Extractive Text Summarization

BY

Md. Jamaner Rahaman

ID: 172-25-595

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science and Engineering

Supervised By

Dr. Sheak Rashed Haider Noori

Associate Professor and Associate Head

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

AUG 2018

APPROVAL

This Thesis titled **“Extractive Text Summarization**”, submitted by *Md. Jamaner Rahaman* to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 15th August 2018.

BOARD OF EXAMINERS

____

Dr. Syed Akhter Hossain **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

____

Dr. Sheak Rashed Haider Noori **Internal Examiner**
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

____

Md. Zahid Hasan **Internal Examiner**
Assistant Professor & Coordinator of MIS
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

____

Dr. Mohammad Shorif Uddin **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that, this thesis has been done by me under the supervision of **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head** of CSE Daffodil International University. I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of CSE
Daffodil International University

Submitted by:

Md. Jamaner Rahaman
ID: 172-25-595
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the M.Sc. thesis successfully.

I really grateful and wish my profound my indebtedness to **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head,** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "*Extractive Text Summarization*" to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

I would like to express my heartiest gratitude to **Dr. Syed Akhter Hossain, Professor and Head,** Department of CSE, for his kind help to finish my thesis and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Selecting the important words by counting the word score it is one of the ways of extractive text summarization. Summarizer helps the people to reduce the time destroying. People can easily understand by seeing the summary. This thesis based on bag of words model. Using this model I created a text summarizer for generating the summary follow the extractive text summarization. Inside this model I additionally added weighted histogram for counting the word score. Finally, this summarizer is able to create a summary of the full text file.

Keywords: Bag of Words model, Extractive summary, Histogram, Weighted histogram, Tokenization.

TABLE OF CONTENTS

CONTENTS **PAGE NO**

LIST OF FIGURES

FIGURES **PAGE NO**

LIST OF TABLES

TABLES **PAGE NO**

1 Introduction

Summarize the main text into a smaller version with the all necessary information and easily understandable by the user is called text summarization (Mittal, Agarwal, Mantri, Goyal, & Jain, 2014). It shows the importance of full text documents. Sometimes we do not have enough time to read whole text that situation we need gist points of that full text, in this time summarization helps us a lot.

When human generate any summarization it takes time but simple. However, the text summarizers do this automatically and faster. But it is difficult to choose correct meaningful words or sentences by the machine. Another one most important thing is: Are the summaries satisfying the user or not? (KHAN & SALIM, 2014).

There are two types of text summarization: Extractive and Abstractive. Extractive text summarization takes the important words as same as the main text and produce the new summary (M, C, Ganesh, & A, 2016). Abstractive text summarization produces the summary more likely the human. It has to understand the original text and by using linguistic method generate the new words or sentences (KHAN & SALIM, 2014).

In this thesis, I concentrate on extractive text summarization technique. I used bag of words model to create a text summarizer. This model follows the extractive way to collect the important words from the whole text documents.

This repot divided into 5 sections. Section 2 shows the literature review. In section 3 describes the methodology. Result stored in the section 4. Section 5 included the conclusion.

2 Literature review

In present situation text summarization is one of the most valuable researches throughout the world. Many researchers are doing their research on it. It is a significant part of natural language processing area. Basically extractive text summarization quits easier than the abstractive summarization. We can see lots of methods day by day developed by the researchers.

Mittal, Agarwal, Mantri, Goyal, & Jain (2014) discussed a graph based method which removes the unnecessary sentences for generating the extractive text summary of input text. They worked with those types of documents which are doing same discussion again and again, basically their methods useful for those types of documents. This approach worked into two periods one for creating outputs from the inputs, another for reducing the unnecessary summary from the previous outputs. They measured their outputs by using ROUGE-Recall Oriented Understudy and got 60% of similarity result with the main sentences which are collected as a summary [5]. M, C, Ganesh, & A (2016) reviewed the abstractive summarization based on ontology. They tried to show different types of ontology methods for abstractive text summarization already done by the researcher. They also evaluated the methods [4].

Fang, Mu, Deng, & Wu (2016) worked on the technique of sentence scoring for extractive text summarization. Basically they proppposed a model which name is CoRank based on word sentence co-ranking model. It is also related with the graph-based unsupervised ranking model. They also used 600 documents for clarifying their work [2]. Bhargava, Sharma, & Sharma (2016) proposed a graph based technique for abstractive text summarization. They used word graphs for shortening information. Their method does not take any domain knowledge. They worked with two datasets from the conference named Document Understanding Conference (DUC) which was arrenged by the National Institute of Science and Technology (NIST) for their evaluation. They also selected their baselines: Baseline 1, baseline 2 and measurement those baselines by using ROUGE-1, ROUGE-2. ROUGE-1 and ROUGE-2 are for Baseline 1, ROUGE-1 is for Baseline 2. Both the time they used ROUGE-1 and ROUGE-2 for comparing with the Baseline 1 and Baseline 2 [1].

3 Methodology

I am not going to work with deep learning based approach but I have worked with natural language processing based approach. Using my methodologies easily produce the summary from the website. I took help from the bag of words model but the ending part of bag of words model finished by binary matrix. We know that binary means the combination of only 0 and 1 that's why it's difficult to count words score because in this situation one more words score is 0 or 1 [6]. So we are not able to take most important words. At first I have to create words histogram and then I added another weighted histogram to generate individual words scores. Finally I can count the scores of words and by adding those words scores I find out the most important sentences. What our summarizer will do?

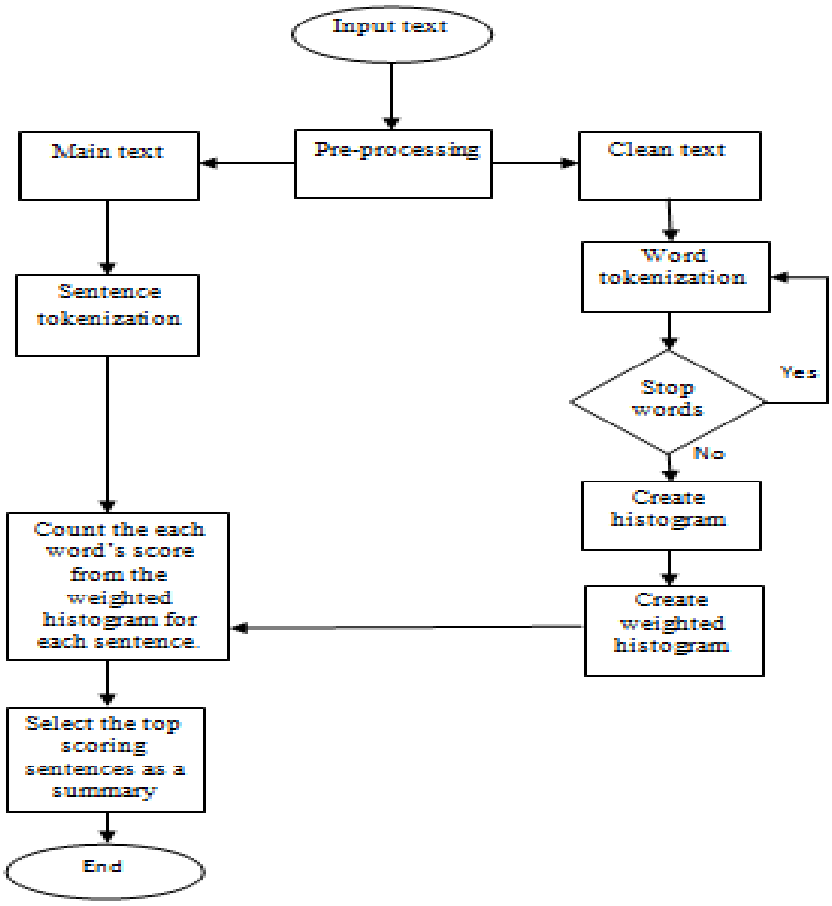
Input text
Pre-processing
Main text
Clean text
Word tokenization
Sentence tokenization
Stop words
Yes
No
Create histogram
Create weighted histogram
Count the each word's score from the weighted histogram for each sentence.
Select the top scoring sentences as a summary
End

Fig. Method overview

3.1 Pre-processing

At first I preprocess the input data and divided into two sections. The two sections are main text and clean text. Inside the main text I have reduced all extra spaces; I have transferred all data upper case to lower case etc. Inside the clean text again I have reduced all extra spaces, transferred all data upper case to lower case, I also have reduced the all full stops (.). This is the preprocessing stage.

Table 1. Input preprocessing.

Normal text	Pre-processed text	
	Main text	Clean text
Jackfruit is my most favorite fruit. It is a delicious fruit. I like to eat Jackfruit.	jackfruit is my most favorite fruit. it is a delicious fruit. i like to eat jackfruit.	jackfruit is my most favorite fruit it is a delicious fruit i like to eat jackfruit

3.2 Tokenization

In the preprocessing stage I divided normal text into two sections because when I am going to create histogram I need all the words so that I generated clean text. Main text I have generated because of sentence tokenization. I need all the sentences for counting words weight and producing the final summary. I divided all the sentences individually from the main text this is called sentence tokenization. I also have separated all the words from the clean text this is called word tokenization. Both examples are given below.

Table 2. Sentence tokenization from the pre-processed data.

Main text	Sentence tokenization
jackfruit is my most favorite fruit. it is a delicious fruit. i like to eat jackfruit.	jackfruit is my most favorite fruit.
	it is a delicious fruit.
	i like to eat jackfruit.

Table 3. Word tokenization from the pre-processed data.

Clean text	Word tokenization
jackfruit is my most favorite fruit it is a delicious fruit i like to eat jackfruit	jackfruit, is, my, most, favorite, fruit, it, is, a, delicious, fruit, I, like, to, eat, jackfruit.

3.3 Stop words

When I am going to create histogram I do not take the stop words. The stop words are: 'the', 'is', 'are' etc. In summarization stop words do not create any valuable meaning and if I consider stop words that time lot of extra words will come for the long text file which will create complexity. So that I do not need to consider any stop words.

3.4 Histogram

Histogram means how many times each word appears in every sentence. Histogram created for scoring the words. It is also needed for generating weighted histogram. In the bag of words model matrix generate from the histogram. But here I am not going to generate matrix.

Table 4. Creating histogram for scoring the words list.

Word	Count
jackfruit	2
is	2
my	1
most	1
favorite	1
fruit	2
it	1
a	1
delicious	1
i	1
like	1
to	1
eat	1

3.5 Weighted Histogram

$$\text{Weighted histogram} = \frac{\text{Each of the word score}}{\text{Maximum word score}}$$

Weighted histogram generated for counting the full sentence score because from the weighted histogram I am able to collect new weight for every word.

Table 5. Creating weighted histogram for counting the words weight.

Word	Count
jackfruit	2/2 =1
is	2/2 =1
my	½ = 0.5
most	½ = 0.5
favorite	½ = 0.5
fruit	2/2 =1
it	½ = 0.5
a	½ = 0.5
delicious	½ = 0.5
i	½ = 0.5
like	½ = 0.5
to	½ = 0.5
eat	½ = 0.5

4 Results

Finally I counted the each sentence score from the weighted histogram and I got the individual sentence score. I just added the each word weight of a sentence.

Table 6: Add the each word's weight to count the each sentence score for selecting the summary.

Sentence	Score
jackfruit is my most favorite fruit.	4.5
it is a delicious fruit.	3.5
i like to eat jackfruit.	3

4.1 Summary

I have taken the top scorer sentences as a final summary. This sentence taken by the importance of words and it is called extractive text summarization technique.

Table 7. Selecting final summary.

Top scores	Produce summary
4.5	jackfruit is my most favorite fruit.

5 Conclusion

There is no binding for researching something. So it is a continuous process to develop techniques day by day. Still now a lot of researches are ongoing in the field of natural language processing. So that extractive text summarization also needs more research. Sometimes we can see extractive text summarization does not match the needs of users because of the lack of the right words or sentences. In this thesis I showed a method for developing an extractive text summarizer and also developed that summarizer. I used bag of words model to develop our summarizer. But at the end of the bag of words model produce a binary matrix that's why variable does not identify the real value of words. So that I additionally used weighted histogram for identifying the words score. Near future I will again work on it for further development.

References

1. Bhargava, R., Sharma, Y., & Sharma, G. (2016). ATSSI: Abstractive Text Summarization using Sentiment Infusion. *Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)* (pp. 404-411). Elsevier B.V.
2. Fang, C., Mu, D., Deng, Z., & Wu, Z. (2016). Word-Sentence Co-Ranking for Automatic Extractive Text Summarization. *Expert Systems With Applications* , 189-195. [doi: 10.1016/j.eswa.2016.12.021]
3. KHAN, A., & SALIM, N. (2014). A REVIEW ON ABSTRACTIVE SUMMARIZATION METHODS. *Journal of Theoretical and Applied Information Technology , 59* (1), 64-72.
4. M, J. M., C, S., Ganesh, A., & A, D. (2016). A Study on Ontology based Abstractive Summarization. *Fourth International Conference on Recent Trends in Computer Science & Engineering* (pp. 32-37). Chennai, India : Elsevier B.V.
5. Mittal, N., Agarwal, B., Mantri, H., Goyal, R. K., & Jain, M. K. (2014). Extractive Text Summarization. *International Journal of Computer Engineering and Technology , 4* (2), 870-872.
6. Brownlee, J. (2017, October 9). *A Gentle Introduction to the Bag-of-Words Model*. Retrieved August 3, 2018, from machinelearningmastery: machinelearningmastery.com