ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

A New Approach for Text Summarizer

[1]Dr. M. Suman, [2]Tharun Maddu, [3]A. Shalini, [4]K. Bhavana.

[1,2,3,4] Department of Electronics and Computer Engineering, K.L University, Vaddeswaram, Guntur

Abstract: Text Summarization is a process of giving the shorter version of a text document. For the research scholars who want to do research on a particular domain have to search a lot of documents. It is very difficult and it takes a lot of time to go through the domain, in this case there is a chance of missing some key word in those documents. To get rid of this problem it is better to have a summary of a document. Summarizer gives the summary of a paper. The summarizer can be developed using some algorithms like Sentence Position, Sentence resemblance to the title, Lexical Similarity etc. The main aim is to reduce the body of the text and maintaining coherence and avoiding redundancy.

Keywords: Summarizer, Sentence Position, Sentence Resemblance, Lexical Similarity, coherence, redundancy.

1. Introduction

The volume of increasing the information in internet is increasing day by day. As a result referring that large amount of data is very difficult in this real world. This text summarization makes easier for the natural language processing applications. Applications such as information retrieval, Question answering and text comprehension etc. Text summarization will help users to manage lot of time and information by extracting the information from various documents of the same domain. These emerging technologies will bring new challenges in this research which is required to be solved. It is very essential to have an overview and analyze about the past and the present progress and highlighting the main advanced achievements and then outlining the limitations. This paper describes the customized algorithms.

2. Literature Survey

To solve this problem the solution is to use a summarizer with some algorithms. The algorithms used represent the original text, perform sentence scoring and having a summary of several sentences. There are multiple ways to get this summarizer such as word frequency, lexical similarity, and word co-occurrence.Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form.An Abstractive summarization attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language [1]. Before creating the summary of a text, first it is pre-processed by segmentation, tokenization, and removal of stop words and stemming. Summarization can also be single document or multiple documents. Single document summarization is summary generated from a document while multiple document summarizations is summary generated from two or more related documents. According to type of summary, different approaches are employed [3]. The original method of Luhn assigned weights to each sentence of the document according to statistical criteria, in fact, a simple function of the number of high frequency words occurring in the sentence [9]. There are multilingual platforms that support, at most, 2 languages by proposing a language independent summarization platform that provides corpus acquisition, language classification, translation and text summarization for different languages [11]. When going through the multiple documents on the same domain redundancy is the major issue to be taken care off and since there are multiple documents there should be only one clear summary. There is less focus on the precision with respect to the text context. Some papers describe a simple approach of searching through a single or a multiple documents is a simple one or searching with the help of a multiple key words and the results are provided in a well-defined format.

3. Proposed Work

To design a Text summarizer there are some specifications such as functional specifications and program specifications.

3.1 Functional Specifications:

The functional specificationof the compendium generator is mainly to generate accurate key phrases after parsing a document. When multiple papers on a same domain is given then the summary generated is based on the level of coherence between the papers given. The redundancy between the papers is eliminated.

The summarizer first uses Text Segmentation. Text segmentation is done to eliminate the stop words. Stop words are the words they are too common and irrelevant such as ‘a’, ‘an’, ‘,’,’’. Text segmentation also stems the words. The sentences can be ranked. The sentences are ranked using some combination of the algorithms. After obtaining the rank the sentences are choose based on the rank of the sentences. This can be done by maintaining the correlation with the base idea. The next step is to summarize the document belongs to the same domain.

The redundancy statements between the documents are eliminated. In the same way the coherence between the documents of the same domain is also presented. The possible links between the multiple papers of the same domain are also presented.

3.2 Program Specifications:

Some algorithms are used:

3.2.1 Sentence Position: The position influences the importance of the sentence. In generally the most important sentence comes at the beginning of the paragraph.

3.2.2 Sentence Resemblance to the Title: This means that there is a vocabulary overlap between the various sentences and the title of the document.

3.2.3 Lexical Similarity: Lexical Similarity is based on the assumption that the important sentences are identified using the strong chains.

3.3 Natural Language Tool kit:

NLTK is the best platform to do the python programs and work with the human language data. It is very easy and interfaced with the 50 corpora and the lexical resources. The Natural Language Toolkit is a suite of program modules, data sets, tutorials and exercises, covering symbolic and statistical natural language processing. NLTK is written in Python and distributed under the GPL open source license. Over the past three years, NLTK has become popular in teaching and research. We describe the toolkit and report on its current state of development. Lexical resources such as the word net along with the text processing libraries for the classification stemming, tokenization, parsing and semantic reasoning. These are active. These are computational linguistics. This natural language tool kit is useful for linguistics, engineers, students, researchers and educators. This tool kit is available for Windows, Mac OS, and Linux operating systems. NLTK is free, open source.

4. Methodology

The input is given as a text file of any size and contains any number of words to the text summarizer then the text segmentation is done. Under text segmentation stop words (The set of words which are present in the stop words package if we want any specific words we are able to add those words) are removed and stemmer is used. After the text segmentation the algorithms such as Sentence Positioning and sentence scoring algorithms are used. These algorithms are described below of this paper. Some sentences will be repeated which is known as redundancy. The repetition of the statements is to be eliminated so redundancy removal is used. At last the summary for the given input is generated.

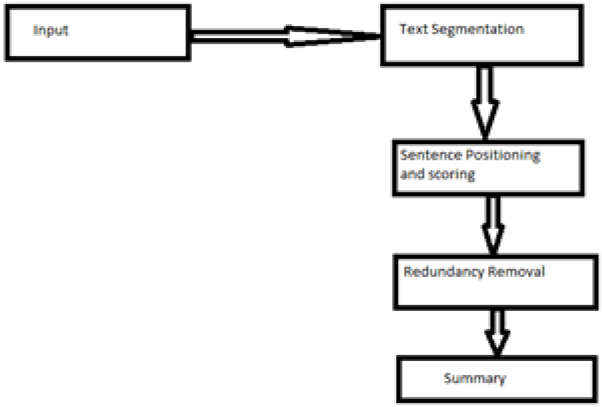


Figure 1: Methodology

5. Implementation

The whole project is divided into four modules. Among the modules the first module is Text

Segmentation. Three main steps take place in this module are:

Lemmatization:It is a process of converting the word to its root form. For example words like says, said belongs to the root word say.

Stemmer: It is similar to a lemmatized. Stemmer stems the words but a lemmatize will convert the word to its root form. For example laughed, laughing words will be stemmed as laugh

Stop Words Removal: Stop words are removed or eliminated in this step which means that continuously repeated words are removed.

The second module is Sentence Ranking. Since the sentences are tokenized the sentences are ranked using the Sentence Positioning and Ranking algorithms.

Sentence Position: The position influences the importance of the sentence. In generally the most important sentence comes at the beginning of the paragraph.

Sentence Resemblance to the Title: This means that there is a vocabulary overlap between the various sentences and the title of the document.

Lexical Similarity: Lexical Similarity is based on the assumption that the important sentences are identified using the strong chains.

5.1 Text Summarizer design

ALGORITHM:

1. Importthe different packages from NLTK and others like sent_tokenize, work_tokenize, storpwords, defaultdict, punctuation, nlargest

2. Initialize the text summarizer. Words that have a frequency term lower than min_cut or higher than max_cut will be ignored.

3. Compute the frequency of each of word.

Input: word_sent, a list of sentences already tokenized.

Output:

Frequency, a dictionary where frequency[w] is the frequency of w.

4. Return a list of n sentences which represent the summary of text.

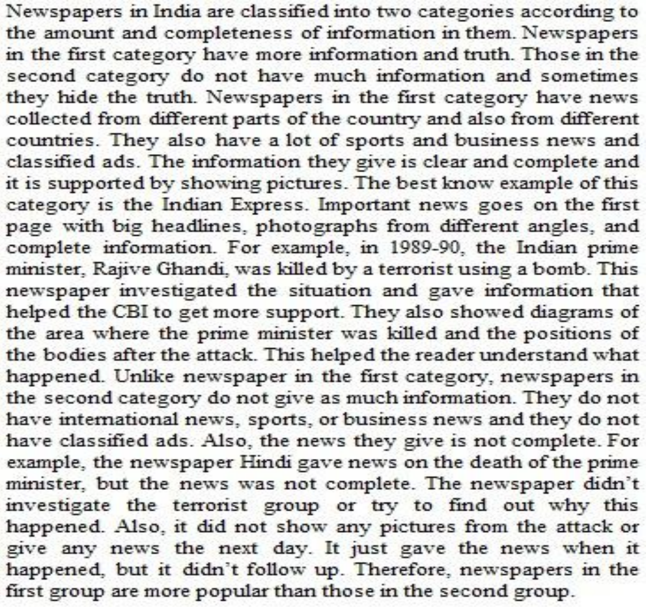
5. Print the title.

6. Return the first n sentences with highest ranking.

6. Experimental Results

In this text summarization input is given as text document with an extension ".txt" After successful compilation the execution is to be done in command prompt of Kubuntu operating system. When the command prompt appears the output can be viewed by giving the file name with an extension of ".py" since it is code is written in python.

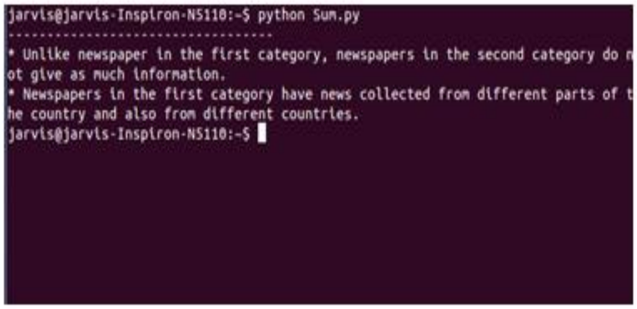
Input:



Newspapers in India are classified into two categories according to the amount and completeness of information in them. Newspapers in the first category have more information and truth. Those in the second category do not have much information and sometimes they hide the truth. Newspapers in the first category have news collected from different parts of the country and also from different countries. They also have a lot of sports and business news and classified ads. The information they give is clear and complete and it is supported by showing pictures. The best know example of this category is the Indian Express. Important news goes on the first page with big headlines, photographs from different angles, and complete information. For example, in 1989-90, the Indian prime minister, Rajive Ghandi, was killed by a terrorist using a bomb. This newspaper investigated the situation and gave information that helped the CBI to get more support. They also showed diagrams of the area where the prime minister was killed and the positions of the bodies after the attack. This helped the reader understand what happened. Unlike newspaper in the first category, newspapers in the second category do not give as much information. They do not have international news, sports, or business news and they do not have classified ads. Also, the news they give is not complete. For example, the newspaper Hindi gave news on the death of the prime minister, but the news was not complete. The newspaper didn't investigate the terrorist group or try to find out why this happened. Also, it did not show any pictures from the attack or give any news the next day. It just gave the news when it happened, but it didn't follow up. Therefore, newspapers in the first group are more popular than those in the second group.

Figure 2: Input Text Document

Output:



```
jarvis@jarvis-Inspiron-N5110:~$ python Sum.py
----------------------------------
* Unlike newspaper in the first category, newspapers in the second category do n
ot give as much information.
* Newspapers in the first category have news collected from different parts of t
he country and also from different countries.
jarvis@jarvis-Inspiron-N5110:~$ █
```

Output of the text document that is taken as input in a .txt file and execute it using Python.

Figure 3: output after summarizing the document

7. Conclusions and Future Work

We have employed new packages of python in text summarizer. As expected, the summarizer works best on all the text documents. Our approach exploits the best out of many possibly complimentary techniques.

In future we plan to design a summarizer called **Compendium Generator**which gives the summary of an IEEE paper and also checks the coherence of the different set of IEEE papers.

8. References

[1] I. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer", In Proceedings of the 18th ACMSIGIR Conference, pages 68-73, 1995.

[2] Lal, Partha. "Text Summarization.," (2002).

[3]Goldstein, J. , Mittal, V. , Carbonell, J. , and Kantrowitz, M. (2000, April). Multi-document summarization by sentence extraction. In Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization-Volume 4 (pp. 40-48). Association for Computational Linguistics.

[4] R. Ferreira, L. de Souza Cabral, R. D. Lins, G. P. e Silva, F. Freitas, G. D.Cavalcanti, R. Lima, S. J. Simske, and L. Favaro, “Assessing sentencescoring techniques for extractive text summarization,” *Expert Systemswith Applications*, vol. 40, no. 14, pp. 5755 – 5764, 2013.

[5]YunFei Yi, Cheng Hua Li, Lijun Liu, Wei Song,Shuai Liu ”Machine Learning Algorithms with Co-occurrence based Term Association for Text Mining”

[6]R. Ferreira, L. de Souza Cabral, R. D. Lins, G. P. e Silva, F. Freitas, G. D.Cavalcanti, R. Lima, S. J. Simske, and L. Favaro, “Assessing sentencescoring techniques for extractive text summarization,” *Expert Systemswith Applications*, vol. 40, no. 14, pp. 5755 – 5764, 2013.

[7]A. Nenkova and R. Passonneau, "Evaluating Content Selection in Summarization: The Pyramid Method," In Proc. HLT/NAACL 2004.

[8]J. Steinberger and J. Karel, "Evaluation Measures for Text Summarization," In Computing and Informatics, Vol. 28, pp. 1001–1026, Mar. 2009.

[9] H. P. Edmundson, “New methods in automatic extracting,” *J. ACM*,vol. 16, no. 2, pp. 264–285, Apr. 1969.

[10] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design patterns:elements of reusable object-oriented software*. Addison-Wesley Professional,1995.

[11] R. D. Lins, S. J. Simske, L. de Souza Cabral, G. de Silva, R. Lima, R. F. Mello, and L. Favaro, “A multi-tool scheme for summarizing textual documents,” in *Proc. of 11st IADIS International Conference WWW/INTERNET 2012*, July 2012, pp. 1–8.

9. Authors Profile

Dr. M. Suman professor (Signals and Systems) in department of Electronics and Computer Engineering (ECM) has extended his services as HOD in ECM department, K L University. He was awarded with Ph.D. from JNTUH, Hyderabad for the thesis entitled "ENHANCEMENT OF COMPRESSED NOISY SPEECH SIGNAL".He is also the life member of Computer Society of India (CSI).

Tharun Maddu student of Electronics and Computer Engineering (ECM) pursuing 4th year of B.TECH in K L University. My previous research works are based on datamining. The present work is related to NLTK on which the present paper research is done.

A.Shalini student of B.TECH (Electronics and Computer Engineering) at K L University. Her areas of interest includes Computer Networks, Data Base Management System, Web technologies, Data mining. Previously she had done research paper in area of Data mining titled as? Privacy Preserving Data Mining Using LBG and ELBG? She participated in various National and International conferences and seminars related to her Subjects of interest.

K.Bhavana studying IV/IV B-Tech in KL University in the Department of Electronics and Computer Engineering and published a paper on Rouge Access Point in wireless sensor network. The present research area is NLTK.