International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 4 524 -530

Summarizing Text Using Lexical Chains

Pooja Jain
Dept. of CE & IT
Suresh Gyan Vihar University, Jaipur
Raj., India
erpoojajain92@gmail.com

Sachin Jain
Assistant Professor, Dept. of CE & IT
Suresh Gyan Vihar University, Jaipur
Raj., India
sachin.jain@mygyanvihar.com

Abstract: The current technology of automatic text summarization imparts an important role in the information retrieval and text classification, and it provides the best solution to the information overload problem. And the text summarization is a process of reducing the size of a text while protecting its information content. When taking into consideration the size and number of documents which are available on the Internet and from the other sources, the requirement for a highly efficient tool on which produces usable summaries is clear. We present a better algorithm using lexical chain computation. The algorithm one which makes lexical chains a computationally feasible for the user. And using these lexical chains the user will generate a summary, which is much more effective compared to the solutions available and also closer to the human generated summary.

Index Terms— Text Summarization, Lexical Chains, Summary Generation

*****

I. INTRODUCTION

A summary may be defined as a text that's created from one or a lot of texts, that contains a major portion of the data within the original text(s), which isn't any longer than half of the initial text(s) . Text summarization [1] is the method of distilling the foremost important data from a source (or sources) to provide a short version for a specific user (or users) and task (or tasks).

When this can be done by means of a pc, i.e. automatically, they call this Automatic Text summarization. Despite the actual fact that text summarization has historically been targeted on text input, the input to the summarization method also can be multi-media info, like pictures, video or audio, in addition as on-line info or hypertexts. Moreover, they will refer summarizing just one document or multiple ones. In this case, this method is understood as Multi-document summarization (MDS) [1] and also the source documents in this case are often in a very single-language (monolingual) or several languages (trans-lingual or multilingual).

Fig 1.1: Text Summarization

.

II. CLASSIFICATION OF TEXT SUMMARIZATION

Text summarization strategies are often classified into extractive and abstractive summarization [2]. An extractive summarization technique consists of choosing necessary sentences, paragraphs etc. from the original document and concatenating them into shorter kind. The importance of sentences is determined based on statistical and linguistic characteristics [2] of sentences.

An abstractive summarization [2] attempts to develop an understanding of the main concepts in every document and then specifies those ideas in clear natural language. It uses linguistic strategies to look at and interpret the text and so to search out the new concepts and expressions to best describe it by generating a brand new shorter text that conveys the most necessary info from the initial text document.

Extractive summaries [2] are developed by extracting key text segments (sentences or passages) from the text, based mostly on statistical analysis of individual or mixed surface level options like word/phrase frequency, location or cue words to find the sentences to be extracted. The "most important" content is treated as the "most frequent" or the "most favorably positioned" content. Such an approach therefore avoids any efforts on deep text understanding. They're conceptually easy, simple to implement.

Extractive text summarization [2] methods are often divided into 2 steps:

1. Pre processing step and
2. Processing step.

Pre processing is structured illustration of the initial text. It usually includes:

a) Sentences boundary identification [2]:- In English, sentence boundary is known with presence of dot at the end of sentence.
b) Stop-Word Elimination [2]:- Common words with no semantics and that don't combine relevant info to the task is eliminated.
c) Stemming [2]:- The purpose of stemming is to get the stem or base form of every word that emphasize its semantics.

Prime hierarchal sentences are elected for final summary.

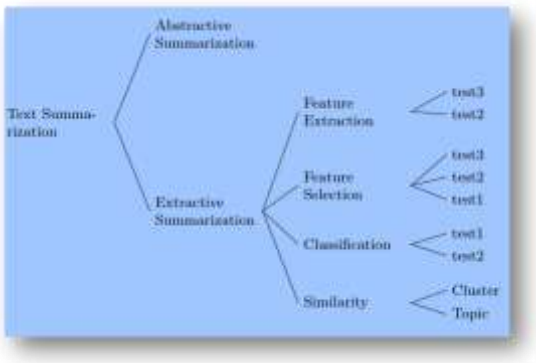

Fig 2.1: Classification of Text Summarization

Importance and relevance of study:

Text summarization [1] is the method of creating a condensed version of original document. This condensed version ought to have vital content of the initial document. Analysis is being done since many years to get coherent and indicative summaries [3] using totally different techniques. As Per (Jones, 1993) the text summarization is represented as 2 step method

(i) Building a source representation from the initial document.

(ii) Generating summary from the source representation

Text summarization is generally classified into 2 types: Single document summarization [8] and multi-document summarization [4]. This paper focuses on single document summarization that generates summary of single document. The text summarization is classified into extractive and abstractive depending on the nature of text illustration within the summary, detail is defined in previous chapter. We used extractive summarization in our proposed work.

Automatic text summarization (ATS) [3] is considered as method of reducing a text document with a computer program so as to form a summary that retains the main or important details of the initial document. As the drawback of info overload has increased, and because the amount of data has risen, therefore has interest in automatic summarization [3] .Technologies which will create a coherent summary take into consideration variables like length, writing style and syntax. Automatic data summarization may be a vital area among machine learning and data mining. Summarization technologies are used nowadays, in a very large number of sectors in business nowadays. An example of the usage of summarization technology is search engines like Google. An alternative example comprises document summarization, image collection summarization and video summarization. The main concept of summarization is to search out a representative subset of the data, that contains the information of the whole set. Document summarization, tries to automatically produce a representative summary or abstract of the whole document, by finding the important and main informative sentences. Similarly, in image summarization the system finds the important and main representative and vital (or salient) pictures. Similarly, in consumer videos one would need to get rid of the boring or repetitive scenes, and extract out a far shorter and compact version of the video. This can also used say for investigation videos, where one would possibly need to extract out only necessary events from the recorded video, since most of the events are uninteresting with nothing going on.

Due to substantial increase in the quantity of info on the web, it's become very troublesome to go looking for relevant documents required by the users to resolve this drawback, Text summarization is employed which produces the summary of documents in a way that the summary contains vital content of the document. Lexical chains [8] are created via WordNet. The score of every Lexical chain is calculated supported keyword strength & alternative features. The main concept of implementing lexical chains helps to research the document semantically and therefore the concept of correlation of sentences helps to think about the relation of sentence with preceding or succeeding sentence. This improves the standard of summary generated [3].

Berzilay & Elhada[5] given an improved algorithmic rule that constructs all possible interpretations of the source text using lexical chains. It's an efficient methodology for text summarization as lexical chains establish and capture necessary ideas of the document while not going into deep semantic analyses. Lexical chains are made using some knowledge base that contains nouns and its numerous associations.

Next merge chains between segments that contain a word within the same sense in common. The algorithm then calculates score of lexical chains, determines the strongest chain and uses this to get a summary. They conjointly used the idea of correlation of sentences to get a decent quality summary. The terms that occur within the strongest lexical chains are thought of as key terms and also the score of sentences is calculated on the basis of the presence of key terms in it. All the sentences are graded on the basis of their score and top n sentences are chosen for inclusion within the summary. Then the correlation of sentences is checked and if any sentence has correlation with the previous sentence, then the previous sentence must be enclosed within the summary based on condition. From this paper we have inspired by the concept of WordNet, a library of the words and their senses. We have then used in our dissertation the concept of WordNet for finding the base forms.

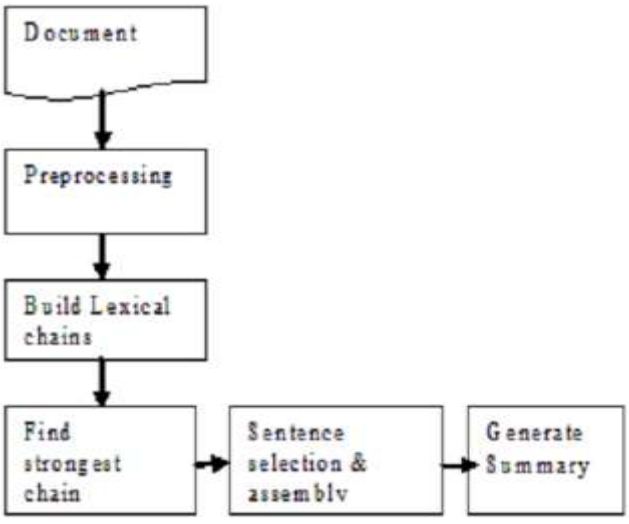

Fig 2.2 Architecture of Text Summarization

Lexical Chains can also be considered as the method in which certain words or the grammatical features of the sentence can connect it to its predecessors and the successors in the text. Apart from that the Cohesion occurs when the interpretation of some element in the discourse is depends upon the interpretation of other elements of the text [9]. Cohesion is also considered as a device for "sticking together" the different portions or sections of the text. And it is attained through the use of the semantically related terms, the references, the ellipsis and also the conjunctions in the text. Among these of the types, lexical cohesion, which is then created by making use of the semantically related words, is the most frequent one [9]. Lexical cohesion can also be classified into the reiteration category and the collocation category. And the reiteration occurs when one of the lexical items recalls the meaning of an earlier item in the text. And it can be obtained by using repetition, synonyms and hyponyms. Collocation refers to the words on which tend to co-occur in the text. And Lexical cohesion can also occur not only between the two terms but also within the sequences of the related words. And then these sequences of the words are known as lexical chains. Lexical chains can be distributed over sentences and different text parts. Words may be grouped in the same lexical chain when: [10]

• Two noun instances are identical and are used in the same sense. (The house on the wood is large. The house is made of wood.)

• Two noun instances are used in the same sense (i.e., synonyms). (The car is fast. My automobile is faster.)

• The senses of two noun instances have the hypernym/hyponym relation in between them. (John owns a car. It is a Toyota.)

• The senses of two noun instances are siblings in the hypernym/hyponym tree. (The truck is fast. The car is faster.)

Barzilay and Elhadad used lexical chains first [5], as an intermediate step in the text summarization process to extract important concepts from a document. They showed that cohesion is one of the surface signs of discourse structure and lexical chains can be used to identify it. They relied on WordNet [6] to provide sense possibilities for word instances as well as semantic relations among them. Senses in the WordNet database are represented relationally by synonym sets (synsets) are considered as the collection of all the words sharing a common sense. Words of the same category are linked through semantic relations like synonymy and hyponymy. Lexical chains were constructed in three steps:

(i) Select a set of candidate words
(ii) For each candidate word, find an appropriate chain according to a relatedness criterion
(iii) If such a chain is found, insert the word into the chain and update the chain

Once the chains were constructed, they showed that picking the ideas which is presented by the strong lexical chains gives a far better understanding of the central topic of a text than picking only the most frequent words in the text. Finally they used these strong chains to extract sentences from the original text to construct a summary. After Barzilay and Elhadad, many researchers followed this approach to use lexical chains in text summarization. Silber-McCoy proposed a new algorithm to compute lexical chains that was based on Barzilay-Elhadad method but was linear in space and time. Since the method proposed in had exponential complexity [5], it was hard to compute lexical chains for large documents. For this purpose, Silber and McCoy recompiled the WordNet noun database into a binary format and memory-mapped it. Then, they created "meta chains" that represent every possible representation of the text. These "meta chains" were used to disambiguate word senses and to create the lexical chains. Since WordNet was recompiled into a new format, it could be accessed as a large array and this allowed the algorithm to compute the lexical chains in linear time. After the chains were computed, the strong chains were selected, and summary sentences were extracted like did.

The increased in the growth of the net has resulted in huge amounts of information that has become tougher to access with efficiency. Web users need tools to manage this immense amount of information. The main goal of this analysis is to form an economical and effective tool that's able to summarize quite large documents quickly. This analysis presents a linear time algorithmic rule [10] for finding out lexical chains that could be a technique of capturing the "aboutness" of a document. This technique is compared to previous, less efficient strategies of lexical chain extraction. They additionally give different strategies for extracting and evaluation lexical chains. They show that their technique provides similar results to previous analysis, however is considerably quite more efficient. This efficiency is important in web search applications where several quite large documents might have to be summarized promptly, and where the reaction time to the end user is very vital.

This initial part of their implementation constructs an array of "meta chains" [10]. Every Meta chain contains a score and a data structure that encapsulates the meta-chain. The score is computed as every word is inserted into the chain. Whereas the implementation creates a flat illustration of the source text, all interpretations of the source text are implicit among the structure. Every line represents a semantic association [10] between 2 word senses. Every set of connected dots and lines represents a meta-chain. The gray ovals represent the list of chains to that a word will belong. The dashed box indicates the strongest chain in their illustration show in figure 2.2.

Notice that in some senses of the word machine, it's semantically like friend, whereas in different senses, it's semantically like computer (i.e. within the same meta-chain). The algorithmic rule continues by making an attempt to search out the "best" interpretation from among their flat illustration. They consider the illustration as a group of transitively closed graphs whose vertices are shared. In figure, the sets of lines and dots represent five such graphs. The set of dots among an oval represent a single shared node. That's to mention, that whereas two of those graphs could share a node, the individual graphs aren't connected. The " best" interpretation are going to be the set of graphs that may be created from the initial set mentioned above, by deleting nodes from every of the graphs in order that no two graphs share a node, and also the overall " score" [10] of all the meta-chains is largest.

Form this paper, we have learned and inspired by the concept of the lexical chains, and how they are created and applied in the field of the text summarization. We have also learned the concept of how to score the chain and find the usability of the chains for text summarization. We have also understand, how to use the WordNet, a library of the words and their senses.

They investigate one technique to supply a summary of an original text while not requiring its full semantic interpretation [11], however instead hoping on a model of the topic progression within the text derived from lexical chains. They present a new algorithmic program to find out lexical chains in a text, merging many robust knowledge sources: the WordNet thesaurus, a part-of-speech tagger, shallow parser for the identification of nominal teams, and a segmentation algorithmic program. Summarization is carried out in four steps: the initial step is, text is segmented, lexical chains are made, strong chains are marked or identified and vital sentences are extracted.

Text summarization is among one application of natural language processing and is now becoming much common for info condensation. Text summarization could be a method of reducing the size of original document and results a summary by holding necessary info of original document. This paper provides comparative study of varied text summarization strategies based on differing kinds of application. The paper discusses well two main classes of text summarization strategies these are extractive and abstractive summarization strategies [12]. The paper conjointly presents taxonomy of summarization systems and statistical and linguistic approaches [12] for summarization.

III. PROPOSED ALGORITHM

Proposed methodology: Automatic text summarization using lexical chains.

Step 1: Input

Input Original document for generating summary (.txt file).

Step 2: Segmentation

Divide the document into sentences using segmentation [13].

Step 3: Tokenization

Each sentence is divided into tokens i.e. an example:

Friends, has been colouring and roman lend me, your field;

Hence after tokenization we get: Friends has been colouring and roman lend me your field.

Basically we need to omit the commas, punctuations, (carefully apostrophes), question marks etc [14].

Step 4: POS tagging

The pos tagging is the tagger which specify the token as nouns, verbs, adverbs, adjectives [15].

Step 5: Nouns and compound nouns filtering

In this we need to filter the (nouns and compound nouns ex: computer-science) should be extracted.

Step 6: Word senses of nouns and compound nouns

In this we need identify all the senses of nouns and compound nouns and get more appropriate sense regarding the document. It will be done by WordNet dictionary.

Step 7: Collection of candidate words:

After getting an exact meaning of a noun like: PLANT: A living thing (tree) or industrial plants or anything else regarding that text. After ensuring the sense of each noun we collect the nouns and compound nouns as CANDIDATE WORDS. (Count frequency of words)

Step 8: WordNet dictionary

The WorldNet dictionary is most important part of our research. In this dictionary we need to run some java API's that are available on net. By them we need to identify the semantic relations among the candidate's words.

There are some relations that have been provided by WordNet:

Identity relation: the relation among same synset (set of synonyms) like: Red and red are same.

Synonym relation: The two words should be synonym to each other like: Intelligent and brilliant should be present in same synset or car and vehicle in same synset.

Hypernyms/ hyponyms: There are some relations like: hyponyms are like: oak is a hyponyms of tree or dog is a hyponyms of animal and the vice versa (opposite of hyponyms is hypernyms).

Meronyms: Is a part of whole: like "finger" is a part of "hand" or "memory" is a part of "computer".

There are some more relations but we consider these four relations.

Hence these are the four types of relations we need to identify among the each candidate words.

Step 9: Lexical chains

After finding the relations we need to make lexical chains of candidate words for that document. We will have a no. of lexical chains. Changes are applied only two relations identity and synonyms but we will work on four relations for making chains.

Step 10: Scoring the chains

After getting the lexical chains we need to score them up to utility of the chains, related formulas describe in reference paper [3]. In that paper they use global set for many documents but we apply a single document and for significance of lexical chains we need to compare a single chain by all other chains for that document. And then find out the utility formula. Then we will calculate threshold value.

Step 10: Accepting the chains

If utility is greater than threshold value, then those chains are accepted. We will find out the words which are presented in the accepted chains.

Step11: Generate summary

Then we will compare in our original document, each line contains how many words from the group of accepted words. Then sort the lines on the basis of number of accepted words contained in the line. Then we will extract the percentage of summary lines.

Step 12: Evaluation--After getting summary we need to evaluate by using Recall Method.

Formula Applied in the Solution:

1. Significance of the chain

For each chain find the significance of each chain

$$sig(L) = -\frac{length\ (L)}{\sum_{LEC} Lenth\ (l)} . \log_2 \frac{length\ (L)}{\sum_{LEC} Lenth\ (l)}$$

Fig 3.8 Formula for Significance of the chain

Where length (L) is the length of a particular chain L. And Length (l) is the sum of all chains length in the text. In this way each chain has its significance.

2. Formula for Utility of the Chain

The utility of a lexical chain L to a document D is defined as

$$util(L,D) = sig(L).length(L)$$

Fig 3.9 Formula for Utility of the Chain

3. Computing the Threshold value

This formula is used for computing the accepted chains and the formula is as follows ,

$$\text{Threshold} = \sum_{k=0} \text{util}(L, D) / (\text{TotalChains} * 2)$$

Fig 3.10 Formula for Threshold

4. Computing the recall

This formula is used to calculate the percentage of match from the human generated summary and our algorithm generated summary

RECALL = (TOTAL_WORDS_MATCHED_IN_HUMAN_SUMMARY/TOTALWORDS);

5. Computing the time difference

Here we will take the start time which is the time the process started and end time when the process stopped or ended after generated recall and summary.

Total_Seconds=End_Time-Start_Time.

IV. Evaluation and Experimental

We have run our program on around 40 sample summaries and from those we have presented around 4 samples and the result of the comparison is presented in the form of the graph , we have taken documents with name SampleData.txt,SampleData2.txt,SampleData3.txt and SampleData4.txt.

SAMPLE DATA 1: Input file SampleData.txt:

Most San Francisco-area homeowners may have to pay for damage from Tuesday's earthquake out of their own pockets, while insurance companies may reap long-term benefits from higher rates, industry spokesmen and analysts said Wednesday.

Only 15 percent to 20 percent of California homeowners have earthquake insurance, which typically requires a 10 percent deductible and costs between $200 to $400 a year for a $100,000 home, according to industry spokesmen.

The Association of California Insurance Cos. in Sacramento said that in the San Francisco area roughly 25 to 30 percent of the homes have earthquake insurance.

The organization estimated residential damages from Tuesday's earthquake at $500 million in the Bay area, with between $100 million to $150 million insured. Insured homeowners without earthquake protection will get reimbursed only if their homes were ravaged by fire, which is covered under basic homeowner insurance polices, said Hugh Strawn, director of catastrophe services at the Property Loss Research Bureau in Schaumburg, Ill.

Insurance companies attempted Wednesday to assess the amount of quake-related damages they're likely to have to pay.

In addition to home damage, the companies likely will get claims for automobile damage, broken glass, theft and burglary, business interruption due to electrical outages, water damage and, possibly, workers compensation.

Some estimated that insurers might face bills totaling $1 billion or more from the quake.

But industry observers said they don't expect any company to suffer serious financial damage from quake-related claims. ``We don't think any company is going to have problems paying claims," said Elisa Siegal, public affairs manager for the American Insurance Association, a Washington-based trade group.

The insurers actually could benefit. Industry analysts predicted insurers would be able to reverse three years of declining rates and win rate hikes from state regulators due to the quake damages and the estimated $4 billion in damages from Hurricane Hugo, which hammered South Carolina and other parts of the southeastern United States earlier this month. An increase in insurance rates could translate into greater profitability in the long term, the analysts said. ``There's a perception that this could turn the cycle ... that this could be enough to firm pricing," said Gloria L. Vogel, an analyst with Bear, Stearns & Co. Inc. in New York.

Despite their predictions for the long run, the analysts warned that fourth-quarter earnings among insurance companies are likely to be disappointing. ``It's going to be a meaningful loss, perhaps as big as Hugo," said Robert Glasspiegel, an analyst with Hartford-based Langen McAlenney.

On the New York Stock Exchange, some insurance company stock rose on Wednesday.

Aetna Life & Casualty Co. rose $2.37 to $59.50 a share; ITT Corp., parent of The Hartford, was up 37 cents to $59; and the Travelers Cos. rose $1 to $40.87.

Reinsurance companies, which absorb risk from policy writers, did especially well…….etc

Human Generated Summary Original:

The Association of California Insurance Cos. in Sacramento said that in the San Francisco area roughly 25 to 30 percent of the homes have earthquake insurance. Under a 1985 California law insurers are required to offer earthquake insurance to

homebuyers but homebuyers are not required to buy the coverage. He is estimating this week's disaster will generate insured losses of $2 billion to $4 billion following about $4 billion in costs to insurers from Hurricane Hugo . They expect to have a preliminary estimate of the damages in a day or two. The governor is wrong however in his campaign to distance himself from the California Department of Transportation on the issue of what caused the Nimitz Freeway in Oakland to collapse and what could or should have been done to have prevented it.

Our Generated BasePaper Summary:

Most San Francisco-area homeowners may have to pay for damage from Tuesday's earthquake out of their own pockets while insurance companies may reap long-term benefits from higher rates industry spokesmen and analysts said Wednesday. The Association of California Insurance Cos. in Sacramento said that in the San Francisco area roughly 25 to 30 percent of the homes have earthquake insurance..Insurance companies attempted Wednesday to assess the amount of quake-related damages they're likely to have to pay.In addition to home damage the companies likely will get claims for automobile damage broken glass theft and burglary business interruption due to electrical outages water damage and possibly workers compensation.Some estimated that insurers might face bills totaling $1 billion or more from the quake.

Our Generated Proposed Work Summary:

Most San Francisco-area homeowners may have to pay for damage from Tuesday's earthquake out of their own pockets while insurance companies may reap long-term benefits from higher rates industry spokesmen and analysts said Wednesday. The Association of California Insurance Cos. in Sacramento said that in the San Francisco area roughly 25 to 30 percent of the homes have earthquake insurance public affairs manager for the American Insurance Association a Washington-based trade group. That. In New York. Insurers aren't required to offer earthquake insurance to commercial property owners but the percentage of business property with the coverage is very high industry spokesmen said.

We used 4 Sample Data to generate base paper & proposed work base summary & compare to human generated summary and find the value of recall, total words matched & time taken to generate summary. We used the 30% of lexical chains for summary of sample data. We also show the result in below.

S.NO	Input File	Recall		Total Words Matched		Time Taken to generaete summary	
		Basepaper	Proposed work	Base paper	Proposed work	Base papa	Proposed work
1							
2							
3							
4							

Table 5.1 Results of text summary generation

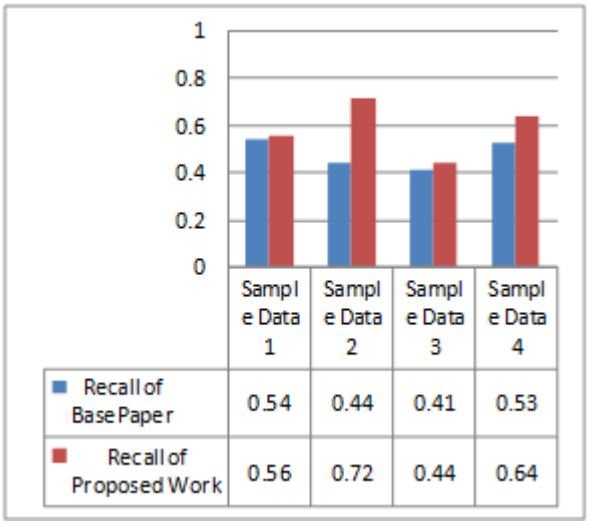

Fig 5.2 Show the recall value of Summary generation

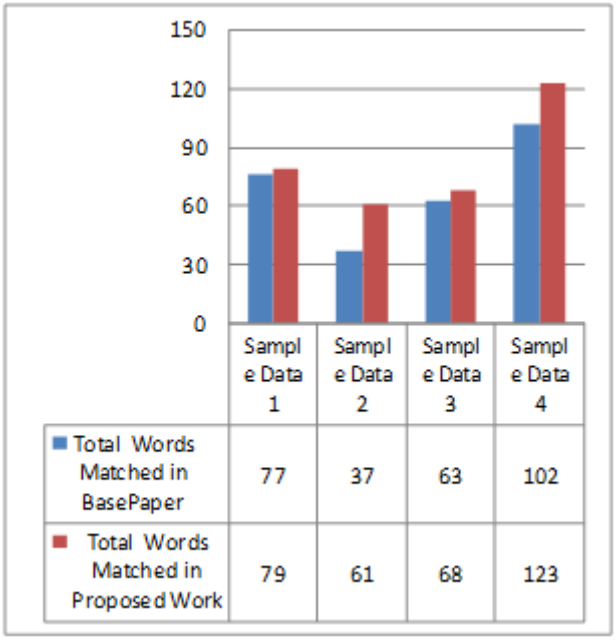

Fig 6.2 Show the Total words matched in Summary generation

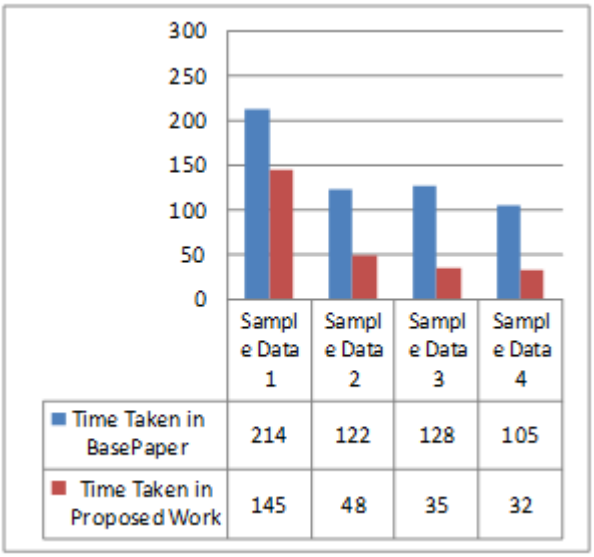

Fig 5.4 Show the Time taken in summary generation

V. CONCLUSION

The document summarization problem is a very important problem due to its impact on the information retrieval methods as well as on the efficiency of the decision making processes, and particularly in the age of Big Data Analysis. Though a good kind of text summarization techniques and algorithms are developed there's a requirement for developing new approaches to supply precise and reliable document summaries that may tolerate variations in document characteristics.

In this thesis, we presented a method to find out the lexical chains as an efficient intermediate representation of our document. Along with WordNet API, our method also included the nouns and proper nouns in the computation of lexical chains. And the statistical calculations in our proposed methodology resulted in the better output as compared to the base paper.

.

REFERENCES

[1] "Text Summarization: An Overview" ,Elena Lloret, paper supported by the spanish government under the project TEXT-MESS (TIN2006-15265-C06-01), 2008

[2] "A Survey of Text Summarization Extractive Techniques", Vishal Gupta, *Gurpreet Singh Lehal*, *Journal of Emerging Technologies in Web Intelligence*, Vol 2, No 3 (2010), 258-268, Aug 2010, doi:10.4304/jetwi.2.3.258-268

[3] "An Automatic Text Summarization Using Lexical Cohesion and Correlation of Sentences" ,A.R.Kulkarni, S.S.Apte,2014, IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308.

[4] "A Pandect of Different Text Summarization Techniques" ,A. N. Gulati, Dr. S. D. Sawarkar, Datta Meghe College of Engineering, Airoli, Navi Mumbai, Maharashtra, India", International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 4, April 2015 ISSN: 2277 128X .

[5] Barzilay, R. and M. Elhadad, "Using Lexical Chains for Text Summarization", ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization, pp. 10–17, 1997.

[6] "WORDNET", http://wordnet.princeton.edu/.

[7] "Corpus based Automatic Text Summarization System with HMM Tagger",M.Suneetha, S. Sameen Fatima, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-3, July 2011.

[8] "A Comparative Study of Stemming Algorithms", Dr. Anjali Ganesh Jivani, 2011, Journal Int. J. Comp. Tech. Appl Volume 2, Issue 6, Pages 1930-1938.

[9] Barzilay, R., "Lexical Chains for Summarization", M.Sc. thesis, Ben-Gurion University of the Negev, Department of Mathematics & Computer Science, 1997.

[10] Silber, H. G. and K. F. McCoy, "Efficient Text Summarization Using Lexical Chains", Proceedings of the 5th international conference on Intelligent user interfaces, pp. 252–255, January 2000

[11] "Using Domain Knowledge for Text, Summarization in Medical Domain",Kamal Sarkar, International Journal of Recent Trends in Engineering, Vol 1, No. 1, May 2009

[12] "Survey on Extractive Text Summarization Approaches", M S Patil, M S Bewoor, S H Patil , NCI2 TM: 2014 ISBN: 978-81-927230-0-6

[13] DILEK HAKKANI-TUR¨ , SLAV PETROV, DAN KLEIN "EFFICIENT SENTENCE SEGMENTATION USING SYNTACTIC FEATURES", BENOIT FAVRE, PUBLISHED IN: SPOKEN LANGUAGE TECHNOLOGY WORKSHOP, 2008. SLT 2008. IEEE , DATE OF CONFERENCE: 15-19 DEC. 2008 PAGE(S): 77 – 80 E-ISBN : 978-1-4244-3472-5 CONFERENCE LOCATION : GOA DOI: 10.1109/SLT.2008.4777844.

[14] RM Kaplan, "A method for tokenizing text" - Inquiries into words, constraints and contexts, 2005 - stanford.edu.

[15] A Practical Part-of-Speech Tagger Doug Cutting and Julian Kupiec and Jan Pedersen and Penelope Sibun Xerox Palo Alto Research Center 3333 Coyote Hill Road, Palo Alto, CA 94304, USA.

[16] Jayarajan, Dinakar and Deodhare, Dipti and Ravindran, Balaraman (2008) "*Lexical Chains as Document Features"*. In: Proceedings of the Third International Joint Conference on Natural Language processing (IJCNLP2008).