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

Automated Text Summarization Using RR and PST Method

[1.]K. Divya, [2.] Mr. Anil Patidar
[1,2.] Department of Computer Science & Engineering,
Medi-Caps Indore

***Abstract*— The most common use of Natural Language processing is for the summary generation. The denser representation of documents is called a summary. The automated text summarization is the process in which the input is to the computer is the text, whereas the output is the concise extract of the input data. The entire process of automatic text summarisation takes four stages. They are tokenization, feature identification, characterization, tagging and summarization. The work can be used in much practical application. Most of the data on the web today don't have summaries. System as well as for human the job of summarization is challenging task. As a result of which the task of automating i.e. to get an indicative and informative summaries became researcher's main concern. The process of automated text summarization constructs a summary automatically for some text.**

Keywords— Tokenization, POS Tagging, WordNet, and Lucene.

I. INTRODUCTION

Natural Language Processing is the study of the analysis of the human language i.e. the spoken language [1]. There are two main tasks of Natural Language Processing. The primary task deals with the computer system that performs necessary as well as interesting tasks with human languages. Secondly, Natural Language Processing is concerned with the better understanding of human language. Natural Language Processing (NLP) comprises of two main components. They are as follow[2].

- Natural Language Understanding : The natural language understanding identifies the input data i.e. is the natural language(human language).This phase comprises of different analysis levels They are : morphological, syntactical, semantic, and discourse analysis
- Natural Language Generation: The Natural Language Generation generates the output of the input data (human language).

To create a summary the researchers have given three different stages which when performed gives a summary. The increased data on the web needs to be summarized. But due to large amount of data, the process of summarization has become very complex. So to make the process easy and to get the summary these stages are performed. Hence the summarization is the combination of topic identification, topic interpretation, and summary generation. The brief description of the above stages is discussed as below:

- Topic Identification: With the amount of textual data that is available and exponentially increasing there is a need to automatically process the same. One way of doing this is by Topic Identification, which is the process of assigning one or more labels to text. These labels are chosen from a pre-defined list of topics. Topic Identification has various applications such as document categorization, e-mail routing and so on.
- Interpretation or Topic Fusion: In the process of interpretation the topics considered as important are coupled together. These fused topics are represented in the form given by new formulations such that without using word that is in the original text. A system cannot perform interpretation without having the prior knowledge of the domain. The system analyzes the input text in such a manner that is irrelevant to the text input.
- Summary Generation: The final and the most important stage of the process is the summary generation. The content of the summary can be generated by two ways. It could be either by abstracting and/or information extraction. To perform this task the system requires the techniques of natural language generation, namely sentence (micro-) planning, and sentence realization.

II. RELATED WORK

Automated text summarization is not a new research idea. It is been a research topic since last few years .Many techniques have been tried during 1950's and 60's.

A. Automated Text Summarization in SUMMARIST

One such similar pioneer work is by Eduard Hovy and Chin-Yew-Lin. He introduced a summarist, an attempt to create a robust automated text summarization system [3]; He gave an equation for the summarist system that is:

Summarization = Topic Identification +Interpretation + Generation

All the above three stages the SUMMARIST equation uses a combination of symbolic world knowledge and statistical or IR-based techniques. Each stages different and complementary methods. But before all these stages pre-processing stage takes place. Each module either performs certain pre-processing tasks (such as tokenization) or attaches additional features (such as part-of-speech tags) to the input texts.

B. Automated Text Categorization and Summarization using Rule Reduction

This paper suggests the summary of the document or data by using rule reduction [4].The text analyser parses the given input into tokens and recognizes the features of the alphabets and group them into Noun Phrase or Verb Phrase or Prepositional Phrase. The system generates some rules for the respective phrase. Generation of Rules: The rules are generated to identify the noun phrase, verb phrase, Prepositional phrase and adjective phrase.

The process of summarization takes palace through four stages. These are: Tokenization, Feature Identification, Categorization and Summary Generation. These are explained below:

1. Create Tokens for the given input.

In this step, the input is broken into three major categories of tokens namely alphabets, white spaces and punctuation symbols.

2. Recognize the feature of the created tokens

In this step, the features of alphabet tokens are identified namely as Determiner, Preposition, Noun, Verb, Adjective etc based on the rules defined in the text analyser.

3. Categorize the alpha tokens and summarize it to a sentence

In this step, depending on the rules, the analyzer categorizes the tokens into Noun Phrase, Possessive Pass, Prepositional Phrase or Verb Phrase based on its feature and then summarizes them to formulate a sentence PROPOSED WORK

III. PROPOSED WORK

The proposed solution is to first read the text document. Then the tokens of the sentence are created, as well as the stop words are removed. Once the stop words are removed, the outputs is stop word free token collection. These tokens are then named according to their behavior as noun, verb, adjective etc. These tokens contain some phases which are implemented with the addition of some assembly libraries or application to make the work easier. The phases are given and described below:

1. Tokenization
2. Tagging
3. Feature Identification
4. Weighting
5. Summary Generation

These stages are implemented using application which makes it easy to implement the stages and to get the better results.

A. WordNet

In WordNet the major relation among words is synonymy, for example, between the words shut and close or car and automobile [5]. The words that denote the same concept and can be interchanged in many contexts usually called as the synonyms are coupled into unordered set called synsets. The same part of speech is the main and large amounts of connection that are found in WordNet's relations. Hence, WordNet consists of four sub-nets, one for each, nouns, verbs, adjectives, and adverbs, respectively, each with few cross-POS pointers. Cross-POS relations consist of the "morphosemantic" links are those that are semantically similar words sharing the same meaning. For example: observe (verb), observant (adjective), observation (nouns), observatory (nouns). Some of the noun-verb pair examples can be considered to understand the semantic role of noun and verb respectively in the pair. For example: {car, van} where the car and van is the THING for {driving} and another example could be {chef} who is the agent whereas {cooking, food} are the RESULT.

B. Fitness Assessment

To create a summary of a given text data it is very important to identify the synonyms, antonyms etc. To identify them is a tedious job. The figure 1 shows a sample article which needs to be summarized

> Over the past thirty years, *research* in the health *arena* has attracted psychologists, anthropologists and the sociologists. The *focus* of psychological research in this *area* is concerned with the individual motives, attitudes and beliefs in the relation to both health and *illness*. Anthropological studies, however, are concerned with culture and health care. Such studies *concentrate* on a conception of *disease* as a culture product and on the way social and culture life in the past affect beliefs about health and illness in sociological *studies*, the *emphasis* is similar but focused more on social relation with in a particular social structure with respect to medical care.

Figure 1. Sample article to see the use of RiTaWN in the system

For the summarization of the above article the system needs to read the entire data to generate the summary. The WordNet helps to identify the words and their meaning but if there are words with similar meaning say as the synonym or opposite meaning i.e. antonym and so it is important to identify the words to generate a relevant summary. In the above figure 1, the article has the synonym word that needs to be rectified. To perform this task the system uses the application RiTaWN. The synonym words are:

Research → Studies
Focus → Concentrate and emphasis
Arena → Area
Illness → Disease

C. Stanford POS Tagger

Stanford part of speech tagger is the software that reads the text in some language and assigns part of speech to each token [6]. The sentences are broken into tokens and each token contains a word. Before the tagger it is important to know about the concept part of speech. The syntactic and the morphological behavior of the lexical item is defined by Part of speech in grammar, it shows the linguistic category of words. Mostly all languages have the lexical categories that are noun and verb, but there were significant variations in different languages. A simple example without ambiguous words/tokens is given below with simple part of speech tagging process.

Input: John hits the ball
Output: NN VB DT NN

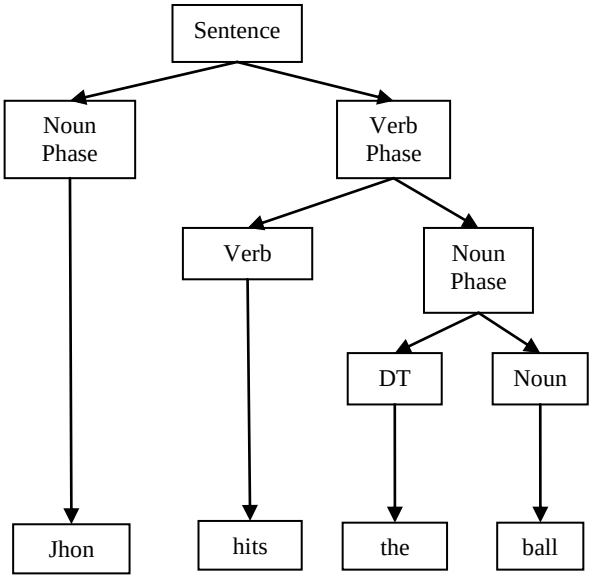

Figure 2. The Part Of Speech Tagging Process.

D. Lucene

Once the words are recognized and tagged with noun, verb, and adjective phrases. It's time to assign weight to the tagged words. After tagging the words the weights are assigned to the words. Each token i.e. word is assigned a particular weight. The weight assignment of the similar meaning words will be in such a manner that when the system comes across the word say "research" in a sample input document, it assigns '1' to it as weight but later when the system comes across the word "study" it increases the assigned the weight of the word "research" by one and now the weight of the word "research" becomes '2'. Similarly all the other synonym words are assigned weights. This is where Lucene is help full to find out the synonym words and assign them the weights as it is a full text search engine[7].To understand a simple example is considered which has two sentences. The example is given below in figure 3

> Over few years, *research* in natural language has increased widely. The *study* emphasizes on how natural language processed and the natural language is processing method.

Figure 3. The sample article with the use of Lucene in the system

The synonym word is:

Research ⟶ study

The weights are assigned in the manner given below in table 1

Table 1 The sample of tagging while using Lucene

Words	Weights
Over	1
Few	1
Years	1
,	0
Research, study	2
In	0
Natural	3
Language	3
Has	0
Increased	1
Widely	1
.	0
The	0
Emphasizes	1
On	0
How	0
Processed, processing	2
And	0
The	0
Is	0
Methods	1

E. Algorithm with Proposed Automated Text Summarization Method

1) An input string (n) is taken
2) The string is broken into tokens.
3) The tokens are read and the white spaces, stop words, repeated words and the expressions are removed.

4) A sequence of valid tokens is created.
5) Each token is rectified whether it is noun (NN), verb (VB), and adjective (Adj) and so on as well as the synonym, anonyms are also identified.
6) The tokens are then tagged with noun, verb, and adjective and so on respectively.
7) The tagged tokens are then assigned weights.
8) The weights are then calculated.
9) The sentences with the highest weight are included in the summary.
10) Summary is generated.

F. Flow Chart

To understand the process of a system, an easiest way is to create a flow chart of the system. A flow chart is the diagram that depicts the algorithm or the process of the system.

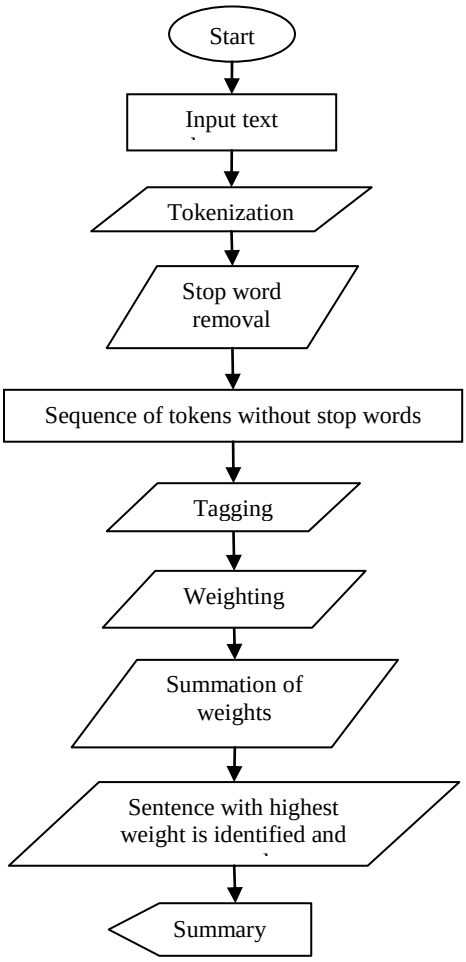

Figure 4. Flow chart of the proposed solution

IV. EXPERIMENTAL RESULTS

The summarizer created using specific application. Where never the text input is given to the summarizer, it will create the tokens, find the synonyms, tagged the token, identify the sentences, weight each token. At the end weight over the tokens in the sentence are summed up and the sentence having the highest sum value be included to create a summary.

A. Datasets

Three datasets are described in this section.

1) **Plain text:** The Plain text could be any article containing only data.

These example datasets are processed and result is shown below.

B. Plaintext

The sample plaintext is shown in the figure 5 below.

> Over the past thirty years, *research* in the health *arena* has attracted psychologists, anthropologists and the sociologists. The *focus* of psychological research in this *area* is concerned with the individual motives, attitudes and beliefs in the relation to both health and *illness*. Anthropological studies, however, are concerned with culture and health care. Such studies *concentrate* on a conception of *disease* as a culture product and on the way social and culture life in the past affect beliefs about health and illness in sociological *studies*, the *emphasis* is similar but focused more on social relation with in a particular social structure with respect to medical care.

Figure 5. The Sample Plaintext

The tagged and weighted words of the text document through the system is shown in figure 6

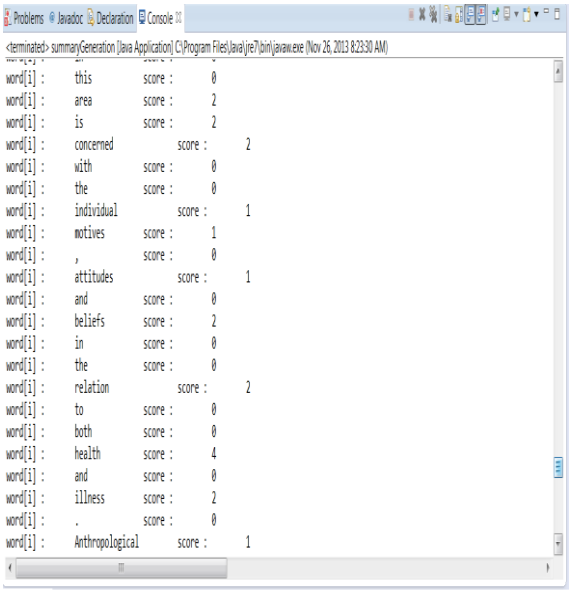

Figure 6. The tokens and the weights assigned to the text input to the system

Then the generated scores of the sentences and the summary of the input text document is given in figure 7

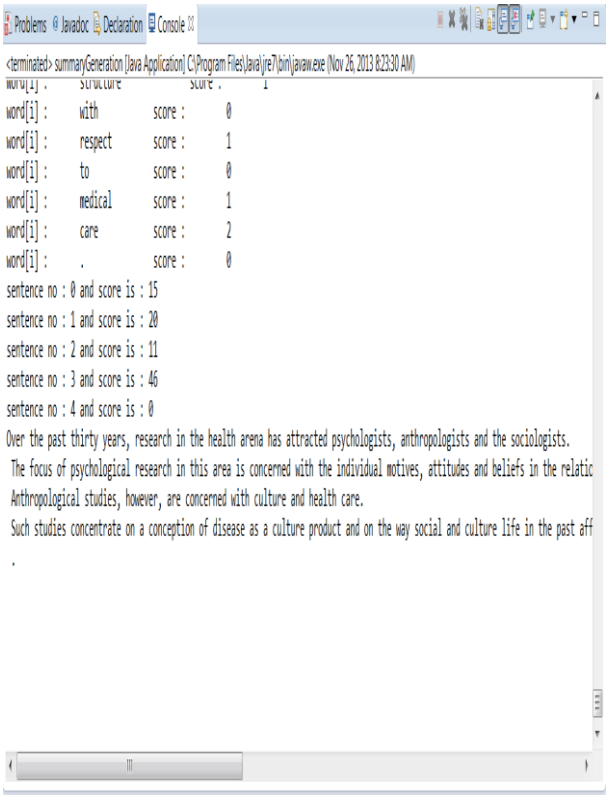

Figure 7. The scores assigned to the sentences and the summary of the text input to the system

V. CONCLUSION

In this paper concentrates on automated text summarization of the input text data. Thins this work the summarizer is developed to the granular structure of the input text data to analyze the semantically and syntactically the input text data. The summarizer performs few steps. They are tokenization, tagging, weighting and summarization.

REFERENCES

[1] K.R.Chowdhary."Natural Language Processing". K.R Chowdhary, *'Natural Language Processing'*, 2nd ed.,April 29 , 2012.

[2] Ted Briscoe, "Introduction to Linguistics for Natural Language Processing," *Computer Laboratory University of Cambridge*, October 4, 2011.

[3] Eduard Hovy and Chin-Yew Lin ,"Automated Text Summarization in SUMMARIST" In I. Mani and M. Maybury (eds), Advances in Automated Text Summarization. MIT Press

[4] C.Lakshmi Devasena[1] and M.Hemalatha[2] , " Automatic Text Categorization and Summarization using Rule Reduction",[1]Research Scholar ,[2]Professor , Department of Computer Science, IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012) Karpagam University , Coimbatore , India,March 30,31,2012

[5] The Princeton University website – WordNet software, (2012). http://www.wordnet.princeton.edu/

[6] The Part-Of-Speech [Online] Available: en. http://wikipedia.org/wiki / Part_of _speech./

[7] The LuceneTM Features[online] – Lucence software, (2012). http://www.lucene.apache.org/core/

[8] Eduard Hovy , "Text Summarization"

[9] M. A. Hearst. Untangling text data mining. In Proceedings of the 37th Meeting of the Association for Computational Linguistics (ACL-99), pages 3–10, College Park, MD, June 1999.Annual

[10] Christopher D.Manning, Prabhar Raghavan , Hinrich Schutze "An Introduction to Information Retrieval" Cambridge University Press Cambridge , England , Online edition (c) 2009.

[11] R. Feldman and I. Dagan, " *Kdt - knowledge discovery in texts*," In Proc. of the First Int. Conf. on Knowledge Discovery (KDD), 1995, pp

[12] PDF written by mMembers of The National Center for text mining and produced ND EDITED BY Judy Redfearn and the JISC Communications team

[13] Dipan Das AndreF.T.Marins "A Survey on Automatic Text Summarization" Language Technologies Institute Carnegie Mellon University, November 21,2007

[14] Elena Lloret ,"Text Summarizatoin : An Overview", Dept lenguajes y Sistemas Informaticos Universidad De Alicante , Spain. (TIN2006-15265-C06-01)