Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

Texterizer

Divya Santwani, Akash Bedi, Mohit Bahrani
Department of Information Technology, Vivekananda Education Society of
Institute of Technology, Chembur

Abstract- **Text summarization refers to the method of shortening lengthy portions of textual content. The goal is to create a coherent and fluent summary having only the principle factors outlined in the record. With the exponential growth of on line records and text documents, summarization of text has grown to be a bigger problem and a domain of interest to keep and display the true meaning of information contained in text files. Massive text documents are really hard for human beings to summarize manually. Summarization of text is considered to be an automatic task in which the given massive text in converted into a shorter text containing the basic but precise meaning of original content. It preserves the original meaning of the actual text of whose summary is to be created. With such a major measure of information coursing in the advanced space, there is a need to develop deep learning algorithms that can naturally abbreviate longer messages and convey precise outlines that can easily pass the planned messages. Besides, applying content outline diminishes understanding time, quickens the way toward inquiring about for data, and expands the measure of data that can fit in a zone. Summarization may be very exciting and beneficial assignment that offers aid to many other responsibilities as nicely because it takes advantage of strategies evolved for associated Natural Language Processing tasks.**

Keywords- Text Summarization, Deep learning algorithms, extraction-based summarization.

I. INTRODUCTION

According to the NOP World Culture Score Index, India is the country that reads the most, whether it is online or offline books. Because of busy lifestyle, people gradually loose interest in completing books or stories that contains pages more than 1000. But Science and Technology have improvised the lifestyle of each person. one cannot carry books with themselves, so technology make adequate online books i.e e-books like Amazon Kindle. This makes ease in carrying a lot of books in just one tab. However, it does not resolve the time consuming issue for reading books. Technology further design a software to ease the reading of books. This proposed system is based on NLP(Natural Language Processing) and NLU(Natural Language Understanding) that comprises of two methods i.e Extraction Strategy to summarize a document and an Abstraction Strategy. Also with the assist of Artificial Intelligence and Deep Learning. Each methodology contains different types of algorithm which results in different type of
accuracy of Summarization. An Extractive method composed of selecting important sentences from an original document whereas an Abstractive method composed of generating new and meaningful sentences from the original document. The algorithm will be chosen accordingly to make it more accurate. The result has emergence of an automatic text summarization. Automatic text summarization makes it feasible to quick the document containing a hundred's of sentences to a given threshold quantity of sentences.

There are two basic tasks to be done in the text summarization (i) How to determine the essential content from a massive text. (ii) how to represent the selected essential parts of a document in the form of a condensed summary. Text Research has centered on the commodity several times more: the description, and less on the cognitive basis of the comprehension and processing of texts which is the basis of human resumption. A clearer understanding of the cognitive foundations would be of benefit to some of the shortcomings of existing systems. However, it remains the research challenge to formalize the content of open-domain documents, since most frameworks are focused only on a collection from the original documents.

Text summarization methods are divided into classes: extractive and abstractive. Extractive summarization extracts crucial sentences from supply documents and group them collectively to generate summary. Abstractive summarization creates a brief useful precis via generating new sentences. In this paper we propose a better abstractive approach for greater accuracy of textual content summarization.

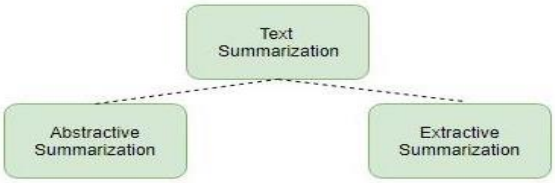


Fig 1. Text Summarization Methods

II. OBJECTIVE

The essential goal of this proposed system is to boom accuracy of text summarizer model by using methods based on deep learning. The model used for developing this system is TextRank Algorithm . TextRank is nicely acceptable forpackages related to whole sentences. The main objective is to summarize a input which is relevant to content of document by adding Indexing and Sorting to the algorithm. The objective is to consume less time of user while using this summarizer. This system tries to summarize the document as much as possible above 50%.

III. LITERATURE SURVEY

In the field of natural language interpretation and natural language processing, text summarization is a major task [6]. This focuses on reducing the scale of the text document to less relevant phrases. It is aimed at creating a simplified representation of an input text that accurately captures the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

fundamental meaning of the original document. Extractive and abstractive approaches to describe a text are two distinct.

(1) Ranjitha and Kallimani in [8] shows findings and results on multi-document Summarization. Multi-document summary consists of automatic summary from multiple documents which tells about the same topic or event. There are three main methods of summarizing multi-documents which are compression-based, abstraction-based, Extraction-based. The extraction-based approach is the least effective method as it generates a summary by picking sentences as it is from the inputs but often leads to redundancy. Some of the problems faced in extractive based approach is overcome by compression based method by generating a summary which is less redundant by removing phrases or words but fails to merge sentences from different sources. The most effective approach is an abstractive-based approach which creates sentences which are not present in original text or documents. The two main approaches towards abstractive multi document summarization are Phrase selection and merging, Phrase Salience Calculation. The collection and fusion of sentences will first and foremost be accomplished by exploring fine-grained nouns and verb phrases. They also argue that Clustered Semantic Graph can be accomplished. It is achieved by content and layout monitoring. The value of this is high in terms of efficiency. The high salience benefit is given by the elimination of redundancy by agglomeration. This paper also reveals findings on the newly produced sentence's improvised standard grammar. Phrase Selection and Merging technique creates new sentences from documents by exploring combinations of syntactic units such as Noun Phrase/Verb Phrase.

(2) Bhatia and Jaiswal in [10] discusses that Automatic Text Summarization is the method of filtering which significant facts that document can be shortened so that it can consume less time of customers. There are summarizers that are available in the market which summarizes single-document or multi-document. The summaries can be created depending on the categories like movies, outlines, biographies etc. This paper shows different method-wise approaches for Automatic Text Summarization. Three types of automated text description are used frequently: Abstract versus extract, single document versus multi document, common versus query versus. Some of the methods are:

- Term Frequency based method which calculates related terms in sentences and by the quantity of frequency which a document contains.
- Graph based methods represent sentences as a set of nodes and give weights to the edges. This method shows best results in areas like image captions, biomedical documents or newswire etc.
- Time based method is the advanced version of graph based method which improves the method and is known as Text Rank.
- Topic based method is used to identify related sentences in the form of Axioms.
- Latent Semantic Analysis based method is used to identify significant topics in documents without any assistance from WordNet.
- There are approaches which are based on fuzzy logic or lexical chains. Lexical chain methods use databases like WordNet to determine the relation of cohesion chains. Fuzzy logic takes into account similarity in keywords or sentence location or indicative expressions.

(3) Merchant and Pande in [11] shows findings which are based on legal documents of court. It is difficult to read long case studies and pick out useful information related to cases. This will not only benefit the lawyers and judiciary system but also to ordinary citizens. Deep learning approach is not used in this model as it uses previously occurring data and doesn't consider new documents but for legal text summarization every case is a new case and thus could not rely on previously seen data. This paper focuses therefore on the summary documents of legal documents using the methodology called Latent Semantic Analysis (LSA) which is a Natural Language Processing (NLP) technique. Dataset which has been used for this system is Indian Judiciary System, data has also been collected from the High Court and District Court Supreme Court. The system uses LSA which is completely automated unsupervised statistical-algebraic summarization. The proposed model consists of pre-processing of data, applying LSA which is widely used forretrieval of information, sentence selection picks the top sentences to create a meaningful summary from the analysis results of LSA. This model is successfully implemented and a short summary is generated having a ROUGE-1 Score of 0.58. However, the evaluation approach (ROUGE) in this program is not entirely successful because specific words are taken into consideration by the test which leads to discontinuity in the generated summary.

(4) An Overview on Extractive Text Summarization shows the relationship with Text Mining with Text Summarization. In this findings it shows use Fuzzy text summarization, statistical text summarization: Text Clustering. This paper also shows finding on different criteria in the text summarization system designing like Summarization based on output summary, Summarization based on details, Summarization based on contents, Summarization based on limitations, Summarization based on number of input document, Summarization based on language acceptance. It also shows different approaches to summarize a text based on extractive method i.e Fuzzy logic based approaches, Cluster based approaches, Graph based approach, Lexical chain approach. This paper shows the Assessment and Evaluation of Summarizing Methods.

IV IMPLEMENTATION

As there are two strategies to sum up an archive there are numerous ways on based which sum up should be possible. Like Summarization on substance, subtleties, input and so forth. There are a great deal of calculations utilized in both the strategies to sum up a report. This proposed framework

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

utilizes Extractive strategy for outline of books which data about specific theme or a story.

Extractive Summarization

These techniques depend on removing a few sections, for example, expressions and sentences, from a bit of content and stack them together to make a synopsis. Along these lines, recognizing the correct sentences for rundown is of most extreme significance in an extractive technique. This system uses PageRank Algorithm implemented in colab to summarize a document.

TextRank Algorithm

Page Rank algorithm is basically used in ranking web pages. It is generated by the probability of the user visiting the page and it generates a probability matrix and tells us which page is going to be visited. This algorithm is the base for the TextRank algorithm. As page rank works on web pages TextRank works on sentences and text.

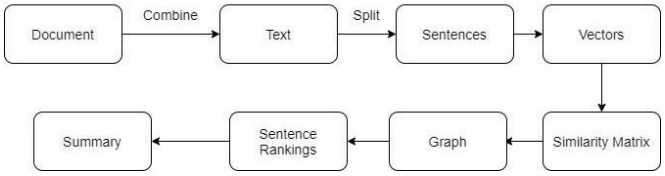
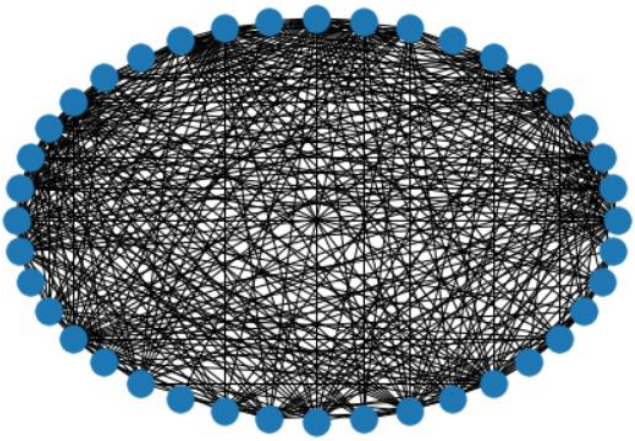


Fig 2. TextRank Algorithm

The above outline shows the flowchart of PageRank Algorithm where the archive is perused as full content without containing any pictures in input record. First it will check the arrangement of info archive whether it is txt or pdf The record is perused from the drive where it is first mounted. In the wake of perusing information the information content is splitted into discrete sentences apply word installing module, for example, Number Batch or Glove Embedding. This framework utilizes Number clump for inserting module. It figures number of lines or bytes present in input report. It is very important to pre-process the data as it gives better result. The steps to pre-process the data in extractive method is

- Removing Punctuations.
- Removing numbers and special characters.
- Removing stop words using nltk library.

After Text Pre-Processing it utilizes vector portrayal. To make comparability lattice. This will store the scores of grid in cosine structure. Changing over the grid into graphical portrayal. Each information has distinctive graphical portrayal as per the lattice framed or scores of archive. Applying PageRank Algorithm with the goal that sentences can be shown according to their likelihood and significance.



Finding MaxRank() :- This will give the rank of sentences having highest score in matrix.

Finding Minrank() :- This will give the rank of sentences having lowest score in matrix.

If some sentences having equal number of ranks, then they all are same and all sentences give the summary but first sentence will be considered. The sentences are then sorted to print summarize.

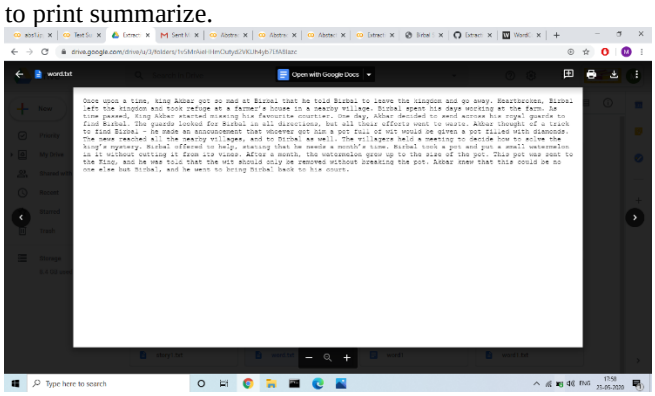


Fig 3. Input file for Extractive Summary

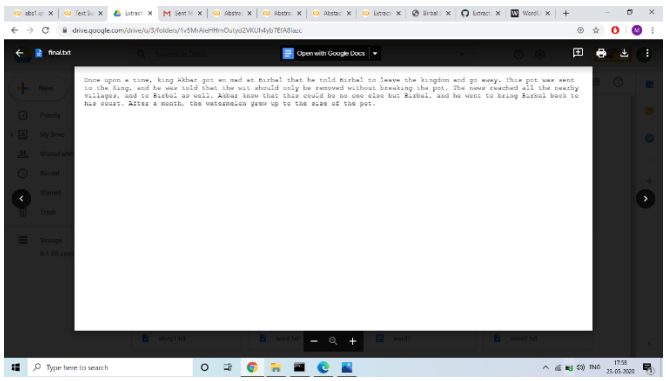


Fig 4. Generated Summary file

V CONCLUSION

We have successfully implemented the desired TextRank architecture for our Extractive Text Summarizer Model. We have used the Concept net Number batch's pre-trained vector, which is better than GloVe pre-trained vector, for representing the words in the input text. We have used Stories of Akbar and Birbal. Text summarization can help humans save a lot of time as time required to read and understand a summary of any given text is far less than

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

the time required to read and understand the entire given text. Although this model performs well with just a subset of the data, we will try to expand the architecture to improve the quality of the generated summaries. Research on this topic is still ongoing since the work in the field of Text Summarization is not yet completed..

VI ANALYSIS

There are other websites available for summarizing a text document. For eg:- http://textsummarization.net/text-summarizer

http://textsummarization.net/text-summarizer

http://textsummarization.net/text-summarizer

http://textsummarization.net/text-summarizer

These all websites work on Extractive based Method which uses TextRank Algorithm. But the results if being compared with our Proposed system is less. Our proposed system gives much accurate results than these websites though in ordered form of summarization. If we summarize a story based on Akbar Birbal using our Texterizer the results would be more accurate also shows the decrement in words comparing with the original document. It will display the number of input/output words and percentage of output words to input words. The results also shows the percentage of Summarized text which will analyze the time saved by the reader.

VI REFERENCES

[1] Kamal Sarkar. "Automatic Text Summarization Using Internal and External Information".

[2] LvCuiling. "Text Automatic Summarization Generation Algorithm for English Teaching", 2016 International Conference on Intelligent Transportation, Big Data & Smart City.

[3] Michael J. Garbade. "A Quick Introduction to Text Summarization in Machine Learning", 2018 from A Quick Introduction to Text Summarization in Machine Learning

[4] Mahsa Afsharizadeh, Hossein Ebreahimpour Komleh, Ayoub Bagheri. "Query-oriented text summarization using sentence extraction technique", 2018 4th International Conference on Web Research (ICWR), 2018

[5] M. Indu, Kavitha K. V. "Review on Text Summarization Evaluation Methods", International Conference on Research Advances in Integrated Navigation Systems (RAINS - 2016)

[6] Alaa F. Alsaqer, SreelaSasi. "Movie Review Summarization and Sentiment Analysis Using Rapidminer", 2017 International Conference on Networks & Advances in Computational Technologies (NetACT), 2017

[7] Nithin Raphal, Hemanta Duwarah, Philemon Daniel. "Survey on abstractive Text Summarization", 2018 International Conference on Communication and Signal Processing (ICCSP), 2018

[8] N.S. Ranjitha, Jagadish S Kallimani. "Abstractive multi-document summarization", 2017 International Conference on Advances in Computing, Communications and informatics (ICACCI), 2017

[9] Romain Paulus, Caiming Xiong, Richard Socherar. "A Deep Reinforced Model for Abstractive Summarization",Xiv:1705.04304v3 [cs.CL] 13 Nov 2017.

[10] Neelima Bhatia, Arunima Jaiswal. "Automatic Text Summarization and its Methods – A Review", 2016 6th International Conference – Cloud System and Big Data Engineering (Confluence).

[11] Kaiz Merchant, Yash Pande. "NLP Based Latent Semantic Analysis for Legal Text Summarization", 2018 IEEE.

[12] Aravind Pai. "Comprehensive Guide to Text Summarization using Deep Learning in Python", 2019 from https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/

[13] Shohreh Rad Rahimi, Ali Toofanzadeh Mozhdehi, Mohamad Abdolahi. "An Overview on Extractive Text Summarization", 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI).

[14] Sumit Chopra, Michael Auli, Alexander M. Rush. "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks", San Diego, California, June 12-17, 2016. c 2016 Association for Computational Linguistics.

[15] Soumye Singhal, Arnab Bhattacharya. "Abstractive Text Summarization".

[16] Ramesh Nallapati, Bowen Zhou, CaglarGulcehre, Bing Xiang. "Abstractive text summarization using sequence-to-sequence rnns and beyond", 2016.

[17] Xinyu Zhang, Zhiqiang Ge. "Automatic Deep Extraction of Robust Dynamic Features for Industrial Big Data Modeling and Soft Sensor application", IEEE Transactions on Industrial Informatics, 2019

[18] Faizan Shaikh. "Essentials of Deep Learning – Sequence to Sequence modelling with Attention", 2018 from https://www.analyticsvidhya.com/blog/2018/03/essentials-of-deep-learning-sequence-to-sequence-modelling-with-attention-part-i/?utm_source=blog&utm_medium=comprehensive-guide-text-summarization-using-deep-learning-python

[19] Chris Olah, Shan Carter. "Attention and Augmented Recurrent Neural Networks", 2016

[20] "Understanding LSTM Networks", 2015 from http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[21] Pranjal Srivastava. "Essentials of Deep Learning: Introduction to Long Short Term Memory", 2017 from https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/?utm_source=blog&utm_medium=comprehensive-guide-text-summarization-using-deep-learning-python

[22] https://www.google.com/url?q=https://github.com/tensorflow/models/tree/master/research/textsum&sa=D&ust=1587987102398000&usg=AFQjCNFnGS3Z8II9E_P0le62HHXoHqQIrw

[23] https://www.google.com/url?q=https://towardsdatascience.com/text-summarization-with-amazon-reviews-41801c2210b&sa=D&ust=1587987102398000&usg=AFQjCNGtvT4AQ-L92MvlUZlVZtLXKAMZHw

[24] Shivangiraj Singh, Aayush Singh, Sudip Majumder, Anmol Sawhney, Deepa Krishnan, Sanjay Deshmukh. "Extractive Text Summarization Techniques of News Articles: Issues, Challenges and Approaches", 2019 International Conference on vision Towards Emerging Trends in Communication and Networking (ViTECoN), 2019

[25] Kiyoumarsi, Farshad. "Evaluation of Automatic Text Summarizations based on Human Summaries", Procedia – Social and Behavioral Sciences, 2015