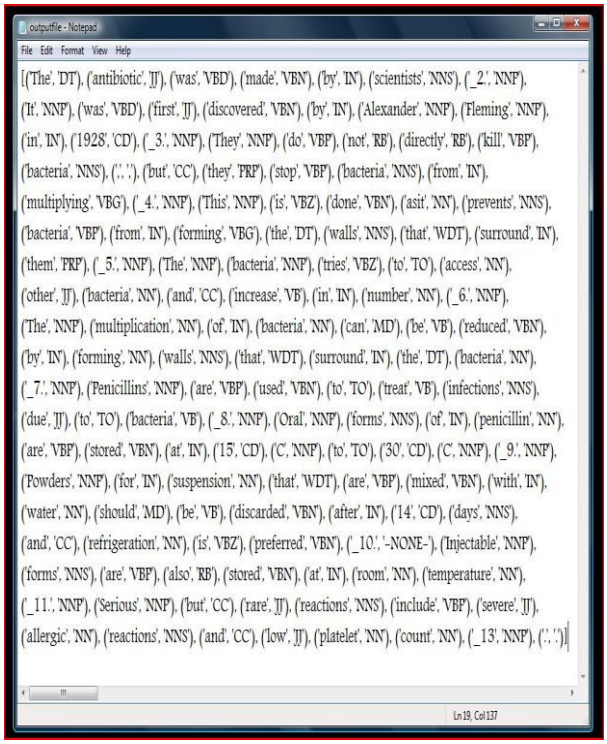


Fig.5: NLTK Tagged Non Informative File

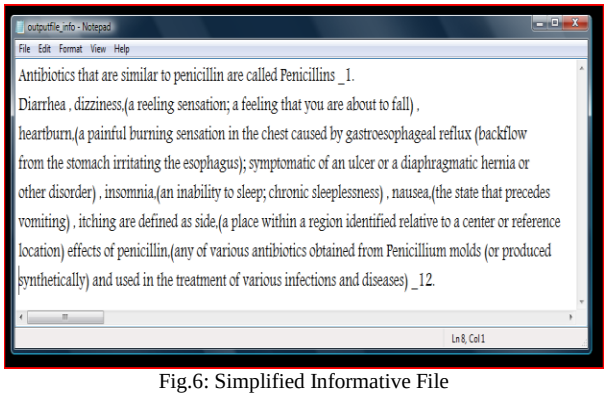


Fig.6: Simplified Informative File

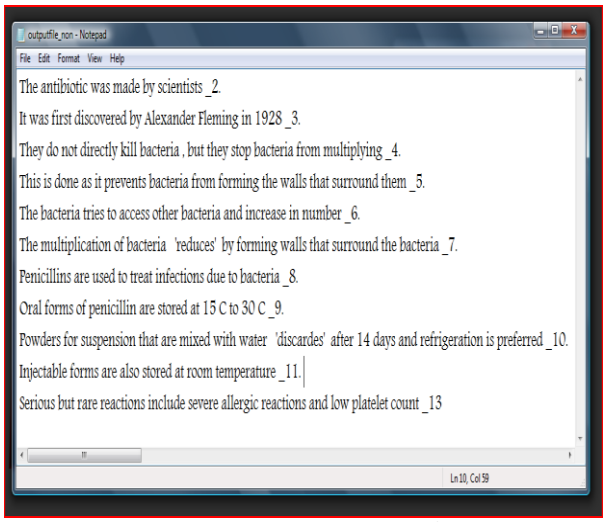


Fig.6: Sentence Length Reduced file

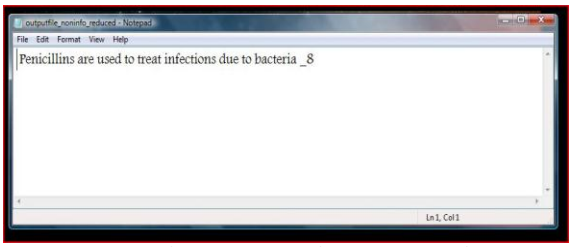


Fig 7.Summarized Informative File (After application of Keyword Selection)

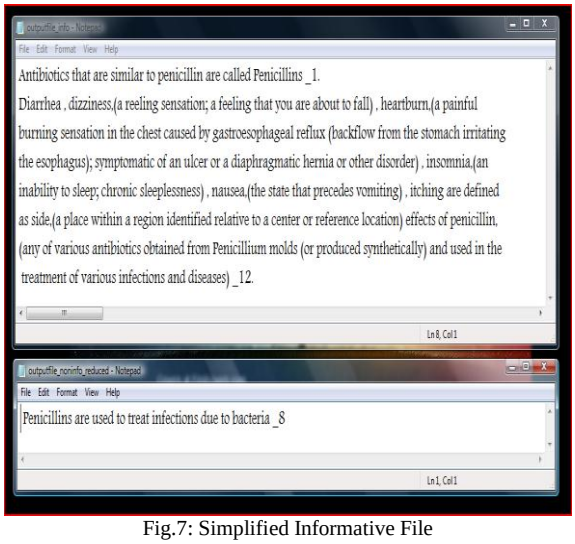
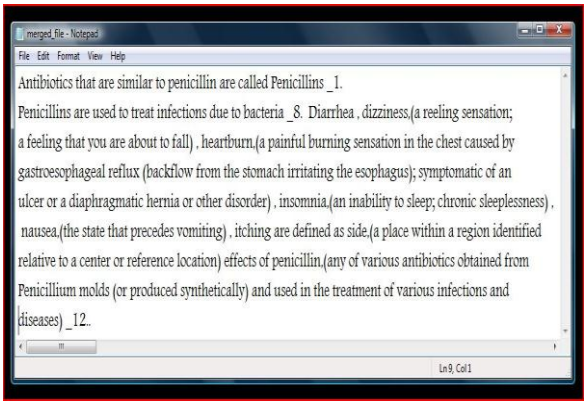


Fig.7: Simplified Informative File



VII. Evaluation Parameters

Evaluation methods are useful in evaluating the usefulness and trustfulness of the summary.
Following are the Parameters on which it is evaluated

Precision: It evaluates correctness for the sentences in the summary

$$P = \frac{\text{Retrieved Sentences} \cap \text{Relevant Sentences}}{\text{Retrieved Sentences}}$$

Recall: It evaluates proportion of relevant sentences included in summary

$$R = \frac{\text{Retrieved Sentences} \cap \text{Relevant Sentences}}{\text{Relevant Sentences}}$$

Where Retrieved Sentences: Retrieved from system Relevant Sentences: Identified by human

Overall Fitness Measure (F) : Combination of P & R

$$F = \frac{2 \times \text{Recall} \times \text{Precision}}{(\text{Recall} + \text{Precision})}$$

Compression Ratio: Small (Tending towards zero)

$$CR: \text{length S} / \text{length T}$$

Retention Ratio : Large (Tending to Unity)

RR = info in S / info in T

Where S is summarized text & T is original text

in this paper we perform evaluation using formula 4 and 5 as under:

- Compression ratio: 4/15=0.26
- Retention Ratio: 60/958=0.063

VIII. Future Scope & Conclusion:

The work is currently focused on single document domain specific text summarization. This work can be extended to

- Anaphoric Resolution
- Multi Document Summarization
- Implementation of NER
- Multi-Lingual Summarization

Currently, text summarization is one of the hot areas of research and attracts lots of attentions from different fields. Text summarizations systems can be categorized in to various groups based on different approaches were presented in this paper. As discussed earlier, paper focuses on different types of summarization methods, which might be used in a system for generating a summary; It also discuss the most important issues in evaluating a summary and present common criterion for evaluating a summarization system

References

[1] Hovy, E. H. Automated Text Summarization. In R. Mitkov (ed), The Ox-ford Handbook of Computational Lin-guistics, chapter 32, pages 583-598.*Oxford University Press*, 2005.

[2] Dehkordi, P.K., Khosravi, H., Kumarci, F., 2009. Text Summarization Based on Genetic programming. *International Journal of Computing* and ICT Research, 3(1), 57-64.

[3] Te-Min Chang Wen-Feng Hsiao "A hybrid approach to automatic text summarization" IEEE 2008

[4] *Vishal Gupta & Gurpreet Singh Lehal (2012):* Paper Review on Text Summarization Extractive Techniquesjournal of emerging technologies in web intelligence, vol. 2, no. 3, august 2010

[5] Atif Khan, Naomie Salim (2014): Paper Review on Abstractive Summarization Methods journal of Theoretical andApplied Information Technology,10,th,January 2014. Vol. 59 No.1 ,© 2005 - 2014 JATIT & LLS. All rights reserve

[6] Md. Majharul Haque, Suraiya Pervin, and Zerina Begum(2013**)** Md. Majharul Haque, Suraiya Pervin, andZerina Begum, "Literature Review of Automatic Multiple Documents Text Summarization," International Journal of Innovation and Applied Studies, vol. 3, no. 1, pp. 121–129, May 2013.

[7] Kavita Ganesan and ChengXiang Zhai and Jiawe Han "Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions

[8] Daniel Marcu "Discourse Tree Good Indicator for Important Text" S. Zhang, C. Zhu, J. K. O. Sin, and low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.