TextSumarization using TextRank

Anjali Sharma
19MCA0210
anjali.sharma2019@vitstudent.ac.in

Vishakha Singh
19MCA0231
vishakha.singh2019@vitstudent.ac.in

Akhya Omar
19MCA0205
akhya.omar2019@vitstudent.ac.in

Dr. Shynu P.G
Guide(SITE)
pgshynu@vit.ac.in

Abstract*— *A wide popularity has been enjoyed by the text summarization in the Natural Language Processing because of its potential in Information Access Software. A huge amount of data is present around us in unstructured form that makes irrelevance sense to the users. Text Summarization produce precise and fluent summary of the document while preserving the key points. In this review paper we are going to conceptualize the two approach of the Text Summarization which is abstractive Summarization and extractive Summarization. We identify important phrases and sentences from the original document and extract only those in Extractive Summarization. But in Abstractive Summarization we generate new sentences and text from the original one which act as the summary of the given text. In this paper we are going to study analyze and review various advancement in this field and we will use coding languages such as "java" for implementing text summarizing.

Keywords—extractive, abstractive, indicative, informative.

Introduction

In this new era, where huge data is accessible on the internet, it is most critical to give the improved instrument to separate the data rapidly and most proficiently. It is hard for people to physically separate the outline of vast reports of content. There are a lot of content materials accessible on the web. So there is an issue of looking for significant records from the quantity of reports accessible, and engrossing pertinent data from it. In request to take care of the over two issues, the programmed content synopsis is particularly necessary. Text summarizing is the way toward distinguishing the most significant important data in an archive or set of related records and packing them into a shorter form saving its general implications.

Text Summarization are classified on 3 basis:

On the basis of input: On the type of input there are 2 types of summary:

1. **Single Document:** In single document [6] summary data
   is quite small i.e.; data is in a single document..

2. **Multiple Document:** In this is the document is quite large i.e.; the data is not confined to single document, data is stored in multiple documents [6] and summary has to be prepared from all the documents.

On the basis of content: On the basis of content summary are classified under 3 categories:

1: Domain Specific: Domain-specific[5], where the model uses domain-specific knowledge to form a more accurate summary.

2: **Query based**: Query-based[5], where the summary only contains information which answers natural language questions about the input text.

3: **Generic:** Generic[5], where the model makes no presumptions about the topic or the content to be condensed and regards all inputs as homogeneous. Most of the work that has been done revolves around generic summarization.

On the basis of output: On the basis of output there are 2 types of summary:

1. **Extractive text summarization[4]:** In this summarizing no new text will be added only the text will be summarized.

2. **Abstractive text summarization[4]:** In this summarizing the whole original text is understood and then retells it in fewer words.

Problem Description:

1. How to select the most relevant information?
   (Optimizing topic coverage)

2. How to express it in Final Summary?
   (Optimizing readability)

Steps involved in text summarization:

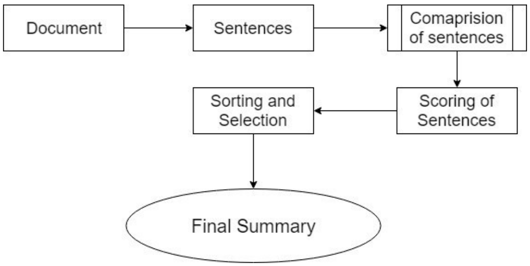


Figure 1: Process of text summarization[1]

There are 4 steps in the text summarization:

1. Document division
   First of all the document is divided into sentences.

2. Comparison of sentences
   After dividing the document into sentences these sentences are then compared with other sentences.
3. Scoring of sentences:
   After comparison sentences are assigned scores.
4. Selection and sorting of sentences:
   After scoring of sentences we have to select the most important sentence which is decided by various algorithms.
5. Creation of Final Summary:
   After all these steps final summary is created by merging all the selected statements.

RELATED WORK

Statistical based or Numerical based method make it easier for developing summaries that are extractive in nature. Numerical or statistical is in a trend or in use in a summarization method.

Based on title method.

Luhn[4] entitled that when the words are used more in any of the text paragraph it entitles or shows the great significance and it highlights the whole part of the paragraph text or the abstract idea. When the occurrence of word is more it is only referred as title.

Based on cue method

Edmondson[4] methods that is based on location, composition or title and sign or cue words. The composition or title will be signified through the means of quality number or words used. It takes the title where words are used in average number means that the word used more or less won’t be considered.

Based on rank method.

Goldstein[4] signified query based summarization directed at extracting necessary information that answers the query from the real text. Query is said to be that the data is represented in a relevant way but it is represented in a short and meaningful way. Whenever there is new article given or represented summarization used the linguistic characteristic which will be based on the rank or the number of the sentences.

Based on cluster and extraction method

ZHANG Pei-ying[4] & LI Cun[4] directed an approach based on collection or gathering said to be clustering and extraction. Clustering is a method of searching form or structure from where the document is represented and given, so among all the given documents , the same types of documents are grouped into categories. When the required data is collected or gathered together than it is extracted through extraction method. K-means algorithm is used in the cluster method.

Based on lexical chain method

Morris and Hirst introduced lexical chain concept. Generation of lexical chains id done on the basis of grouping of data or the words that are semantically related. The more lexical chains will be strong the good related data will be chosen as a summary.

Example - When any buyer purchases a novel or a book he goes through the summary of the book. And how well the matter is explained in precise form, on that basis he purchases the book.

Analysis of various algorithms:

Comparison of summarization methods

Types	Sub type	Concept	Advantages	Disadvantages
Approach	**Abstractive[4]**	It lessens a text or paragraph in order to generate summary of it.	Data is compressed on ratio. When text is reduced more it is related semantically.	Hard to culminate or calculate.
	Extractive[4]	Selects sentences of the higher priority from the real text document based on numerical characteristics.	Computation is easy because it is easy to deal with explanation and symbolic language.	Results are lengthy text. Unpleasant from consistencies.
Details	**Indicative[4]**	Only the real idea of the text is displayed to user. Quickly decide whether a text is worth reading or not.	Supports the users to read the main document in deep. Used for fast classification and easier to crop.	Well defined Data information not present
	Informative[4]	Gives Proper short summary.	Provides as a replacing for the whole document	Unable to serves fast overview
Content	**Generic[4]**	General represent of the summary regardless of any user. Information is equally important.	Any type of the user can use it.	Authors view is more important than the users view here.
	Query Based[4]	Determination of the original text is more important by the user, so it should be easy for the system to extract the required information.	The information searched based on the user choice. Reflects their area of interest.	It is used by similar type of user not by any type of user.
	Domain Dependent [4]	Short brief statement of required points of the document which their subject can be defined in the fixed domain	They are aware of the special domain on which they are dependent	Limited to the subject of the text document.

Methods	**Cue method [4]**	Set of cue phrases and sentences is a part of the source abstraction	Grammatically correct and less redundant summary	It is limited to single document
	Graph method [4]	The edges represent the semantic similarity and vertices represent the sentences. The vertices got the highest edges that sentences selected as summary.	The sentence scoring algorithm performs well in agglutinative languages	Less Semantics
	Rank method [4]	Each sentence is ranked. High scored sentences are selected as summary sentences	Simple method and easy to implement	It cannot handle synonymy and polysemy (a word with different or multiple meaning in different context).
	Lexical Chain[4]	Grouping of data or the words that are semantically related. The more lexical chains will be strong the good related data will be chosen as a summary.	It is good for semantically appropriate summary	Grammatical error will occur

PROPOSED APPROACH:

In this proposed approach[6] we are going to use extractive summary approach for summarization. For extracting summary from the data we follow these steps:

1. First all of divide the whole document into paragraphs and then divide these paragraph into sentences.
2. After this each and every sentence is compared with each other, this is done by counting the number of similar word and then dividing by average number of words per sentence to normalize it.
3. These values are stored in a intersection matrix.
4. Then store these values in a key value dictionary.
   Key: sentence
   Value: Sum of intersection values
5. From these paragraphs, find out the highest score.
6. Then sort the selected sentences as to preserve the meaning of the original document.

And at last the final summary is copied to output document and our system will also calculate the efficiency of the summary by calculating the total words in the original document and in our summary document and also find the percentage of compression.

METHODOLOGY

In the extractive summarization, the summarizer takes input as word file or text file and the document is then split into tokens as to remove the redundant text and also to remove the stop word.

Step 1: weights are assigned to them. The term weight is calculated as follows:

$$wt = \frac{Frequencyofterm}{Totalno.ofterms \in maindocument}$$

Step 2: The weighted frequency of the document is calculated by this:

$$wtf = \frac{Frequencyoft\,h\,eterms}{Maximumfrequencyoft\,h\,eterms}$$

Step 3: Sentences are arranged in descending order of frequency and then from these paragraphs, find out the highest score. Then sort the selected sentences as to preserve the meaning of the original document.

ALGORITHM

This algorithm is a improved version of algorithm by

Shlomi Babluki [13]**(TextRank)**[13]**.**

The intersection function[13]**:**

This module gets two sentences, and returns a score for the intersection point [13]between them. We simply split each sentence into words/tokens[13], check what number of regular tokens we have, and afterward we standardize the outcome with the normal length of the two sentences. [13]

The sentences dictionary[13]**:**

This module[13] is the main part of the algorithm. It takes text as the input and assigns score[13] to the sentences. It consists of the following steps:

1. In the primary step we part the content into sentences, and store the crossing point value between each two sentences in a matrix (two-dimensional array) [13]. So values[2][5] will hold the intersection score between sentence #3 and sentence #6. Computer Science:

In truth, [13]we fair changed over our content into a fully-connected weighted graph! [13]Each sentence could be a node within the graph and the two-dimensional array[13] holds the weight of each edge!

2. In the second step, we calculate an individual score [13]for each sentence and store it in a key-value word reference, where the sentence itself is the[13] key and the value is the overall score. We do that just by summing up all its intersections with the[13] other sentences within the content (not counting itself). [13]Computer Science: We calculate the score for each node in our graph. [13]We basically do that by summing all the edges that are associated to the node[13].

Building the summary:

The ultimate step of our algorithm is producing the ultimate summary. We do that by part our text into sections, and after that we select the best sentence from each section according to our sentences dictionary. Computer Science: The Idea here is that each the paragraph withinthe content represents some co herent subsets of our graph, so we just choose the foremost valuable node from each subset.

Implementation

To run the code we have to type the following commands inside the command prompt as we have not used any IDE the whole work is based on command prompt and java.

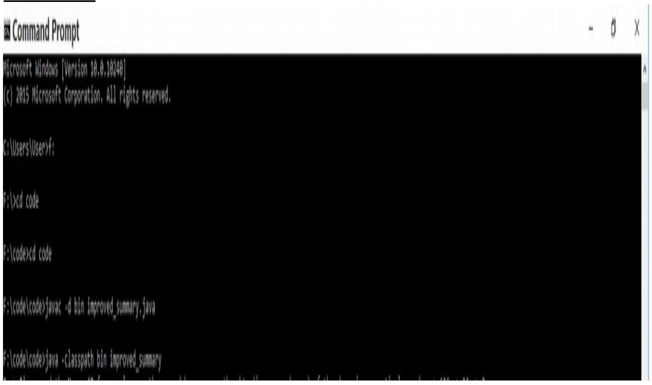
```
F:\code\code>javac -d bin improved_summary.java
```

This command will place all the class into bin folder .The input file is the code folder. We have to specify whole location of the input file in the init() module. Option d specifies the path to store the class files.So now all files are in the bin folder. We execute the main class file by specifying the class path-the bin folder in this case.

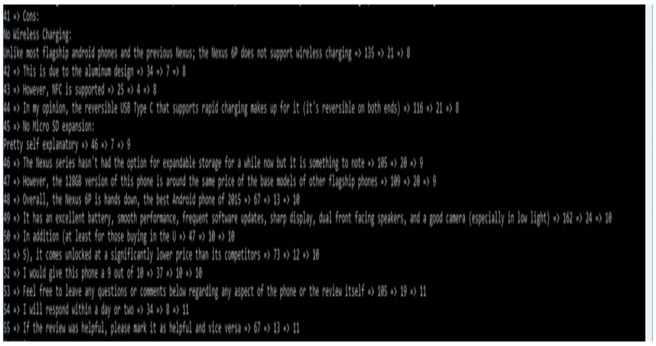
```
F:\code\code>java -classpath bin improved_summary
```

This command will show data from the bin , we have use classpath to specify the folder name in which all the class files are stored.

INPUT:

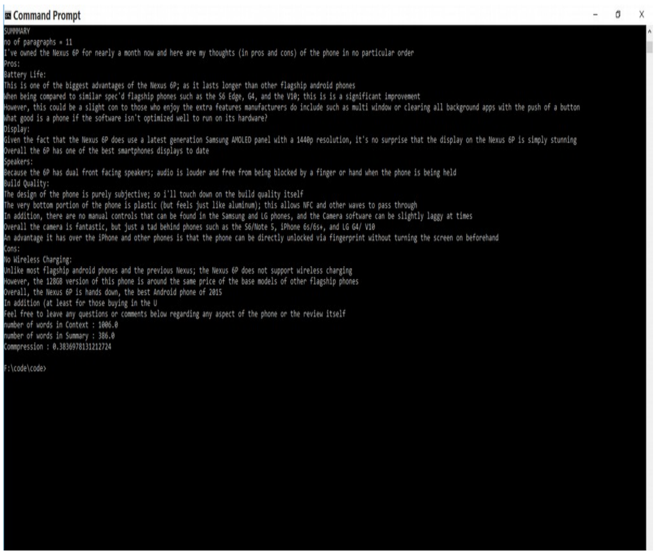


```
Microsoft Windows [Version 10.0.10240]
(c) 2015 Microsoft Corporation. All rights reserved.

C:\Users\User>f:

F:\>cd code

F:\code>cd code

F:\code\code>javac -d bin improved_summary.java

F:\code\code>java -classpath bin improved_summary
```



```
41 => Cons:
No Wireless Charging:
Unlike most flagship android phones and the previous Nexus; the Nexus 6P does not support wireless charging => 135 => 21 => 8
42 => This is due to the aluminum design => 34 => 7 => 8
43 => However, NFC is supported => 25 => 4 => 8
44 => In my opinion, the reversible USB Type C that supports rapid charging makes up for it (it's reversible on both ends) => 116 => 21 => 8
45 => No Micro SD expansion:
Pretty self explanatory => 46 => 7 => 9
46 => The Nexus series hasn't had the option for expandable storage for a while now but it is something to note => 105 => 20 => 9
47 => However, the 128GB version of this phone is around the same price of the base models of other flagship phones => 109 => 20 => 9
48 => Overall, the Nexus 6P is hands down, the best Android phone of 2015 => 67 => 13 => 10
49 => It has an excellent battery, smooth performance, frequent software updates, sharp display, dual front facing speakers, and a good camera (especially in low light) => 162 => 24 => 10
50 => In addition (at least for those buying in the U => 47 => 10 => 10
51 => S), it comes unlocked at a significantly lower price than its competitors => 73 => 12 => 10
52 => I would give this phone a 9 out of 10 => 37 => 10 => 10
53 => Feel free to leave any questions or comments below regarding any aspect of the phone or the review itself => 105 => 19 => 11
54 => I will respond within a day or two => 34 => 8 => 11
55 => If the review was helpful, please mark it as helpful and vice versa => 67 => 13 => 11
```

Output:



```
SUMMMARY
no of paragraphs = 11
I've owned the Nexus 6P for nearly a month now and here are my thoughts (in pros and cons) of the phone in no particular order
Pros:
Battery Life:
This is one of the biggest advantages of the Nexus 6P; as it lasts longer than other flagship android phones
When being compared to similar spec'd flagship phones such as the S6 Edge, G4, and the V10; this is is a significant improvement
However, this could be a slight con to those who enjoy the extra features manufacturers do include such as multi window or clearing all background apps with the push of a button
What good is a phone if the software isn't optimized well to run on its hardware?
Display:
Given the fact that the Nexus 6P does use a latest generation Samsung AMOLED panel with a 1440p resolution, it's no surprise that the display on the Nexus 6P is simply stunning
Overall the 6P has one of the best smartphones displays to date
Speakers:
Because the 6P has dual front facing speakers; audio is louder and free from being blocked by a finger or hand when the phone is being held
Build Quality:
The design of the phone is purely subjective; so i'll touch down on the build quality itself
The very bottom portion of the phone is plastic (but feels just like aluminum); this allows NFC and other waves to pass through
In addition, there are no manual controls that can be found in the Samsung and LG phones, and the Camera software can be slightly laggy at times
Overall the camera is fantastic, but just a tad behind phones such as the S6/Note 5, iPhone 6s/6s+, and LG G4/ V10
An advantage it has over the iPhone and other phones is that the phone can be directly unlocked via fingerprint without turning the screen on beforehand
Cons:
No Wireless Charging:
Unlike most flagship android phones and the previous Nexus; the Nexus 6P does not support wireless charging
However, the 128GB version of this phone is around the same price of the base models of other flagship phones
Overall, the Nexus 6P is hands down, the best Android phone of 2015
In addition (at least for those buying in the U
Feel free to leave any questions or comments below regarding any aspect of the phone or the review itself
number of words in Context : 1006.0
number of words in Summary : 386.0
Commpression : 0.3836978131212724

F:\code\code>
```

VII. Challenges and Issues:

1. Summarizing techniques and evaluation of summary differs.

2. The Code Quantity Principle", which is a linguistic theory about how humans codify the information in a text, depending on what they want a reader to pay more attention to.

3. The main problem arising while implementing text summarization is the sentence compression

4. Lack of readability.

Future Scope:

1. Researchers are working on combining NLP with AI to create abstractive summary.

2. Speech recognition is the new age text summarization technique.

Conclusion:

The computers will be able to know online information based on natural language. Computer will be able to handle data and will be capable of receiving and sending instructions. As the technology is getting developing fast and more internet users are there, there is an information overload. Problems can be resolved through strong text summarizers that presents a short and understandable summary. To have summarized document then you need to develop an efficient system.. Extractive or Abstractive methods can be used for summarization. Extractive is an effortless and simple to implement. Abstractive technique is strong because it produces short and simple summary as it is not possible to produce. Pros and cons are discussed in types of summarization method.

References:

[1] Weigo Fan, Linda Wallace, Stephanie Rich and Zhongju Zang,"Tapping the power of text mining", Journal ofACM,Blacksburg 2005

[2] Baxendale,P.(1958). "Machine-made index for technical literature" –an experiment. IBM Journal of Research developement354-361

[3] Luhn, H (1958). "The automatic creation of literature abs IBM Journal of Research Development, 2(2):159-165.

[4]International Journal of Computer Applications Volume 102– No.12, September 2014 Authors-Nikita Munot Department of Computer Engineering PIIT,New Panvel, India Sharvari S. Govilkar Department of Computer Engineering PIIT, New Panvel, India.

[5]International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-7 Issue-3, July 2017

Upendra Mishra, Chandra Prakash.: MAULIK: "An Effective Stemmer for Hindi Language", International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397, Vol. 4 No. 05 May 2014)

[6]2019 IEEE 4th International Conference on Computer and Communication Systems

Chuleepohn Yongkiatpanich and Duangdao Wichadakul Department of Computer Engineering Chulalongkorn University Bangkok, Thailand

[7] Extractive Text Summarization Using Sentence Ranking

J.N.Madhuri Dept. of Computer science and Engineering CHRIST (Deemed to be University) Bangalore, India Ganesh Kumar.R, Associate professor, Dept. of Computer science and Engineering CHRIST (Deemed to be University) Bangalore, India

[8]A Comparative Study of Supervised and Unsupervised Classifiers Utilizing Extractive Text Summarization Techniques to Support Automated Customer Query Question-Answering

Kevin Lanyo, Agnes Wausi School of Computing and Informatics University of Nairobi Nairobi, Kenya

[9]Extractive Text Summarization Using Ontology and Graph-Based Method Chuleepohn Yongkiatpanich and Duangdao Wichadakul Department of Computer Engineering Chulalongkorn University Bangkok, Thailand

[10]2018 21st International Conference on Computer and Information Technology (ICCIT), 21-23 December, 2018

978-1-5386-9242-4/18/$31.00 ©2018 IEEE

[11]Automatic Bangla Text Summarization Using Term Frequency and Semantic Similarity Approach Avik Sarkar, Md. Sharif Hossen Department of Information and Communication Technology Comilla University, Comilla-3506, Bangladesh

[12] Extractive Text Summarization using Deep Learning

Nikhil S. Shirwandkar Student, Mtech Electronics Engineering K. J. Somaiya College of Engineering, Vidyavihar Mumbai, India

Dr. Samidha Kulkarni Associate Professor, Dept. of Electronics Engineering K. J. Somaiya College of Engineering, Vidyavihar Mumbai, India

[13] https://thetokenizer.com/author/shlomibabluki Shlomi Babluki Naive Language Detector June 29, 2014