AUTOMATIC TEXT SUMMARIZATION

Prakhar Srivastav
Dept. of Information technology
Bharati Vidyapeeth (Deemed to be University)
College Of Engineering
Pune,India

Shivangi Motwani
Dept. of Information technology
Bharati Vidyapeeth (Deemed to be University)
College Of Engineering
Pune,India

P.S.Chavan
Dept. of Information technology
Bharati Vidyapeeth (Deemed to be University)
College Of Engineering
Pune,India

***Abstract*-In today's fast growing world we are blessed with infinite knowledge and thus infinite bundle of text, especially in the web. So it is important to finite the text so it becomes easy to ingest the knowledge. Summaries of long documents, news articles, or even conversations will help us to consume content faster and more efficiently. It isn't easy to hit the bull's eye after reading thousands of lines and paragraphs. Thus we aim to use an algorithm that extract the important information from the the given document file/text file and make another file that consist of all the key information. Our main goal is to reduce the text in fraction of time and thus absorb maximum knowledge in minimum amount of time. This has now become an paramount research area. Being part of 21st century, that is a fast-growing information age, we have vast chronicles to go through and limited time to read, understand, observe, conclude and then start some bona fide creation. Thus this research plays a very important role in the fast growth for the upcoming stage. Absorbing maximum knowledge in minimum time span is the main motive of this technology. Therefore the task is to reduce the size of text by maintaining coherence and semantics.**

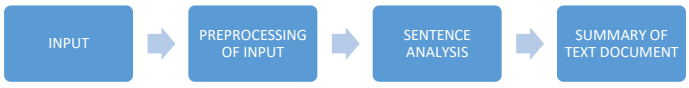
Keywords: Text Visualization, Food Reviews, Opinion analysis, Sentiment analysis.

I.INTRODUCTION

Automatic Text Summarization is a process of condensing document into a summarized form for easy digestion by user, with the help of Algorithm. Summary of a single document or multiple documents can be produced. Summary is supposed to be short and preserve important information. The reduced file should maintain coherence and semantics of the original text. The most challenging part of summarizing tool is choosing which sentences are important and need to be preserved for summary. This main goal of this summarizing tool is that there be no loss of information during summarization phase.
Extractive and Abstractive are 2 methods for texts summarization. Extractive summarization picks sentences from the text that can best represent its summary. It learns to understand the importance of every sentence and their relation with each other rather than understanding the content of the text.
Whereas, Abstractive summarization, is trying to understand the content of the text and then providing a summary based on that, it may or may not have the same sentence as of the original text. It tries to create its own sentences and looks somewhere like human generated summary.

The task of understanding becomes easy when we read summaries.condensed documents help users to manage large amount of information. Text summarisation happens in three steps : analysis, transformation, and synthesis.

First is the input step, in which we input the text file, input can be single or multiple.
Preprocessing is the next step, in this step we perform tokenization and stop words are removed.
Sentence analysis step includes sentences raking on the basis of sentence scores. All these steps sum up to the generation of summary, that is the final step of the system.



Challenges**:**

As the text summariser gives the reduced/condensed form of the original content, there many complications that can occur during the complete process.
Some of the challenges of the text summariser are discussed bellow:

Extract hidden semantic relationship:

A text file consist of many sentences. There are few sentences that are related to other sentences. Some sentences have semantic relationship with other sentence between concept in the text. So, capturing such sentences in the summary is always challenging task.

Relevance detection:

The most important task of a summariser is to find the relevant sentences in our text document so that they cam makeup to the the perfect summary. The summary is highly affected by this. It is always a challenging task to find the relevant sentences out of the huge text document to generate a good quality summary.

II.LITERATURE SURVEY

We have a good supply of data through web, however it's not that straightforward to collect helpful info once reading thousands of lines and paragraphs, therefore from here want of automatic text report rise.

Automatic text summarization account mechanically retrieves the data from the system and gather it at one place by saving our precious time.

H.P. Luhnn was the first one who gets the idea of automatic text summarization of text in 1958.

Various different researchers were also their who planned diffrent techniques in document summarization:

In 1961 G.J. Rath used lexical indicators to

find and summarise the relevant info from documents.

- In 1995 Julian Kupiec used algebraical technique tofind completely different options like uppercase words,length, position of words by using naïve-bayesclassifier.
- In 1997 ChinYew Lin verify the position ofsentences by using algebraical technique.
- In 1999 Eduard Hovy used symbolic wordknowledge with strong NLP processing to indicate theconcepts relevancy.
- In 2005 S.P Yong used neural netwok. Heshowed Summarization = Text pre-processing subsystem + Keywords Extraction sub-system +Summary production sub-system.
- In 1976 M.A. K. Halliday used lexicalsemantic relationships to build lexical cohesionblocks and their patterns.
- In 1984 Ruqaiya Hasan used lexical cohesionto identify similarity chains.
- In 1988 William C.Mann used RST (rhetoricalstructure theory) to encode the terminal nodes of atree.
- In 1991 Jane Morris used cohesion chains todetermine the sequence of associated words.
- In 1997 BranimirBoguraev used saliencybased content characterization to rank the importantsentences in unstructured document.
- In 2010 Li Chengcheng used RST to analyzecandidate sentence, discover rhetoric relations andgive the construction.
- In 2000 Hongyan Jing used human abstractionconcept by taking the closely realated sentences andeliminating the extra ones.
- G.Salton in 1989 used TFI X IDFI techniquesto evaluate the frequency.
- Jun'ichi Fukumoto in 2004 generate abstract byusing TF/IDF for single and multiple documents.
- You Ouyang in 2009 used word hierarchialtechnique for most frequent terms at the top.
- Vikrant Gupta in 2012 used kernel which serveas a guideline to choose other sentences forsummary by using statistical measures.
- Inderjeet Mani in 1997 used graph basedmethod to discover the nodes by applying aspreading activation technique.
- Rada Mihalcea in 2004 used graph basedmethod by adding a vertex for every sentence bycreating links for similar sentences.
- Xiaojun Wan in 2008 used graph based methodby introducing two-link graph for both sentencesand documents.
- Kathleen McKeown in 1995 used time basedtechnique which focuses on how the trends ofevents change with respect to time.
- ShanmugasundaramHariharan in 2012 usedsentence co-relation method where sentences areextracted on the basis of vote casting, scores andpositions to get extracts.
- Tiedan Zhu in 2012 emphasized on logicalcloseness rather than topical-closeness usingsentence co-relation method
- Jade Goldstein in 2000 used clustering,coverage, anti redundancy and summary cohesionfor minimizing redundancy and maximizing bothrelevance and diversity,
- Judith D.Schlesinger in 2008 combinesclustering, linguistics, statistics for summarizationby using clustering based method.
- Nitin Agarwal in 2011 used query-orientedapproach with unsupervised approach with the helpof clustering based method.

III.DIFFERENT METHODOLOGIES

Text summarization can be done by various different methods. Today they are incresingly popular topic with NLP and deep learning.
Few approaches that are taken are:

1. Sentence Scoring based on Word Frequency
2. TextRank using Universal Sentence Encoder

3. Unsupervised Learning using Skip-Thought Vectors

Sentence Scoring based on Word Frequency:

In this method we assign weights to different word based on frequency of the word in passage.

These weights that are assigned to each word, creates score for each sentence. We take the score of the 'N' sentences for the summary. Then we will normalize the scores of each sentence by dividing by its length.

TextRank using Universal Sentence Encoder:

The results generated onceduring universal sentence embeddings and TextRank to come up with summaries. Few measures are necessary to discuss in this part:

TextRank

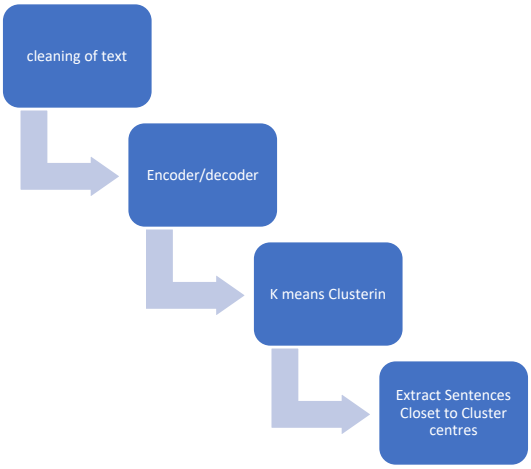
In this we generate a matrix that calculates the probability that a user will move from one page to another.

Universal Sentence Embeddings:

In universal sentence embeddings it encode sentence, word and paragraph into semantic vectors.

We create a cosine similarity matrix thst is used to build our graph. Textrank algorithm is used in this graph to evaluate the importance of each sentence.

Unsupervised Learning using Skip-Thought Vectors:



Here in this there are two main concepts:

Skip Thought Vectors

1. Encoder Network:

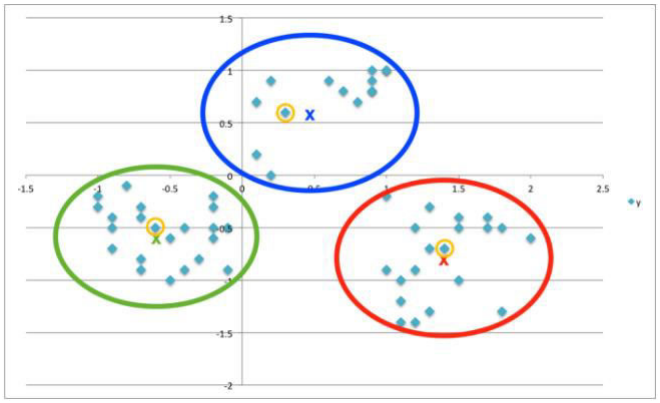
The encoder is generally a GRU-RNN that generatesa fixed length vector representation h(i) for each sentence $S(i)$ within the input.

2. Decoder Network:

The decoder network takes this vector representation h(i) as input and tries to comeup with two sentences — $S(i-1)$ and $S(i+1)$, that might occur before and once the inputs sentence respectively.

K-Means Clustering

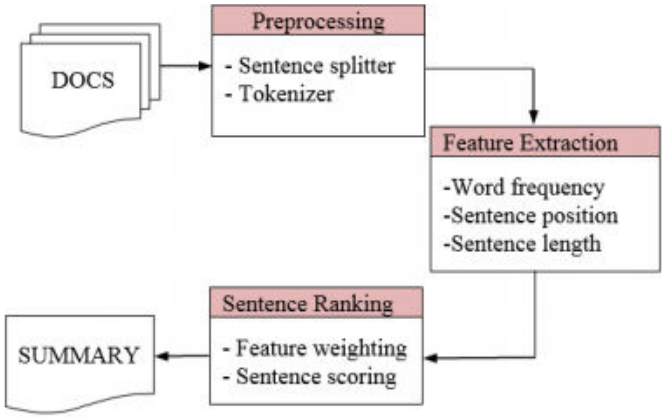
Each cluster have center point in them,in vector space there are point which represents the theme of the cluster. when attempting to make a summary, we must solely want the sentence that is the closest to the center of that cluster.



IV.IMPLEMENTATION DETAILS AND SYSTEM ARCHITECTURE

The system proposed by us does Abstractive text Summarization with help of algorithm by SholmiBabluki.

System Architecture -



Here a text file is sent by the user to the software and the software applies the SholmiBabluki algorithm.

Which then returns the abstractive summary of the test given by the user.

The architectural design of the algorithm is as illustrated by the above diagram. The User Interface is the front-end of the algorithm which is used for the upload of the text files and to provide the results of the summarizer. The pre-processor performs certain preprocessing tasks on the document, such as splitting into sentences, removal of stop-words.However, the process of feature extraction and sentence raking of the summary based on the semantics is the main focus of this architecture for the algorithm.

V ALGORITHM

A. *Sholmi Babluki*

ShlomiBabluki is a Data Scientist and an experienced researcher. Shlomi has vast experience in research in the field of Machine Learning and Natural Language Processing.Shlomi gave a useful algorithms for summarising the content of a text file.

Algorithm:

- Split Context into Paragraphs:
  There is no limit in the size of a text file. In order to summarise the content the entire text file has to be broken into smaller units, thus, the algorithm split the text file into paragraphs, or smaller paragraphs.

- Split Paragraphs into Sentences:
  We need to break the content into more smaller units, thus the paragraphs are further split into sentences.

- Find number of common words [w]:
  To find the difference in 2 statements there need to be proper evaluation of each sentences, so, after chopping the entire file into paragraphs and paragraphs into sentences, we need to observe the most smaller units, that is, sentences, where,

- In sentence s1 and s2 Where w word is in S1 and S2/[(len(s1)+len(s2))/2]

- Intermediate values are stored:
  After we have successfully chopped the text file and recording the common words in each sentences, we store the values in intersect ion matrix.

- Final score stored in Dictionary:
  The final score after the formation of intersect ion matrix is stored the the dictionary:
- Key: Sentence
- Value: Total Sentence Score

- Extract sentence with highest score:
  The score of all the sentences that are recorded are compared with each other, than the sentences those have higher number of score is extracted.

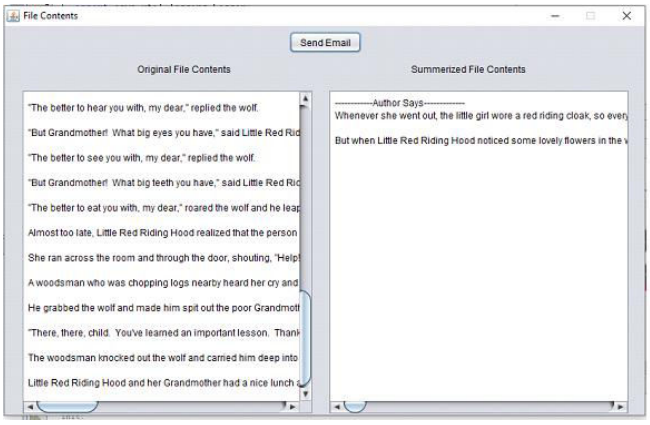
- Sort selected sentences:
  The main and final step is to arrange the sentences into correct order. We have to take special care that meaning of or the context doesn't get changed or gets moulded. The meaning of sentences has to be stationary. Thus the last step if to sort the sentences chronologically.

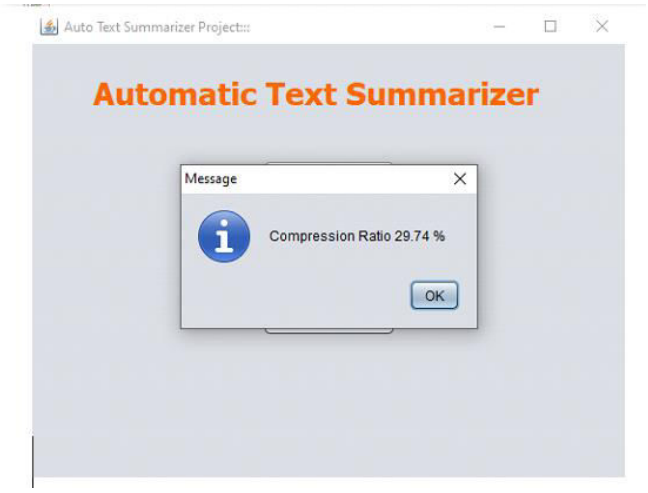
VI.RESULT AND CONCLUSION

With the increasing day by day availability of the data in the form of text day, it becomes arduous to read the entire text data in order to find the the important and relevant information which is both difficult and time consuming for humans.
So, at that time ATS performs an important role by providing a summary of a whole text document by extracting only the useful information and sentences. There are different approaches of text summarization. The real-world applications of text summarization can be: documents summarization, news and articles summarization, review systems, recommendation systems, social media monitoring, survey responses systems. The paper provides a literature review of various research works in the field of automatic text summarization. This research area can be explored more by looking in existing systems and working on different and new techniques.

With the help of this algorithm we could find 60%-70% accuracy while summarising our text document.
This algorithm seems to work good on news articles, technical documents and encyclopaedic entries. However, on essays, fiction and documents with a lot of direct speech.
By using the above explained algorithm by SholmiBabluki we could successfully get the abstractive summary of text.

VII.TEST CASE





REFERENCES

1. SaiyedSaziyabegum, Priti S. Sajja, "Literature Review on Extractive Text Summarization Approaches" International Journal of Computer Applications (0975- 8887) Volume 156- No 12, December 2016.

2. Kang Wu, Ping Shi, Da Pan, "An Approach to automatic summarization for Chinese text based on the combination of spectral clustering and LexRank." IEEE Access 2016.

3. Ibrahim F. Moawad, Mostafa Aref, "Semantic graph reduction approach for abstractive Text Summarization." Seventh International Conference on Computer Engineering & Systems (ICCES), 2012.

4. Akshi Kumar, Aditi Sharma, Sidhant Sharma, Shashwat Kashyap, "Performance Analysis of Keyword Extraction Algorithms Assessing Extractive Text Summarization." International Conference on Computer, Communication, and Electronics (Comptelix), 2017.

5. Pankaj Gupta, RituTiwari and Nirmal Robert, "Sentiment Analysis and Text Summarization of Online Reviews: A Survey." International Conference on Communication and Signal Processing, 2016.

6. Harsha Dave, Shree Jaswal, "Multiple Text Document Summarization System using Hybrid Summarization Technique." 1st International Conference on Next Generation Computing Technology (NGCT), 2015.

7. Yihong Gong, Xin Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis." Proceeding of the 24th annual international ACM SIGIR Conference on research and development in information retrieval, ACM 2001.