Introduction

In today's rapidly expanding information age, text summarization [1] has emerged as a crucial and useful tool for supporting and analyzing textual material. Humans find it extremely difficult to manually summarize lengthy text documents. A vast amount of textual content can be found in today's world. Automatic text summarization reduces the length of the original text while maintaining its overall meaning and information value.

There are two general approaches to automatic summarization: extraction and abstraction. In extraction-based summarization, content is extracted from the original data, but the extracted content is not modified in any way. Abstractive summarization methods generate new text that did not exist in the original text. [2]

For this project, we chose the extractive summarization method. Extractive summarizing, also known as "surface-level summarization" or "selection-based summarization," extractive summarization can be used to quickly understand a large document's main ideas and key points without reading the entire text. This can save time and enhance efficiency in many industries and applications, such as legal, medical, and business settings. [3]

Extractive text summarization offers a solution by automatically condensing lengthy documents into brief summaries. This project proposal aims to explore and implement various techniques and algorithms in extractive text summarization to develop a robust and efficient system for information retrieval.

We will delve into the realm of natural language processing (NLP) leveraging advancements to create a summarization tool capable of handling diverse text formats and languages. By identifying significant sentences or phrases within a document and assembling them into coherent summaries, our system will empower users to quickly grasp the essence of complex texts, saving time and enhancing productivity.

Through this project, we envision not only developing a state-of-the-art extractive text summarization model but also contributing to the broader landscape of NLP research. By fostering accessibility to information and promoting efficiency in text comprehension, our work aims to address contemporary challenges in information management and knowledge dissemination.

Background and Present State

Text summarization has been a subject of interest in the fields of natural language processing and artificial intelligence for several decades. Traditional methods have relied on heuristic-based approaches or statistical techniques like frequency analysis and graph-based algorithms. However, recent advancements in machine learning, particularly with deep learning architectures, have revolutionized the field. These advances have demonstrated remarkable proficiency in summarization. Despite these advancements, challenges remain in achieving optimal summarization quality. This project seeks to build upon existing research and address these challenges by developing a robust extractive summarization system that balances efficiency with accuracy.

Md. Majharul Haque et al. **[4]** dive deep into some prominent research papers. In their literature review of papers, we can learn about some early works that have been done in this field over the past decade, from as early as 1958's P. B. Baxendale's work to S. P. Yong et al.'s work in 2005. This is useful for understanding the journey; advances were made in the preliminary stage of this field.

Vishal Gupta and Gurpreet Lehal's **[5]** work presents various techniques used for extractive text summarization that can be used. There are some techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) method, Cluster-based method, Graph theoretic approach, Machine Learning approach, LSA Method, Concept-obtained text summarization, Neural networks, fuzzy logic, regression for estimating feature weights, Query based extractive text summarization, Multilingual Extractive Text summarization are briefly explained. It covers most of the available techniques used in the extractive method. Also, this paper explores all the features that can be used for the extractive summarization method. The features briefly mentioned are Content word (Keyword) feature, Title word feature, Sentence location feature, Sentence Length feature, Proper Noun feature, Upper-case word feature, Cue-Phrase Feature, Biased Word Feature, Font based feature, Pronouns, Sentence-to-Sentence Cohesion, Sentence-to-Centroid Cohesion, Occurrence of non-essential information, Discourse analysis.

Dr. Geetha C. Megharaj and Varsha Jituri **[6] have** done summarization using the TF-IDF model. Here, they used the text-rank algorithm for this case.

Yazan Alaya Al-Khassawneh et al. **[7]** used the graph-based method "grpah triangle counting approach." Here, Eigen triangle theorems 1 and 2 are used for this approach. Here Three important stages are involved: i) representing text graphically, where modes in the graph represent sentences in the text and edges represent relationships between the sentences. In this paper, the relationship is the similarity between sentences.

ii) discover the number of triangles in the graph by using an adjacency matrix and De-Morgan's laws, and represent them as a sub-graph of the main graph; iii) by using bit vector values for each edge in iii) it can be decided which nodes are the most important, and then represent every node as a sentence in the text to form the summary.

Acharya and Swapnil **[8]** used an unsupervised machine learning technique where LSA, LDA, and k-means clustering machine learning algorithms were used. And the generated summaries were compared to a reference summary using the ROUGE-N metric.

Mora et al. **[9]** build a summarization system using a graph-based approach that employs Wikipedia concepts to determine the key sentences using the weighted iterative ranking algorithm based on a variation of the HITS algorithm. A generalized bipartite graph framework with the inclusion of concepts ensures coverage; the use of nested-level relationships between sentences and concepts aids in better capturing information; and the weighted iterative ranking algorithm promotes coherency.

Abdelaleem et al. **[10]** used a neural network for summarization. Here, this training is done by using three methods: MLP, PNN, and TDNN. They use a three-layered feed-forward neural network; the number of epochs is 1000 epochs and the cross-validation is terminated after 100 epochs.

Babar **[11]** finds the rank of the system based on word and sentence features. These significant text features are extracted from a given document, and the decision model determines the degree of importance of each sentence based on its rated features. Here, the decision module is modeled using Fuzzy Inference system.

Aristoteles et al.'s **[12]** research involved the text feature weighting of Indonesian text by using binary logistic regression algorithms. Features of the text using text features, where eleven text features are used are sentence position, positive and negative keywords, similarity between sentences, sentences that resemble the title sentence, sentences containing names of entities, sentences that contain numeric data, length of sentence, the connection between sentences, and the sum of the weight of the connection between sentences and sentence semantics. They optimized the summarization text by using the binary logistic regression algorithm and the influence of the eleven features text by using binary logistic regression algorithms.

Aims, Objectives and Possible Outcomes

The proposal work will be carried out with a view to achieving the following objectives:

I. To design a system for summarizing text.
II. To reduce the time and effort spent reading a large document. Thus increasing efficiency.

Outline of Methodology

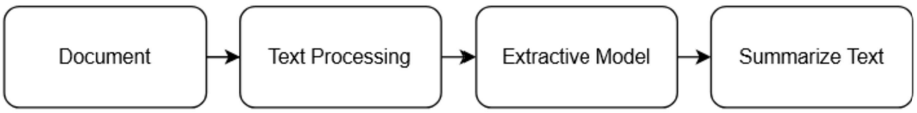

Figure 1: System Architecture for proposed system.

For implementing the extractive summarization system, it has three steps:

Document:

In this stage, a document will be uploaded or imported into the system as the input file, which will be summarized by this proposed system.

Text Processing:

In this step, the input document's text will be cleaned, tagged, ranked, and prepared for the next step.

Extractive Model:

In this phase, the processed text will be inputted into the extractive model, which will give us a summarized output.

Summarized Text:

Finally, we get a summarized output of the large input text document.

References

[1] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.1-12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008.

[2] Wikipedia contributors. "Automatic summarization." *Wikipedia, The Free Encyclopedia*.

[3] Puja Chaudhury, "Getting to the Point: The Benefits of Extractive Summarization", https://catplotlib.medium.com/in-the-field-of-natural-language-processing-nlp-summarization-plays-a-crucial-role-in-reducing-519af0432d96

[4] Md. Majharul Haque, Suraiya Pervin, and Zerina Begum, "Literature Review of Automatic Single Document Text Summarization Using NLP", International Journal of Innovation and Applied Studies ISSN 2028-9324 Vol. 3 No. 3 July 2013, pp. 859-862.

[5] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques" JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010, pp. 260-265

[6] Dr. Geetha C Megharaj, Varsha Jituri, 2022, "TFIDF Model based Text Summerization", INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) RTCSIT – 2022 (Volume 10 – Issue 12), p. 1

[7] Yazan Alaya AL-Khassawneh*, Naomie Salim , Obasa Adekunle Isiaka "Extractive Text Summarisation using Graph Triangle Counting Approach: Proposed Method", 1st International Conference of Recent Trends in Information and Communication Technologies, pp. 304-308

[8] Acharya, Swapnil, "Extractive Text Summarization Using Machine Learning" (2022). Culminating Projects in Computer Science and Information Technology. 39, pp. 18-28

[9] Gopalan, Chitrakala, Moratanch, N., Ramya, B., Raaj, C., Divya, B.. (2018). "Concept-Based Extractive Text Summarization Using Graph Modelling and Weighted Iterative Ranking". 149-160. 10.1007/978-981-10-4741-1_14.

[10] Abdelaleem, Nadeen & Salem, Rashed & Mohamed, Nadeen & diaa.S.AbdElminaam,. (2023). Extractive Text Summarization Using Neural Network. 13119-13131.

[11] Samrat Babar, "Improving Text Summarization Using Fuzzy Logic", Department of Computer Science and Engineering RAJARAMBAPU INSTITUTE OF TECHNOLOGY Rajaramnagar, Islampur - 415 414, pp. 31-32

[12] Aristoteles, Aristoteles & Wibowo, Eko. (2014). Text Feature Weighting for Summarization of Documents Bahasa Indonesia by Using Binary Logistic Regression Algorithm. 29. pp. 30-32