ISSN 2278-3091
Volume 9 No.2, March - April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse206922020.pdf
https://doi.org/10.30534/ijatcse/2020/206922020

Automatic Text Summarization from Unstructured Text using Natural Language Processing

[1]Mamta Aswani,[2] Ashwini V Zadgaonkar
[1]M. Tech Student, Department of Computer Science, RCOEM, Nagpur, Maharashtra, India, aswanimamta@gmail.com
[2]Assistant Professor, Department of Computer Science, RCOEM, Nagpur, Maharashtra, India, zadgaonkarav1@rknec.edu

ABSTRACT

Text summarization is one which creates short, concrete, fluent data from a longer document. Text summarization is of two types: one is abstractive, and another is extractive. In this paper, automatic text summarization is done using extractive summarization. For the idea of generating deserved summary, the unstructured text is first pre-processing which includes – tokenization, parts of speech, chunking. After pre-processing, feature extraction process take place which uses scoring method and a novel reinforcement learning based training algorithm to extract the data, which is greater than threshold value, that extract data is stored for final summary.

Key words: Multi-document summarization, Natural language Processing, Reinforcement.

1. INTRODUCTION

It is at this present time, when large quantities of information are available and produced online daily. It is therefore necessary, provide a better extraction process that is useful information faster and more efficiently. Text summarization is a technique that focuses on the content which define beneficial data, also creates a compact and accurate brief of the huge documents without failing the overall meaning. So, it is a process which converts long texts into shorter text. Machine learning algorithms can be trained to understand the texts and identify paragraphs that convey important facts and information before releasing the necessary summaries.

In Automatic text summarization the problem has two underlying problems that are one single document and second multiple documents. In Single document is considered input information and summary information extracted from that single text. For multiple documents writing multiple texts on the same topic is taken as the input and output must be related to that topic [4].

Automatic text summaries can be classified into two methods: 1)Based on Extraction: According to an extractive-based summary, a set of words representing the most important points is drawn from a piece of text and put together to form a summary [17].

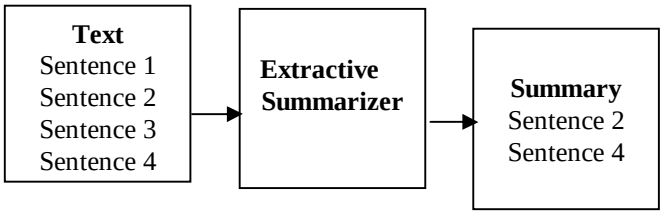

Figure 1: Extractive summarization

2)Based on Abstraction: This method is very interesting. Here, we generate new sentences in the original text. This is contrary to the standard deviation we saw earlier in which we used existing sentences. Sentences with visual summaries may not be in the original text.

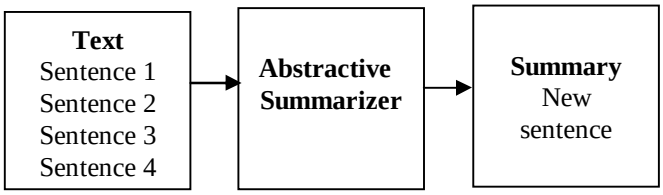

Figure 2: Abstractive summarization

So abstractive summarization seems much more time consuming and more complicated than extractive summarization. Also, summaries can be classified as query summaries and general summaries. In a query-based query, a summary is generated based on a user query, where the text searches to match the user query. While the standard version summarizes the summary that includes the main content of the text. One of the biggest challenges in the standard collection is that there is no title or query available for the summarized process.

A Reinforcement Learning method would be an approach to use in extractive text summarization task to optimize a score function. Emphasis on learning learns how an agent somewhere behaves to get the maximum rewards in the current region [16].

2. LITERATURE REVIEW

J.N.Madhuri and Ganesh Kumar.R,Extractive [1] used, a statistical method for extractive summarization on a single

document. The input text converted in the short form using the idea of sentence extraction method. Sentences are ranked by providing weights and based on their weight they are ranked. From the input high ranked sentences are extracted so that important sentences which have high quality is directed as a summary of the input document.

Ameen Noor, Zuhair Ali and Muntaha Jassim [2] had proposed the method is based on VIKOR algorithm. In First step extract the texts, each sentence has six attributes. In second step weights are find using Cuckoo search algorithm. The next step was VIKOR algorithm which is used for Sentences ranking. Finally, the sentences with less redundant and high scores is an input for a final summary. The work is calculated using Recall-Oriented. The proposed model is checked by using a data set given by the Conference of text analysis for texts. One of the most key points of the given model is the insufficiency of the necessity to plan to set measurement procedures when weights are automatically calculated.

Duke Taeho Jo [3], introduced a text summarization tool that provide a grapgh based on a machine learning algorithm. purpose of this study is three facts: One fact is that a graph is a data graph, the second one is to similar network of graphs are defined and third one is for graph which shows a visualized representation of data, second fact displays various metric which is similar to the graphs are displays and third fact is that the algorithm is valid when summarizing it can be viewed as a division function. The program divides the text into sections, then puts them into graphs where each vertex is a word, and then uses KNN to place the Summarizing text. Then summarized news article domain by domain. For implementation of a text summarization system, it considers the domain granularity and reclassification of each text.

Dijana Kosmajac and Vlado Kešelj [7], introduced an approach which is based on a graph called Text Equalisation with variations in extracting text summaries. The data set used in the research is taken from the online media in Herzegovina and Bosnia. First, it takes a text document of a news article, and then goes through a sentence, after separating the text to get a veteran presentation in each sentence, where it detects a volunteer and analyses the cosine similarity between the punctuation and produces the same matrix. Then it converted into graph. Finally, top ranked sentence forms the final summary. And by using NLP it generated summaries.

Sangaraju Charitha, Nagaratna B. Chittaragi and Shashidhar G. Koolagudi [8], has proposed the multi- document text summarization model of extractive using supervised learning. This model generates the summary by selecting sentences which is done by sentence ranking method.

V V Krishna Kishore and Pramod Kumar Singh [9], introduce a method for generating summaries used to extract a domain based on stack composition. This method uses to calculate the value of numerical symbols depending on the different semantic features and the size of the semantic similarity to select those representative sentences of the document. By using the stack- the algorithm based on the decoder as a template and create on that to generate very close summary. These methods were used which produce a summary of 100 words in a dataset that is accessible as of DUC 2004. The paper defines the algorithms used and their optimization. Spam review scores were found compared to others participating in the competition.

Xiaoping Sun and Hai Zhuge [10], introduced a text-summarization method that generates a Link of semantic Network from a paper containing multilingual units such as communication and communication centers between regions, and then organizes nodes to select high-level sentences to summarize. Here organized six Variety of modes in Experimental Link of semantic Networks. Both aim of tests or tests defines that the position of Network on language units are very helpful to indicate independent sentences. This function not only provides text summaries based on linking semantic links from text but also validates the successful operation of the Semantic reception Link Network and text source. The method used in the use of others to summarize applications such as incremental production.

Krithi Shetty and Jagadish S Kallimani [11], has introduced for the sake of summary, the text was extracted for initial processing including- extracting ASCII characters and stopping words, tokenizing, and Stemming. After which the feature extraction process is performed using tf-IDF, values are compiled, and the pre-processed data is converted into tf-IDF matrix. Sentence are clustered using K-means based on the degree of separation of vector in Euclidean place. Accuracy of summary increase due to increase number of clusters. From each cluster, informative sentences are chosen for the final summary. For the effectiveness of summary recall and precision is measured.

Sumya Akter, Aysa Siddika Asa, Md. Palash Uddin, Md. Delowar Hossain, Shikhor Kumer Roy, and Masud Ibn Afjal [12], proposed the method for summarization which is done in single or multiple text which is in Bengali language. First document is done tokenization process, then stemming etc. then, next process has taken place which is (TF/IDF) method which is used to finding word score.Finally, k means clustering is used for final summary.

Shohreh rad Rahimi and Ali Toofanzadeh [13], had described the text mining and its relationship with text summarization. Then some summarization approaches and their important parameter for extracting sentences, main stage of the summarization process is defined and present the most significant criteria of extraction is reviewed. Finally, the most fundamental proposed evaluation methods are considered. Mr. Krushnadeo Tanaji Belerao and Dr. S. B. Chaudhari [14], introduced the abstract outline generation of enormous variety of documents for giant information is planned which

can think about user input as topic. this technique is intended by works with MapReduce framework for agglomeration and Hidden Markoff Model for summarization victimization DBSCAN algorithmic rule. the strategy follows within which documents area unit scanned with similarity and machine learning technique area unit used. By applying agglomeration, enhances the summarizer system to gather precise words instead of repeating the duplicates words.

3. PROPOSED APPROACH

The summary of the document is the use of natural language (NLP) for the proposed data mining to extract the most important information of the document (s). We use sentence reinforcement to extract a summary from one or more documents. The proposed procedure is used to summarize many documents. In this way, steps include:

3.1. Pre-processing

Pre-processing is an important step for natural language processing (NLP). Converts text into excavation form so that algorithms used in this perform better. It is done in the unstructured text representation. It basically includes: Tokenization, parts of speech, stop words and stemming.

3.1.1. Tokenization

Tokenization is a process of dividing the sequences of string. It separates the sentence in a word count. In the token-making process, some characters such as punctuation are discarded. Tokenization process is done by following a few steps:

- Token are break by white-space or line breaks.
- White space or punctuation may or may not be added as needed.
- All characters within the complete string are part of the symbol.

Tokens can be generated by all numeric, alpha or alphanumeric characters only. The tokens themselves can be dividers.

3.1.2. Stemming

The stemming technique used to improve the accuracy of the text. Stemming process is to lower all words with the same stem as the standard form while lemmatization process removes illegal ends and returns the word's status [15]. For example, A stemming process shorts the words such as “reader”, “reading”, “reads” to the root word, “read” and “walking”, “walk”, “walked” shorts the word to “walk”.

3.1.3. Stop words

Some words which need to be excluded from the vocabulary which are commonly used and it seems like a small amount in useful to consider a sentence, it has to be ignored by both indexing entries for searching and result of a search query when retrieved.

3.2. Scoring Process

After completion of this pre-processing step, now we are calculating the score of the sentence, for that following steps required:

3.2.1. Count Term(ct)

The term represents the frequency i.e., however usually a given term happens in a very document assortment and is given by term $count_i=\sum_{j=1}ni,j$ Where ni,j is that the variety of occurrences of ith term in jth document and |D| is that the variety of documents.

3.2.2. To match the document lines with each other and to normalize the term frequency

For all terms occurring in a very document by most tf to normalize the tf weights in this document. for every document d, let $tf_{max}(d)=max_{ţ\epsilon d}tf_{ţ,d}$, where ţ reaches the document d for all terms t. Then, for every term t we tend to cipher a frequency of normalized term in document d by

$$ntf_{t,d} = a + (1 - a)\frac{tf_{t,d}}{tf_{max}(d)},$$

After getting the word frequency and the sentence frequency, find the score.

3.2.3. Algorithm for the find Score:

Algorithm 1: Find score

a. Find all the words in the sentence(words).
b. Initially declare the titlewords equal to zero and the score equals to zero.
c. Find the score, score equals to score + index[word]/(1+cindex[word]).
d. If the word is present in the title, then the title word increased by 1.
e. Finally, the titlewords equals to 0.1* count of tokens in sentence/ count of those tokens present in title.
f. And, finally score is equals to the sum of score and titlewords.
g. Set the threshold, if the score is greater and equal to the threshold add to the sentence.
h. These sentences are the title of the summary.

3.3. Applying Reinforcement learning based algorithms

Now use the following algorithm to find the final summary from the input document

Algorithm 2: Reinforcement learning algorithm to find summary.

a. Go to each word in the input document(s), and each word in the title.
b. Find the synsets of each of these words (synonyms)
c. Find similarity between synsets of input word with synsets of title words = a
d. Find similarity between synsets of input word with synsets of title words = b
e. Find initial score = Jaccard distance between input word and title word.
f. Find final score = mu * initial_score + (1-mu)*average ((a+b)/2) where, mu is constant.
g. Find the least matching words based on the final score.
h. Find the summary using the minimal matching score.

4. RESULT AND DISCUSSION

The system with multiple documents is tested in the work, which contains multiple sentences. The sentences whose score is greater than threshold is generated as an output by the summarizer. We have used Python 3.7 and NLTK to implement text summarization using automated knowledge extraction and the output of documents without using algorithm is in fig 3 and output of the documents with using algorithm is in fig 4.

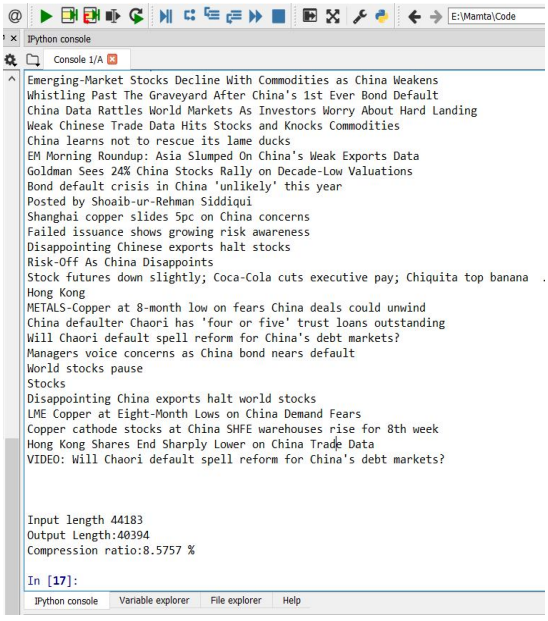

Figure 3: Summarization Output using only score formulation

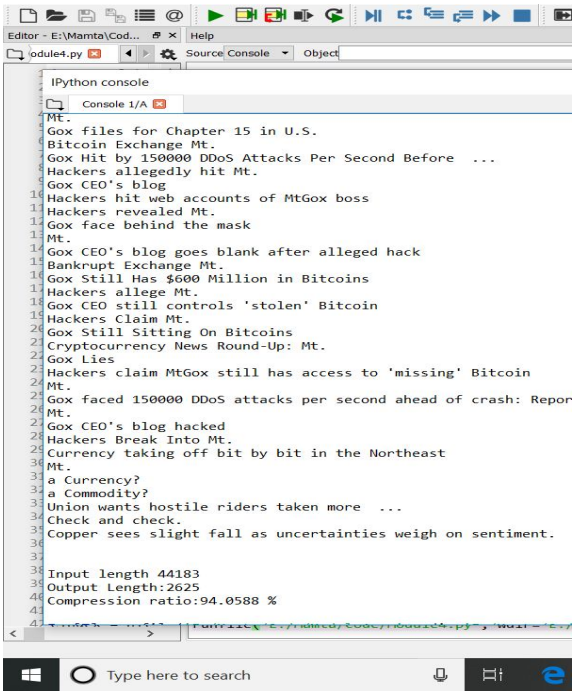

Figure 4: Summarization output using score formulation With reinforcement algorithm

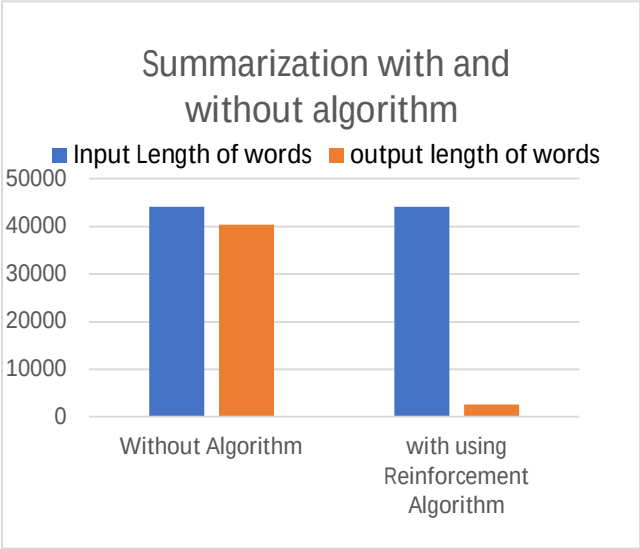

Figure 5: Graph displays the difference of output

Table1: Displays the compression ratio

Summarization	Input Length of words	Output Length of words	Compression Ratio
Without Algorithm	44183	40394	8.057%
With reinforcement	44183	2625	94.058%

5. CONCLUSION AND FUTURE SCOPE

The objective of the automatic text summarization of unstructured text using NLP is most important. In this proposed word we summarize the documents using the reinforcement algorithm. First, we remove the stop words and stemming the text, then find then the score of the sentence. The highest score considers as the important sentence and then using the reinforcement algorithm. Reinforcement ranking algorithm improves the extractive summarization. The paper is used for multiple documents. The approach of the paper shall be used to any unstructured data, also deliver the automatic generation of Summarized text. In future, it can be done with the knowledge extraction and will implement with the abstraction.

REFERENCES

[1] J.N.Madhuri and Ganesh Kumar.R,Extractive, **Text Summarization Using Sentence Ranking**, IEEE, 2019. https://doi.org/10.1109/IconDSC.2019.8817040

[2] Zuhair Hussein Ali, Ameen A. Noor and Muntaha AboodJassim, **Vikor Algorithm based on Cuckoo Search for Multi-Document Text Summarization**, Springer, 2019.

[3] Duke Taeho Jo, **Validation of Graph based K Nearest Neighbor for Summarizing News Articles**, International Conference on Green and Human Information Technology (ICGHIT), 2019.

[4] N. Nazari and M. A. Mahdavi, "**A survey on Automatic Text Summarization**", Journal of AI and Data Mining Vol 7, No 1, 2019.

[5] Arun Krishna Chitturi, Saravanakumar Kandasamy, **Survey on Abstractive Text Summarization using various approaches**, International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), December 2019. https://doi.org/10.30534/ijatcse/2019/45862019

[6] Dian Sa'adillah Maylawati, Yogan Jaya Kumar, Fauziah Binti Kasmin, Basit Raza,**Sequential Pattern Mining and Deep Learning to Enhance Readability of Indonesian Text Summarization,** International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), 2019. https://doi.org/10.30534/ijatcse/2019/78862019

[7] Dijana Kosmajac and Vlado Kešelj, **Automatic Text Summarization of News Articles in Serbian Language**, 18th International Symposium infoteh-jahorina, 2018.

[8] Sangaraju Charitha, Nagaratna B. Chittaragi and Shashidhar G. Koolagudi, **Extractive Document Summarization Using a Supervised Learning Approach**, IEEE, 2018. https://doi.org/10.1109/DISCOVER.2018.8674133

[9] VV Krishna Kishore and Pramod Kumar Singh, **Multiple Data Document Summarization**, Conference on Information and Communication Technology (CICT'17), 2017.

[10] Xiaoping Sun and Hai Zhuge, **Summarization of Scientific Paper through Reinforcement Ranking on Semantic Link Network**, IEEE, 2017.

[11] Krithi Shetty and Jagadish S Kallimani, **Automatic Extractive Text Summarization using K-Means Clustering**, International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 2017.

[12] Sumya Akter, Aysa Siddika Asa, Md. Palash Uddin, Md. Delowar Hossain, Shikhor Kumer Roy, and Masud Ibn Afjal, **An Extractive Text Summarization Technique for Bengali Document(s) using K-means Clustering Algorithm**, IEEE, 2017. https://doi.org/10.1109/ICIVPR.2017.7890883

[13] Shohreh rad Rahimi and Ali Too fanzadeh Mozehdehi, **An overview on Extractive text Summarization**, IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), 2017. https://doi.org/10.1109/KBEI.2017.8324874

[14] Mr. Krushnadeo Tanaji Belerao and Dr. S. B. Chaudhari, **Summarization using Mapreduce Framework based Big Data and Hybrid Algorithm (HMM and DBSCAN)**, IEEE International Conference on Power, Control, Signals and Instrumentation Engineering 2017.

[15] Tiwari Chanchal Chitranjan, Vaibhav Doshi & Rahul Kumar, **Text Summarization: A Review**, Imperial Journal of Interdisciplinary Research (IJIR), 2017.

[16] Gyoung Ho Lee, Kong Joo Lee, **Automatic Text Summarization Using Reinforcement Learning with Embedding Features**, The 8th International Joint Conference on Natural Language Processing, 2017.

[17] Deepali K. Gaikwad and C. Namrata Mahender, **A Review Paper on Text Summarization**, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.