International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Automatic Summary Generation using TextRank based Extractive Text Summarization Technique

Siddhant Upasani[1], Noorul Amin[2], Sahil Damania[3], Ayush Jadhav[4], A. M. Jagtap[5]

[1]*Student, Dept. of Computer Engineering, AISSMS COE, Maharashtra, India*
[2]*Student, Dept. of Computer Engineering, AISSMS COE, Maharashtra, India*
[3]*Student, Dept. of Computer Engineering, AISSMS COE, Maharashtra, India*
[4]*Student, Dept. of Computer Engineering, AISSMS COE, Maharashtra, India*
[5]*Professor, Dept. of Computer Engineering, AISSMS COE, Maharashtra, India*

Abstract - *A summary is nothing but a brief statement or main points of the given textual content. The main goal of an automatic text summarization system is to identify the most important parts of sentences in the text and present it to the user. There are two approaches for automatic text summarization 1) Extractive Text Summarization 2) Abstractive Text Summarization. In this paper, we solely focus on extractive text summarization technique. The input is given in the form of an article and the extractive text summarization approach is followed by identifying the most important sentences in the text using graph based approach for text summarization.*

Key Words: Automatic Text Summarization, TextRank Algorithm, Extractive Summarization.

1. INTRODUCTION

In our day to day life, we come across a huge amount of information available on internet. So there is a need to provide a way to extract only useful and meaningful information for clear understanding. Summary Generation is the method that identifies most useful part of the textual content and presents it in shortened or compressed format without changing the meaning of the original content. It is very useful mechanism as it saves the time for reading entire document as well as saves the space by storing large amount of data in compressed form [1]. The main goal of an automatic summarization system is to reduce the given text into smaller number of sentences without leaving the main idea of text [2]. There are two approaches for Automatic Text Summarization 1) Extractive Summarization and 2) Abstractive Summarization. The Extractive Summarization approach select most important sentences from specific paragraph or document as exactly they appear in the source text on the basis of some criteria for summary generation while Abstractive Summarization approach constructs entirely new sentences just like human being to produce a generalized summary [4].

1.1 TextRank Algorithm

TextRank algorithm is used to assign some rank or score to each sentence in the document. The rank denotes the importance of the particular sentence in the document. Means, higher the rank of the sentence, the sentence is more important. The sentences with the rank above a specific threshold are considered for summary generation. TextRank is an extractive and unsupervised text summarization technique [5].

TextRank algorithm is based on the PageRank algorithm which is primarily used in Google's Search Engine for ranking the web pages in online search result. Suppose, we have 4 web pages W1, W2, W3 and W4, then the PageRank can be applied by creating a graph as shown in Fig-1.

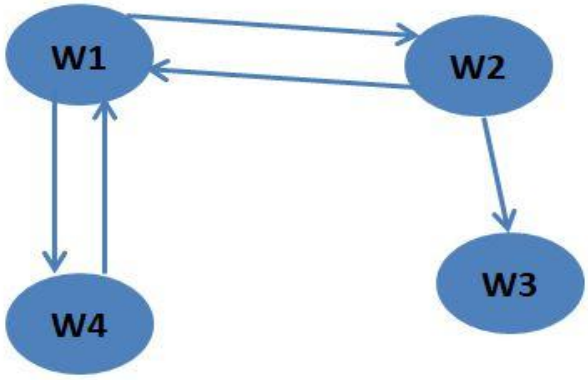

Fig -1: PageRank Algorithm

In the above figure, each node of the graph represents a page and the edges between the nodes denote the links between the pages. In order to rank these pages, we would have to compute a score called the PageRank score. This score is the probability of a user visiting that page. To capture the probabilities of users navigating from one page to another, we will create a square matrix M, having n rows and n columns, where n is the number of web pages. Each element of this matrix denotes the probability of a user transitioning from one web page to another.

In this case, matrix M can be formulated as,

M =

	W1	W2	W3	W4
W1	0	0.5	0	0.5
W2	0.5	0	0.5	0
W3	0.25	0.25	0.25	0.25
W4	1	0	0	0

After matrix formation, the values in the matrix are updated in an iterative manner to get the page rankings.

To convert PageRank into TextRank, consider the following similarities between both algorithms:

- Here, we consider sentences in place of web pages.
- Similarity between two sentences is used as equivalent to the transition probability between two web pages.
- Similarity scores are stored in matrix M for processing to calculate ranks.

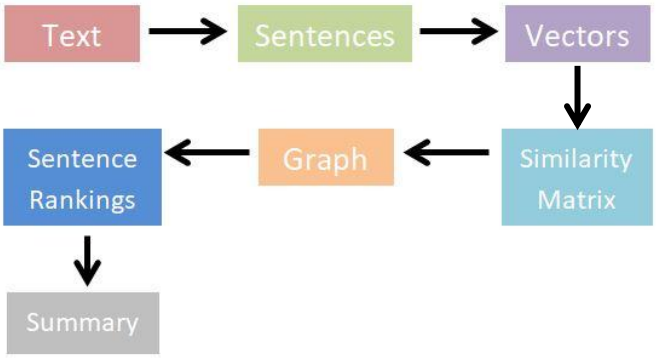

Fig -2: TextRank Algorithm Flow

1.2 Pros and Cons of using TextRank

Pros

- The query-time cost of TextRank to find rank of sentence is low as compared to other algorithms.
- As it computes single measure of quality for a sentence, it provides much more efficiency.
- TextRank is more feasible as it performs operation on current data rather than training a model.

Cons

- Sentences which are not similar to any other sentence in the source text can be called as Dead Ends which can affect summary generation.

2. Proposed System

In proposed system, we solely focus on extractive text summarization. The input to the system is given in the form of text and the summary of the input is obtained by performing steps given below:

Step 1: The given input text is split into sentences.

Step 2: From each sentence, stop words such as I, me, the, etc. are removed as they are of no use in actual processing.

Step 3: Words are brought into their original form by stemming.

Step 4: Similarity of each sentence with other sentence is calculated using Cosine Similarity and similarity matrix is built.

$$\cos\theta = \frac{A.B}{\|A\|\|B\|}$$

Where, A and B are two sentences.

Step 5: Apply TextRank Algorithm to get the score of each sentences.

Step 6: Select sentences with highest scores to be included in the summary.

3. TEXTRANK ALGORITHM TIME COMPLEXITY

As the algorithm is implementing PageRank whose time complexity is O(n+m) (where, n = number of nodes and m = number of edges) and we run it for 'k' iterations until convergence, the time complexity would be O(k*(n+m)) [9].

4. COMPARISON OF TEXTRANK WITH OTHER SUMMARIZATION TECHNIQUES

Clearly, TextRank does the job of Text Summarization very well. But what is its performance as compared to other unsupervised algorithms? The summary generated by the machine is evaluated by referencing to the human generated summary for same textual content. To evaluate the summary generated, the metric known as F1-Score is used.

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

In General:

Bleu measures Precision: How much the words (and/or n-grams) in the machine generated summaries appeared in the human reference summaries.

Rouge measures Recall: How much the words (and/or n-grams) in the human reference summaries appeared in the machine generated summaries [8].

Thus, the formula for F1-Score can be written as:

$$F1 - Score = 2 \times \frac{(Bleu \times Rouge)}{(Bleu + Rouge)}$$

Following table shows the F1-Scores for different summary generation models:

Models	F1 - Score
Cosine	0.02918
WordNet Based Model	0.03354
TextRank	0.03629
Glove-Vec Based Model	0.03054
TF-IDF	0.03371

Table -1: Performance of different models

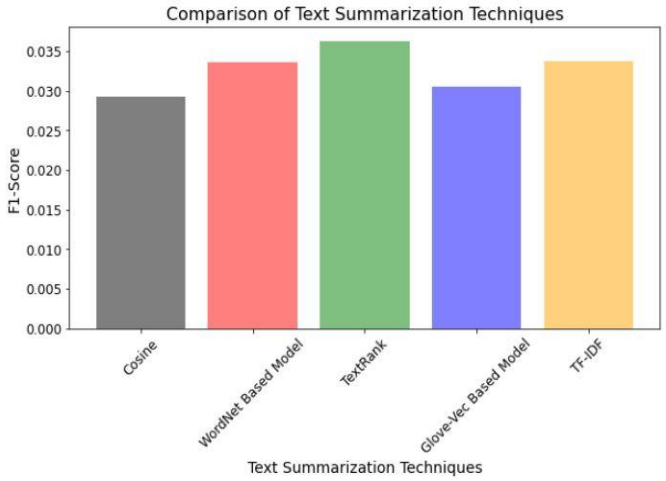

Chart -1: F1-Score of different algorithms

From this chart, we can say that, the TextRank based model gives the highest F1-Score and is more suitable for summarization that the other ones.

5. CONCLUSION

Here, we can obtain the rank or score of each sentence and the sentences with the rank above a particular value can be chosen to be included in the summary.

ACKNOWLEDGEMENT

We would like to thank Prof. A. M. Jagtap Sir, under whose guidance we were able to complete this journal.

REFERENCES

[1] J.N.Madhuri and Ganesh Kumar .R, "Extractive Text Summarization using Sentence Ranking", 2019 International Conference on Data Science and Communication (IconDSC), Year: 2019, Pages: 1-3.

[2] Dharmendra Hingu, Deep Shah and Sandeep S. Udmale, "Automatic Text Summarization of Wikipedia Articles", 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Year: 2015, Pages: 1-4.

[3] Prabhudas Janjanam and CH Pradeep Reddy, "Text Summarization: An Essential Study", 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Year: 2019, Pages: 1-6.

[4] N. Andhale and L. A. Bewoor, "An overview of Text Summarization techniques," 2016 International Conference on Computing Communication Control and automation (ICCUBEA), Year: 2016, Pages: 1-7.

[5] An Introduction to Text Summarization using the TextRank Algorithm. Retrieved 19 September, 2019, from https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

[6] Automatic Text Summarization : Simplified. Retrieved 20 October, 2019, from https://towardsdatascience.com/automatic-text-summarization-simplified-3b7c10c4093a

[7] Automatic summarization. Retrieved 8 December, 2019, from https://en.wikipedia.org/wiki/Automatic_summarization

[8] How do I evaluate a text summarization tool? Retrieved 8 December, 2019, from https://stackoverflow.com/questions/9879276/how-do-i-evaluate-a-text-summarization-tool

[9] TextRank Algorithm Space and Time Complexity. Retrieved 10 December, 2019, from https://stackoverflow.com/questions/50196472/textrank-algorithm-space-and-time-complexity