A Project/Dissertation Review Report

On

TEXT SUMMARIZATION USING NLP

Submitted in partial fulfillment of the requirement for the award of the degree of

BTECH - CSE

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Name of Supervisor : Mr. Kundan Kumar.
Designation (Assistant Professor)

Submitted By

HARSHIT MEHLAWAT, 20SCSE1010943

MD NADEEM SARWAR, 20SCSE1010767

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

ABSTRACT

There is an enormous amount of textual material, and it is only growing every single day. Think of the internet, comprised of web pages, news articles, status updates, blogs and so much more. The data is unstructured and the best that we can do to navigate it is to use search and skim the results. There is a great need to reduce much of this text data to shorter, focused summaries that capture the salient details, both so we can navigate it more effectively as well as check whether the larger documents contain the information that we are looking for. The proposed solution for this problem is a online text summarizer web application which uses abstractive text summarization which produces a paraphrasing of the main contents of the given text, using a vocabulary set different from the original document. Summaries reduce reading time, when researching documents, summaries make the selection process easier. Automatic summarization improves the effectiveness of indexing. Automatic summarization algorithms are less biased than human summarizers. The tools and technologies used in text summarization are Flask, python3, web development (HTML, CSS, BOOTSTRAP), Text summarization API(which uses NLP). One of the future plans may be to apply the topic-focussed summarization framework to news articles or blogs and to extend the work in the machine learning approaches.

List of Tables

List of Figures

Acronyms

B.Tech.	Bachelor of Technology
M.Tech.	Master of Technology
BCA	Bachelor of Computer Applications
MCA	Master of Computer Applications
B.Sc. (CS)	Bachelor of Science in Computer Science
M.Sc. (CS)	Master of Science in Computer Science
SCSE	School of Computing Science and Engineering

Table of Contents

Chapter 1
Introduction

1.1 INTRODUCTION

The Internet is a storehouse of data. Information on news, movies, education, medicine, health, nations, weather, geology, etc. is available on the internet. This could be statistical, numerical, mathematical or text data[1] .Text data is more difficult to interpret due to larger amount of characters. Due to this gigantic amount of information, there must be a system in order to get only the essential parts of the information we access. Text summarization is a way of doing this. Text summarization has been a topic of research and study since decades. Various models have been proposed and tested on different datasets to generate concise summaries. They are compared with different comparison scores. Text summarization can be Extractive or Astractive, single document or multi document. Extractive text summarization is a way of generating summaries by using the same sentences as in the document. ABS is more general and focuses on key concepts of the document. Similarly, single document summarization techniques give summaries of the text of a single document, and multi document generates summaries of multiple documents. Natural language processing (NLP) is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and human language. Natural language processing is a process of developing a system that can process and produce language as good as human can produce. The use of World Wide Web has increased and so the problem of information overload also has increased. Hence there is a need of a system that automatically retrieves, categorize and summarize the document as per users need. Text summarization can be used by various applications; for the entire document or not and for summarizing information searched instance researchers need a tool to generate summaries for deciding whether to read by user on Internet. News groups can use multi document summarization to cluster the information from different media and summarize.

1.1.1 EXTRACTIVE SUMMARIZATION

Extractive text summarization works by selecting a subset of existing words, phrases or sentences from the original text to form summary. Extractive summarization uses statistical approach for selecting important sentences or keyword from document. Extractive summarization uses statistical approach for selecting the important sentences or keyword from document.

1.1.2 ABSTRACTIVE SUMMARIZATION

Abstractive text summarization method generates a sentence from a semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original. It consists of understanding the original text and re-telling it in fewer words.

1.2 FORMULATION OF PROBLEM

There is an enormous amount of textual material, and it is only growing every single day. Think of the internet, comprised of web pages, news articles, status updates, blogs and so much more. The data is unstructured and the best that we can do to navigate it is to use search and skim the results. There is a great need to reduce much of this text data to shorter, focused summaries that capture the salient details, both so we can navigate it more effectively as well as check whether the larger documents contain the information that we are looking for. Be it a researcher, a student, a doctor, a teacher everyone needs summaries of large textual data that are fast and easy to go through and understand, for instance researchers need a tool to generate summaries for deciding whether to read the entire document or not and for summarizing information searched by user on Internet.

1.2.1 TOOLS AND TECHNOLOGY USED

The tools and technology that we will use for the proposed system are:

NLP: Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written -- referred to as natural language. It is a component of artificial intelligence.

TENSORFLOW: TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications.

TEXT SUMMARIZATION API: The summarization API allows you to summarize the meaning of a document extracting its most relevant sentences.

FLASK: Flask is a web framework, it’s a Python module that lets you develop web applications easily. It’s has a small and easy-to-extend core: it’s a microframework that doesn’t include an ORM (Object Relational Manager) or such features.

PYTHON3: Python is an interpreted, object-oriented, high-level programming language with dynamic semantics ..... Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse.

HTML / CSS: HTML (Hyper Text Markup Language) for defining the meaning and structure of web content. Cascading Style Sheets (CSS) is a stylesheet language used to describe the presentation of a document written in HTML.

Chapter 2
Literature Survey

There are many prominent works in Text Summarization from the past few years. Earlier works dealt mainly with Single Document Text Summarization. Now that the technology has increased as well as computing power has increased which paved the path for a faster, more effective and more accurate way of processing documents when compared with the earlier methods.

Niladri Chatterjee, Amol Mittal and Shubham Goyal in [2] proposed an extractive based Text Summarization technique that makes use of Genetic Algorithms. In this paper, they represented the single document as a Directed-Acyclic-Graph. Weight is given to each edge of the DAG-based on a schema explained in the paper. They use an Objective function to express the standard of the summary in terms such as ease of readability (readability factor), how closely sentences are related (cohesion factor) and topic relation factor.

Luhn[3] proposed that the most frequent words represent the most important concept of the text. His idea was to give the score to each sentence based on number of occurrences of the words and then choose the sentence which is having the highest score.

Edmunson[7] proposed methods based on location, title and cue words. He stated that initial few sentences of a document or first paragraph contains the topic information and that should be included in summary. One of the limitation of statistical approach is they do not consider semantic relationship among sentences.

Goldstein [4] proposed a query-based summarization to generate a summary by extracting relevant sentences from a document based on the query fired. The criterion for extraction is given as a query. The probability of being included in a summary increases according to the number of words co occurred in the query and a sentence. Goldstein[4][5] also studied news article summarization and used statistical and linguistic features to rank sentences in the document.

H. Gregory Silber and McCoy [8] developed a liner time algorithm for lexical chain computation. The author follows Barzilay and Elhadad [6] for employing the lexical chains to extract important concepts from the source text by building an intermediate representation.

There is another method for summarization by using graph theory [9]. The author proposed a method based on subject-object-predicate (SOP) triples from individual sentences to create a semantic graph of the original document. The relevant concepts, carrying the meaning, are scattered across clauses. The author [9] suggested that identifying and exploiting links among them could be useful for extracting relevant text.

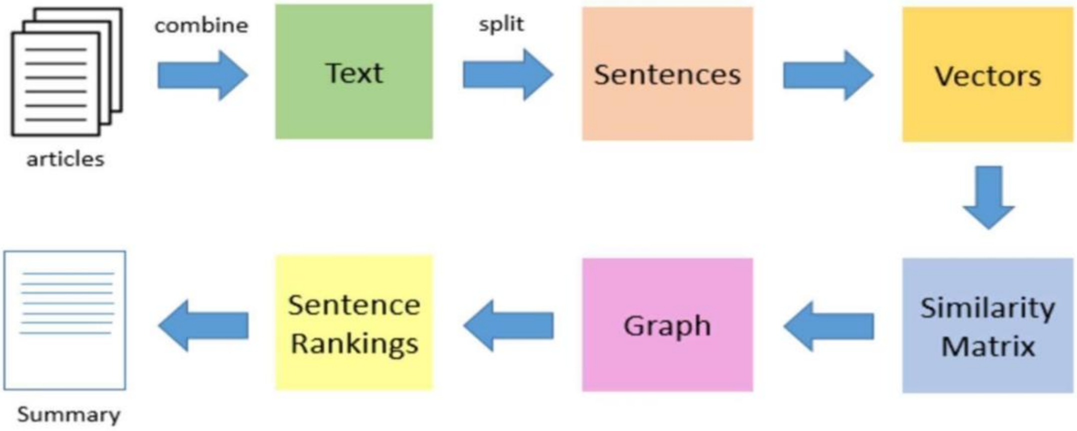
One of the researchers, Pushpak Bhattacharyya [10] from IIT Bombay introduced a Word Net based approach for summarization. The document is summarized by generating a sub-graph from Word-net. Weights are assigned to nodes of the sub-graph with respect to the synsnet using the Word Net. The most common text summarization techniques use either statistical approach or linguistic approach or a combination of both.

COMPARISON OF SUMMARIZATION METHODS:

Type of summarization methods	Subtype	Concept	Advantages	Disadvantages	Application/ Work done
1) Approaches	Abstractive	It is the process of reducing a text document in order to create a summary	Good compression ratio. More reduced text and semantically related summary.	Difficult to compute	SUMMARIST[11]
	Extractive	It consists of selecting important sentences	Easy to compute	Suffers from inconsistency	SURVEY UNIVERSITY[12]

2) Languages	Mono-lingual	Can accept input only with specific language	Need to work with only one language	Cannot handle different languages	FariSum[13]
	Multi-Lingual	Can accept documents in different language	Can deal with multiple language	Difficult to implement	SUMMARIST[11]
3)No. of input document	Single-document	Can accept only one input document	Less overhead	Cannot summarize multiple related documents	Copy & paste system[12]
	Multi-document	Can accept multiple input documents	Multiple document s of same topic can be summarized to single document	Difficult to implement	SUMMONS Designed by Columbia university [12]

BLOCK DIAGRAM:



REFERENCES:

[1] Gonçalves, Luís. 2020. “Automatic Text Summarization with Machine Learning — An overview.” Medium.com. https://medium.com/luisfredgs/au(Gonçalves,2020)Automatic-text summarization-with-machine-learning-an-overview-68ded5717a25.

[2] Niladri Chatterjee, Amol Mittal and Shubham Goyal’s “Single Document Extractive Text Summarization Using Genetic Algorithms” (2012)

[3] Luhn, H.P., 1959. The automatic creation of literature abstracts. IBM J.Res. Develop., 159–165.

[4] Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J., 1999.Summarizing text documents: Sentence selection and evaluation metrics. In: Proc. ACM-SIGIR’99, pp. 121–128.

[5] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, Bahareh Gholamzadeh.A Comprehensive Survey on Text Summarization Systems. 2009 In proceeding of: Computer Science and its Applications, 2nd International Conference.

[6] Barzilay, R., Elhadad, M.Using Lexical Chains for Text Summarization. In Proc. ACL/EACL’97 Workshop on Intelligent Scalable Text summarization, Madrid, Spain,1997, pp. 10–17.

[7] Edmundson, H.P., 1968.New methods in automatic extraction. J. ACM16 (2), 264–285. S.

[8] Silber G.H., Kathleen F. McCoy. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization. Computational Linguistics 28(4): 487-496, 2002.

[9] Kedar Bellare, Anish Das Sharma, Atish Das Sharma, Navneet Loiwal and Pushpak Bhattacharyya.Generic Text Summarization Using Word net. Language Resources Engineering Conference (LREC 2004), Barcelona, May, 2004.

[10] J. Leskovec, M. Grobelnik, N. Milic-Frayling.Extracting Summary Sentences Based on the Document Semantic Graph. Microsoft Research, 2005.

[11] Martin Hassel, Nima Mazdak, “A Persian text summarizer”, International Conference on Computational Linguistics, 2004.