Automatic Text Summarization System Using Extraction Based Technique

[1]Priyanka Gonnade, [2]Disha Gupta

[1,2]Assistant Professor
[1]Department of Computer Science and Engineering, [2]Department of Computer Technology,
[1]Rajiv Gandhi College of Enggineering and Research, Nagpur,India
[2]Yeashwantrao Chavan College of Engineering, Nagpur,India
[1]priyanka.gonnade@gmail.com, [2]disha.g146@gmail.com

***Abstract*— In this new era, where tremendous information is available on the internet. It is most important to provide the improved mechanism to extract the information quickly and accurately. This leads to the difficulty for human beings to manually extract the summary of large documents of text. There is plenty of text material available on web. The difficulty is in looking for the exact document from the number of documents available, and extracting the required one. To solve all the above problems, automatic text summarization arises. Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings.**

***IndexTerms*—** Extractive summary, Abstractive summary, Scores, Text Mining, Ranking, Stop word, Stemming, Cue words

I. INTRODUCTION

Before going to the Text summarization, first we, have to know that what summary making is. A summary is a text that is produced from one or more texts, that follows important information in the original text, and it is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter version with semantics. The most important benefit of using a summary making, which it reduces the reading time. Text Summarization methods can be well-classified into extractive and abstractive summarization. The very first work on automatic text summarization by Luhn (1958) computes salient sentences based on word frequency (number of times a word occurs in a document) and phrase frequency. The subfield of summarization has been investigated by the NLP community for nearly the last five decades. Radev et al (2002) define a summary as “a text that is produced from one or more texts that gives important information in the original text and that is no longer than half of the original text and usually significantly less than that".

Extractive summarization:
An extractive summarization method includes selecting the most important sentences, facts ,paragraphs from the original document provided from user & concatenating them into shorter form that defines whole document.

Abstractive summarization:

An abstractive method forms an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human can think and elaborate.

II. RELATED WORK

The study of text summarization proposed an automatic summarization method combining conventional sentence extraction and trainable classifier based on Support Vector Machine. The study introduces a sentence segmentation process method to make the extraction unit smaller than the original sentence extraction. The evaluation results show that the system achieves closer to the human constructed summaries at 20% summary rate. On the other hand, the system needs to ameliorate readability of its summary output. In the study of proposed to generate synthetic summaries of input documents. These approaches, though similar to human summarization of texts, are limited in the sense that synthesizing the information requires modeling.

The various processes for Automatic Text Summarization are as follows:

A. **Interpretation** of the source text to obtain a text representation. To perform this stage, almost all systems employ several independent modules. Each module assigns a score to each unit of input (word, sentence, or longer passage); then a combination module combines the scores for each unit to assign a single integrated score to it; finally, the system returns the *n* highest-scoring units, according to the summary length requested by the user.

B. **Transformation** of the text representation into a summary representation. The stage of interpretations what differentiate extracts type summarization systems from abstract-type systems. During interpretation, the topics identified as important are coalesce, represented in new terms, and expressed using a new frame using concepts or words not found in the original document.

C. **Generation** of the summary text from the summary representation. The third major stage of summarization is generation. When the summary content has been created through abstracting and/or information extraction, it exists within the computer in internal notation and thus requires the techniques of natural language generation, namely text planning, sentence (micro-) planning and sentence realization.

III. TEXT SUMMARIZATION TECHNIQUES

To perform Text Summarization on input document to Generate Summary. In text automatic summarization system by giving input document to summarizer it will create a summary of input document by using various steps such as preprocessing, interpretation and summary generation.

A. Topic Identification

Preprocessing
The pre- processing is a primary step to load the text into the recommended system and make some processes such as case folding that transfers the text into the lower case state that improve the correctness of the system to differentiate same words. The pre-processing steps are as follows:-

Stop Word Removal
The procedure is to create a filter for those words that remove them from the text. Using the stop list has the advantage of reducing the size of the candidate keywords.

Word Tagging
Word tagging is the process of assigning P.O.S) like (noun, verb, and pronoun, Etc.) to each word in a sentence to give word class. The input to a tagging algorithm is a set of words in a natural language and specified tag to each. The first step in any tagging process is to look for the token in a lookup dictionary. The dictionary that created in the proposed system consists of 230,000 words in order to assign words to its right tag. The dictionary had partitioned into tables for each tag type (class) such as table for (noun, verb, Etc.) based on each P.O.S category. The system searches the tag of the word in the tables and selects the correct tag (if there alternatives) depending on
the tags of the previous and next words in the sentence.

Stemming
Removing suffixes by automatic means is an operation which is especially useful in keyword extraction and information retrieval. The proposed system employs the Porter stemming algorithm with some improvements on its rules for stem. Terms with a common stem will usually have similar meanings, for example: (join, joined, joining, joins) frequently, the performance of a keyword extraction system will be improved if term groups such as these are joined into a single term. This may be done by removal of the various suffixes -ED, -ING, -S to leave the single term joins. In addition, the suffix stripping process will lessen the number of terms in the system, and hence lessen the size and difficulty of the
data in the system, which is always beneficial.

Cue Words
The words which are in the double quotes should be considered as important for summary generation.

B. Topic Interpretation

Existence in the Document Title and Font Type

Existence in the document title and font type is another feature to gain more score for candidate keywords. Since the proposed system gives more weight to the words that exists in the document title because of its importance and indication of relevance. Capital letters and font type can show the importance of the word so the system takes this into account.

Parts of speech approach

After testing the keywords that extracted manually by the authors of articles in field computer science we noted that those keywords fill in one of the patterns. The proposed system improves this approach by discover a new set of patterns about that frequently used in computer science. This linguistic approach draws out the phrases match any of these patterns that used to extract the candidate keywords. These patterns are the most habitual patterns of the key words found when we do experiments.

Key Phrase weight calculation

The proposed system computes the weight for each candidate key phrase using all the features mentioned earlier. The weight represents the strength of the key phrase, the more weight value the more likely to be a good keyword (key phrase). We use these results of the extracted key phrases to be input to the next stage of the text summarization. The range of scores depends on the input text. The system selects N keywords with the highest. (See fig 3.2.1)

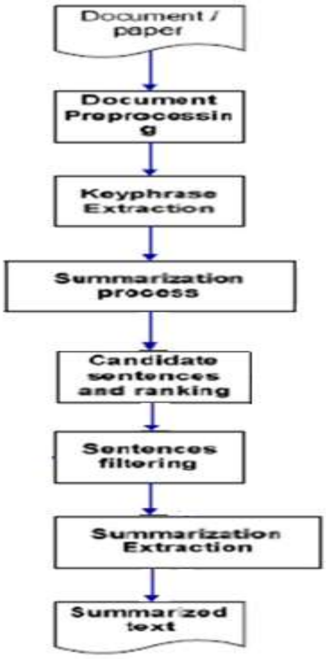


Figure 3.2.1 System Architecture

IV. PROPOSED SYSTEM

There are the following three main steps of summarizing documents that are topic identification, interpretation and summary generation. The most prominent information in the text is identified .There are different techniques for topic identification are used which are Position, Cue Phrases, word frequency. Methods which are based on the position of phrases are the most useful methods for topic identification. Abstract summaries need to go through interpretation step. In This step, different subjects are fused in order to form a general content. In this step, the system uses text generation method.

A. Sentences Selection Features

Sentence position in the document and in the paragraph.
Key phrase existence.
Existence of indicated words.
Sentence length.
Similar sentences of the Document class.

B. Existence of Headings Words:

Sentences occurring under certain headings are positively appropriate; topic sentences tend to occur early or late in a document.

C. Existence of Indicated Words:

By indicated words, we mean that the existence of information that helps to extract important statements. The following is a list of these words:Purpose: Information indicating whether the author's principle intent is to offer original research results, to evaluate the work performed by the others, to present a speculative or theoretical discussion. Methods: Information indicating the methods used in conducting the research. Such statement may refer to experimental procedures, mathematical techniques.

D. Sentence Length Cut-off feature:

Short sentences tend not to be included in summaries. Given a threshold; the feature is true for all sentences longer than the threshold.

V. RESEARCH METHODOLOGIES

Sparck Jones defines summary to be a condensed derivative of source text, i.e. a content trimming through either selection or generalization on what is important in the source document. It is a short version of document with only the significant information. The definition of the summary, though is an obvious one, highlights the fact that summarizing is in general a hard work, because we have to characterize the source text as a whole, we have to capture its most important content, where content is a matter of both information and its expression, and importance is a matter of what is necessary as well as what is important. In general, the functions of a summary include declaration that declare the existence of the original document.

Basic Steps:

1. Screening: determine the relativeness of the original document.
2. Substitution: replace the original document.
3. Retrospection: point to the original document.

Depending on the length and requirement of the summary some of these can be included while discarding others. Different kinds of summaries were identified based on the function of the summary. Indicative summaries provide an idea of what the text is about without conveying specific content and informative ones provide some shortened version of the content. Topic-oriented summaries concentrate on the reader's desired topic/topics of interest, whereas generic summaries reacts the author's point of view. Extracts are summaries created by picking portions (words, sentences, etc.) of the input text, while abstracts are created by regenerating the extracted content. Till now most of the researchers focused on producing extracts, with their concentration being on making the extract either indicative or informative. Indicative summaries usually serve the functions of announcement and screening. By contrast informative summaries are of function of substitution. Critical summaries criticize an approach or an opinion expressed in the text document. Extract can be of announcement and replacement. In general, all of the four types of summaries are retrospective. Indicative and informative summaries are the most important types in the current internet environment. Summaries are influenced by a broad range of factors.

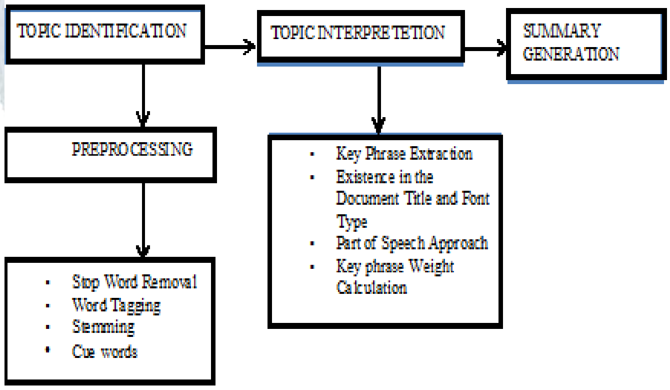


Fig. Summarization Methodology

VI. APPLICATIONS

A **Legal text** is something very different from simple speech. This is especially true of legal texts: those that create, modify, or terminate the rights and obligations of individuals or institutions. Such texts are what J.L. Austin might have called written per formatives. Lawyers often refer to them as operative or dispositive. Authoritative legal texts come in a variety of genres. They include documents such as: - Constitutions, contracts, deeds, orders/judgments/decrees, pleadings

Summarizing **Web Pages** have recently gained much attention from researchers. Until now two main types of approaches have been proposed for this task: content- and context-based methods. Both of them assume fixed content and characteristics of web documents without considering their dynamic nature.

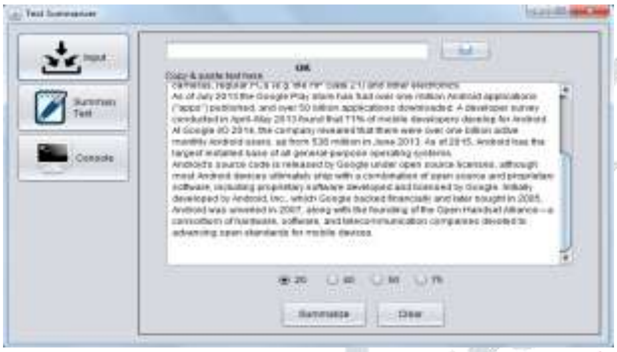
News Paper Reports can also be a example of Automatic Text Summarization System where the rush of data is converted into most important facts out of data junk derived from latest events and information.

Search Engines use this technique of Summarization using multi document summarization process for scanning multiple documents and conclude results.

VII.RESULTS AND DISCUSSION

In text summarization the text is given in the text box given in the home screen. Also text document can also be given by the browse button in the home screen

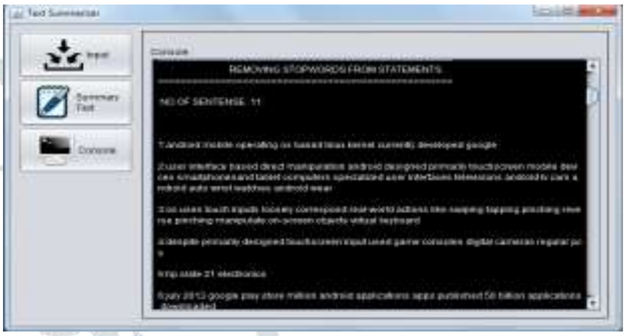
First of all, the input text is given through the "Input Button" to the system by pasting it in the text area or given by browse button. As the user press the "Summarize Button" the summary generation processes starts and the background process of the summary generation is shown in the "Console Button". Finally, the summary is displayed in the "Summary Text Button".The complete process of text summarization is given if following Fig. below



cameras, regular PCs (e.g. the HP Slate 21) and other electronics.
As of July 2013 the Google Play store has had over one million Android applications ("apps") published, and over 50 billion applications downloaded. A developer survey conducted in April–May 2013 found that 71% of mobile developers develop for Android. At Google I/O 2014, the company revealed that there were over one billion active monthly Android users, up from 538 million in June 2013. As of 2015, Android has the largest installed base of all general-purpose operating systems.
Android's source code is released by Google under open source licenses, although most Android devices ultimately ship with a combination of open source and proprietary software, including proprietary software developed and licensed by Google. Initially developed by Android, Inc., which Google backed financially and later bought in 2005, Android was unveiled in 2007, along with the founding of the Open Handset Alliance—a consortium of hardware, software, and telecommunication companies devoted to advancing open standards for mobile devices.

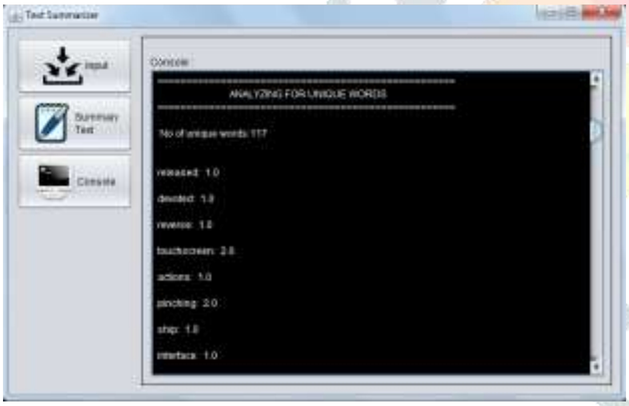
● 20 ○ 40 ○ 50 ○ 75

Fig 7.1. Input Screen



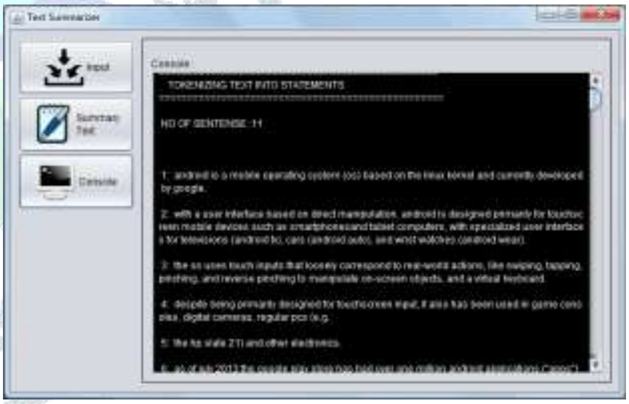
```
            REMOVING STOPWORDS FROM STATEMENTS
============================================================

NO OF SENTENSE :11


1:android mobile operating os based linux kernel currently developed google

2:user interface based direct manipulation android designed primarily touchscreen mobile devi
ces smartphones and tablet computers specialized user interfaces televisions android tv cars a
ndroid auto wrist watches android wear

3:os uses touch inputs loosely correspond real-world actions like swiping tapping pinching reve
rse pinching manipulate on-screen objects virtual keyboard

4:despite primarily designed touchscreen input used game consoles digital cameras regular pc
s

5:hp slate 21 electronics

6:july 2013 google play store million android applications apps published 50 billion applications
 downloaded
```

7.2. Seperation of Sentences



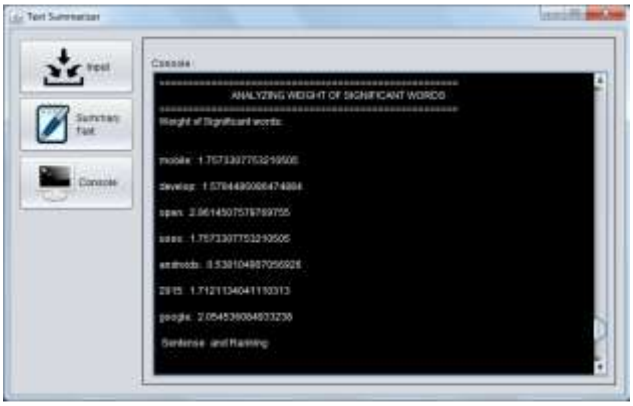
```
============================================================
              ANALYZING FOR UNIQUE WORDS
============================================================

No of unique words:117


released: 1.0

devoted: 1.0

reverse: 1.0

touchscreen: 2.0

actions: 1.0

pinching: 2.0

ship: 1.0

interface: 1.0
```

Fig 7.3 Removing of Stop Words



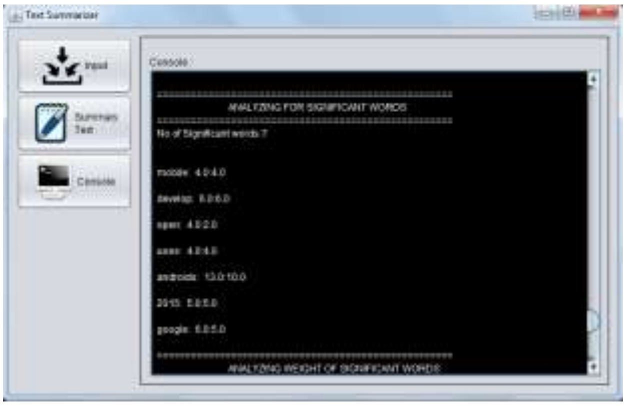
```
  TOKENIZING TEXT INTO STATEMENTS
============================================================

NO OF SENTENSE :11


1: android is a mobile operating system (os) based on the linux kernel and currently developed
by google.

2: with a user interface based on direct manipulation, android is designed primarily for touchsc
reen mobile devices such as smartphones and tablet computers, with specialized user interface
s for televisions (android tv), cars (android auto), and wrist watches (android wear).

3: the os uses touch inputs that loosely correspond to real-world actions, like swiping, tapping,
pinching, and reverse pinching to manipulate on-screen objects, and a virtual keyboard.

4: despite being primarily designed for touchscreen input, it also has been used in game cons
oles, digital cameras, regular pcs (e.g.

5: the hp slate 21) and other electronics.

6: as of july 2013 the google play store has had over one million android applications ("apps"
```

Fig.7.4 Analyzing Unique Words



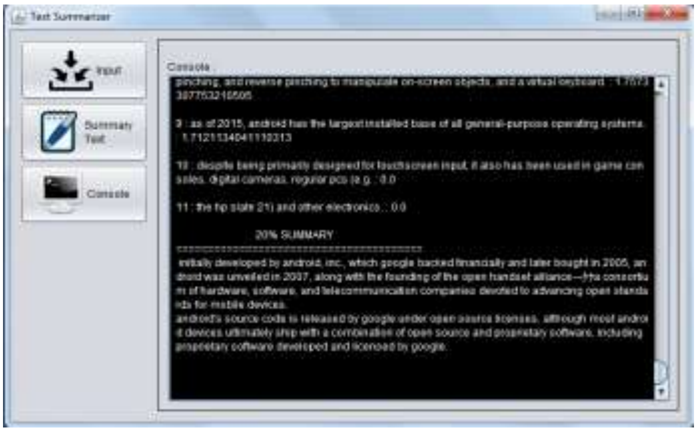
```
============================================================
          ANALYZING WEIGHT OF SIGNIFICANT WORDS
============================================================
Weight of Significant words:


mobile: 1.7573307753210505

develop: 1.5784480006474884

open: 2.0614507576789755

uses: 1.7573307753210505

androids: 0.5301048870569926

2015: 1.7121134041110313

google: 2.054536084033238

 Sentense and Ranking
```

Fig. 7.5 Key Phrase Extraction



```
============================================================
              ANALYZING FOR SIGNIFICANT WORDS
============================================================
No of Significant words:7


mobile: 4.0:4.0

develop: 6.0:6.0

open: 4.0:2.0

uses: 4.0:4.0

androids: 13.0:10.0

2015: 5.0:5.0

google: 6.0:5.0

============================================================
              ANALYZING WEIGHT OF SIGNIFICANT WORDS
```

Fig. 7.6 Key Phrase Weight Calculation



```
pinching, and reverse pinching to manipulate on-screen objects, and a virtual keyboard : 1.7573
387753216505

9 : as of 2015, android has the largest installed base of all general-purpose operating systems.
 1.7121134041110313

10 : despite being primarily designed for touchscreen input, it also has been used in game con
soles, digital cameras, regular pcs (e.g. : 0.0

11 : the hp slate 21) and other electronics. : 0.0

                    20% SUMMARY
==================================================
 initially developed by android, inc., which google backed financially and later bought in 2005, an
droid was unveiled in 2007, along with the founding of the open handset alliance—â€“a consortiu
m of hardware, software, and telecommunication companies devoted to advancing open standa
rds for mobile devices.
android's source code is released by google under open source licenses, although most androi
d devices ultimately ship with a combination of open source and proprietary software, including
proprietary software developed and licensed by google.
```

Fig. 7.7 Text Summary

References

[1] Rafeeq Al-Hashemi,"Text Summarization Extraction System (TSES) Using Extracted Keywords", pg.no.164, International Arab Journal of e-Technology, Vol. 1, No. 4, June 2010.

[2] Elena Lloret ,"Text Summarization Overview", Dept. Lenguajes y Sistemas Informaticos Universidad de Alicante, Spain (TIN2006-15265-C06-01).

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

[4] M.F Porter; "An algorithm for Suffix Stripping"; originally published *in \Program\, \14* no. 3, pp 130-137, July1980.

[5] Jagadeesh J, Prasad Pingali, Vasudeva Varma, "Sentence Extraction Based Single Document Summarization", Workshop on Document Summarization, 19th and 20th March, 2005, IIIT Allahabad Report No: IIIT/TR/2008/97.

[6] Md. Majharul Haque, Suraiya Pervin, and Zerina Begum, **"**Literature Review of Automatic Single Document Text Summarization Using NLP", ISSN 2028-9324 Vol. 3 No. 3 July 2013.