JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

TEXT SUMMARIZATION USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING

[1] Dr. R. S. Khule, [2]Trunali Avhad, [3]Sakshi Aherrao, [4]Pranjali Chavan

[1]Professor, Information Technology Department, Matoshri College of Engineering and Research Centre, Nashik
[2,3,4]Student, Information Technology Department, Matoshri College of Engineering and Research Centre, Nashik

Abstract: This proposed system focuses on the development and implementation of a sophisticated text summarization system using Natural Language Processing (NLP) and Machine Learning, specifically leveraging the Latent Semantic Analysis (LSA) algorithm. Text summarization is a critical aspect of information retrieval and comprehension, enabling users to grasp the essence of large volumes of text quickly and efficiently. Despite the advancements in NLP and ML, there remains a gap in the implementation of a comprehensive summarization system that incorporates LSA. Our project aims to bridge this gap by designing and deploying a novel system that harnesses the power of LSA to extract latent semantic relationships within textual data, thereby generating concise and meaningful summaries. The proposed system will not only enhance the efficiency of information processing but also contribute to the existing body of knowledge in the field of text summarization. Through the integration of LSA, the project seeks to capture the underlying semantic structure of documents, allowing for a more nuanced and contextually relevant summarization. As we embark on the implementation phase, the objective is to address the current absence of a fully realized LSA-based summarization system, providing a valuable tool for researchers, businesses, and individuals dealing with vast amounts of textual information. This project is poised to make a significant contribution to the advancement of text summarization techniques, highlighting the potential of LSA within the broader landscape of NLP and ML applications.

Index Terms **- Summarization, Natural Language Processing (NLP), Machine Learning (ML), Latent Semantic Analysis (LSA), Information Retrieval, Textual Data, Implementation.**

I. INTRODUCTION

Summarization using Natural Language Processing (NLP) and Machine Learning has become a critical area of research and development in the field of artificial intelligence. In an era inundated with vast amounts of textual information, the need for automated summarization techniques has surged to facilitate efficient information retrieval and comprehension. Among the myriad of algorithms employed for text summarization, Latent Semantic Analysis (LSA) stands out as a prominent method that harnesses the power of dimensionality reduction to capture the underlying semantic structure of textual data. At its core, NLP revolves around the intersection of computer science and linguistics, aiming to enable machines to comprehend, interpret, and generate human-like language. Summarization, a subfield of NLP, seeks to distill the essential information from a given text while preserving its core meaning. This process can be extractive, involving the selection of key sentences or phrases from the original text, or abstractive, where the algorithm generates a summary in its own words. Machine Learning techniques play a pivotal role in enhancing the effectiveness of these summarization processes, with algorithms like LSA making notable contributions.

Latent Semantic Analysis is a mathematical approach to uncovering the hidden relationships between words and concepts within a document corpus. LSA operates on the premise that words with similar meanings tend to appear in similar contexts. By representing a document-term matrix in a high-dimensional space, LSA applies singular value decomposition to reduce dimensionality and identify the latent semantic structure. This process enables LSA to capture the underlying meaning of words and documents, facilitating more nuanced and contextually rich summarizations. LSA's application in text summarization involves transforming the document-term matrix into a lower-dimensional space, where each dimension corresponds to a latent semantic feature. By retaining only, the most significant dimensions, LSA effectively filters out noise and irrelevant information, focusing on the essential semantic content. This approach allows LSA to capture the semantic relationships between words and phrases, facilitating a more nuanced understanding of the text.

One of the key advantages of LSA in text summarization is its ability to handle synonymy and polysemy. Synonymy, the existence of multiple words with similar meanings, and polysemy, where a single word has multiple meanings, can pose challenges for traditional summarization algorithms. LSA addresses these challenges by grouping words with similar semantic meanings, providing a more robust foundation for content extraction. LSA contributes to improving the coherence and readability of generated summaries. By incorporating the latent semantic structure, LSA ensures that the selected sentences or phrases in the summary are not only relevant but also contextually cohesive. This enhances the overall quality of the summary, making it more comprehensible and representative of the original text.

Despite its efficacy, LSA does have some limitations. The algorithm relies heavily on the assumption of a linear relationship between terms and may struggle with capturing more complex semantic nuances. Additionally, LSA may not perform optimally when applied to short or highly specialized texts.

II. LITERATURE SURVEY

This paper [1] Latent Semantic Analysis emerges as an innovative approach for distilling the fundamental components of a text corpus, initially applied in the realms of categorization and information retrieval. Its transformative outcomes have elevated LSA beyond mere text analysis, resembling human cognitive processes. In the present study, the utilization of LSA extends to gauging the similarity degree among syslog messages by uncovering concealed relationships within them. Through the examination of authentic syslog message samples, it is demonstrated that LSA effectively discerns the most interconnected messages based on topics. This application presents an alternative to intricate event correlation systems, obviating the necessity for signature or rule set definitions and extensive expertise in configuration. The findings underscore LSA's versatility and efficacy in facilitating nuanced analyses, transcending its conventional applications in text processing.

In the exploration of automatic text summarization challenges, the focus in this work [2] has been on update summarization, a task that involves distilling refined messages from a compilation of new articles, assuming the reader's prior engagement with preceding articles. This reference examines several cutting-edge approaches to extract update summarization, with a specific emphasis on an LSA-based method. Through a thorough analysis of the LSA-based framework, improvements were made to enhance its accuracy. Two key enhancements were introduced. Firstly, the utilization of the TOPIC SIGNATURE algorithm facilitated the extraction of novel information from terms, enriching the evaluation of the topic's novelty score and thereby increasing accuracy. Secondly, the exclusion of the least novel and significant topics during the summary generation process contributed to an overall improvement in summary quality. The evaluation results, as demonstrated in the context of the Text Analysis Conference (TAC) 2008 update summarization task, underscore the validity of these modifications.

Document summarization, a task within the realm of natural language processing (NLP), involves condensing lengthy textual data into concise and coherent summaries that encapsulate all pertinent information present in the document [3]. The specialized branch of NLP dedicated to this task is known as automatic text summarization. Automatic text summarizers play a crucial role in transforming extensive textual documents into brief and well-articulated summaries. Two primary approaches to text summarization employed by automatic text summarizers are extractive and abstractive methods. This research presents an experimental comparison focused on the extractive text summarizer's effectiveness in summarizing text. In parallel, topic modeling, another NLP task, concentrates on extracting relevant topics from textual documents. A method within this domain is Latent Semantic Analysis (LSA), utilizing truncated Singular Value Decomposition (SVD) to extract pertinent topics from text. This study demonstrates an experiment where the proposed research methodology involves summarizing extensive textual documents using LSA for topic modeling. The approach incorporates a TFIDF keyword extractor for each sentence in the text document and utilizes a BERT encoder model to encode sentences, thereby retrieving the positional embedding of topic word vectors. The algorithm proposed in this paper attains a score surpassing that of text summarization achieved through Latent Dirichlet Allocation (LDA) topic modeling. This comparison underscores the efficacy of the proposed algorithm in enhancing the process of summarizing textual information.

Automatic text summarization plays a crucial role in Natural Language Processing (NLP), a subset of the broader field of Artificial Intelligence. The increasing reliance on the internet across various aspects of life has propelled the widespread adoption of text summarization techniques. This research article [4] focuses on applying statistical text summarization methods to Gujarati, a resource-poor South Asian language. The study employs TF-IDF, Latent Semantic Analysis (LSA), and Latent Dirichlet Allocation (LDA) on a custom dataset, evaluating the generated summaries using Rouge scores at compression ratios of 10%, 20%, and 30%. The assessment includes Rouge-1, Rouge-2, Rouge-w, and Rouge-l metrics, with LDA exhibiting the highest Rouge score among the methods investigated. The results are presented in a tabular format, featuring individual Rouge scores and an average score across all methods. The primary objective of this article is to analyze the performance of unsupervised methods in automatic text summarization for the Gujarati language without employing any pre-processing techniques. The selection of sentences is based on a concept-oriented approach utilizing external information. Title matching within the main text is addressed through a topic-based idea, where identical title words receive higher grades, influencing their inclusion in the summary. The cluster-based method organizes comparable sentences depending on the topic, requiring specified cluster counts. The graph-based method relies on the similarity notion, comparing the similarity of words and selecting the best phrases based on these results.

III. METHODOLOGY

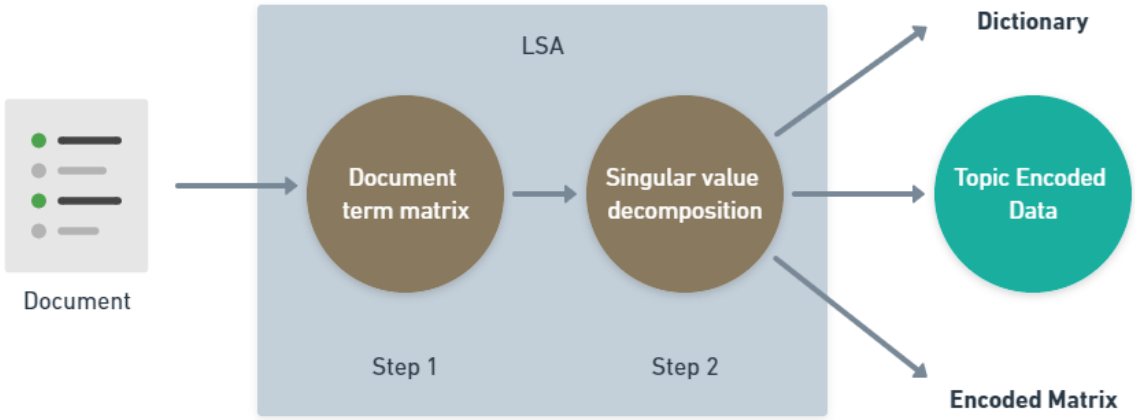

Figure 1 System Architecture

The proposed solution offers various benefits, such as reducing the reading time for lengthy online product documents and providing summaries of product reviews to simplify the selection process. Extractive summarization involves creating a summary based solely on the text's existing content, akin to noting down the key points without modifying them and rearranging the order for coherence. The algorithm for LSA consists of three major steps:

- **Input matrix creation:** The input document is represented as a matrix to understand and perform calculations on it. Thus, a document term matrix is generated. Cells are used to represent the importance of words in sentences. Different approaches can be used for filling out the cell values. There are different approaches to filling out the cell values.

- **Singular Value decomposition (SVD):** In this step, we perform the singular value decomposition on the generated document term matrix. SVD is an algebraic method that can model relationships among words/phrases and sentences. The basic idea behind SVD is that the document term matrix can be represented as points in Euclidean space known as vectors. These vectors are used to display the documents or sentences in our case in this space. Besides having the capability of modeling relationships among words and sentences, SVD has the capability of noise reduction, which helps to improve accuracy.

- **Sentence Selection:** Using the results of SVD different algorithms are used to select important sentences. Here we have used the Topic method to extract concepts and sub-concepts from the SVD calculations which are called topics of the input document. These topics can be sub-topics, and then the sentences are collected from the main topics.

The high number of common words among sentences indicates that the sentences are semantically related. The meaning of a sentence is decided using the word it contains and the meaning of words is decided using the sentences that contain the word. Dictionary and Encoding matrix are the by-products obtained during the execution of LSA processing. A dictionary is a set of all words that occur at least once in our document. While the encoded data represents words of our sentences in terms of their individual strengths. This strength helps us determine the exact effect of each word in our sentence/document.

VI. TRADITIONAL APPROACHES VS MACHINE LEARNING APPROACH

Traditional approaches to text summarization often rely on simple methods such as extracting the initial and concluding sentences or utilizing statistical measures like term frequency and sentence length. While these methods provide a basic overview, they often struggle to capture the nuanced relationships between words and concepts within the text. Additionally, traditional approaches may overlook important contextual information, resulting in less accurate and less informative summaries. In contrast, the Latent Semantic Analysis (LSA) algorithm represents a more advanced approach. LSA employs mathematical and algebraic techniques, specifically Singular Value Decomposition (SVD), to uncover the latent semantic structure within a document. This allows LSA to consider the underlying relationships among words and phrases, providing a more sophisticated understanding of the text. As a result, LSA-based summarization tends to produce summaries that are not only concise but also more contextually relevant and information-rich compared to summaries generated by traditional methods.

V. CONCLUSION

The implementation of the Latent Semantic Analysis (LSA) algorithm for text summarization presents a substantial advancement in the field, offering a sophisticated means to distill key information from documents. By leveraging mathematical techniques like Singular Value Decomposition, LSA captures latent semantic structures, allowing for a nuanced understanding of relationships among words and sentences. The algorithm's ability to discern context and meaning contributes to the generation of concise and contextually relevant summaries. Looking ahead, future endeavors could explore refining LSA through the integration of deep learning methodologies, enhancing its adaptability to diverse linguistic contexts and document types. Additionally, research may

focus on addressing challenges such as handling multimedia content and further optimizing LSA's performance for real-time summarization applications. Continued exploration of advanced algorithms and interdisciplinary collaboration holds the potential to push the boundaries of text summarization, ultimately enhancing information extraction and accessibility in various domains.

REFERENCES

[1] G. Slomovitz, "Latent Semantic Analysis (LSA) for syslog correlation," 2017 International Conference on Electronics, Communications and Computers (CONIELECOMP), Cholula, Mexico, 2017, pp. 1-4, doi: 10.1109/CONIELECOMP.2017.7891819.

[2] Guo-Hua Wu and Yu-Tian Guo, "An enhanced LSA-based approach for update summarization," 2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 2015, pp. 493-497, doi: 10.1109/ICCWAMTIP.2015.7494038.

[3] H. Gupta and M. Patel, "Method Of Text Summarization Using Lsa And Sentence Based Topic Modelling With Bert," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 511-517, doi: 10.1109/ICAIS50930.2021.9395976.

[4] Rahul, S. Adhikari and Monika, "NLP based Machine Learning Approach for Text Summarization," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 535-538.

[5] P. Raundale and H. Shekhar, "Analytical study of Text Summarization Techniques," 2021 Asian Conference on Innovation in Technology (ASIANCON), 2021

[6] X. -y. Jiang, X. -Z. Fan, Z. -F. Wang and K. -L. Jia, "Improving the Performance of Text Categorization Using Automatic Summarization," 2009 International Conference on Computer Modeling and Simulation, 2009

[7] K. D. Garg, V. Khullar and A. K. Agarwal, "Unsupervised Machine Learning Approach for Extractive Punjabi Text Summarization," 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), 2021

[8] M. Ji, R. Fu, T. Xing and F. Yin, "Research on Text Summarization Generation Based on LSTM and Attention Mechanism," 2021 International Conference on Information Science, Parallel and Distributed Systems (ISPDS), 2021

[9] K. S, S. R, S. R and T. S V, "Survey on Automatic Text Summarization using NLP and Deep Learning," 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS), Bangalore, India, 2023, pp. 523-527, doi: 10.1109/ICAECIS58353.2023.10170660.

[10] K. D. Garg, V. Khullar and A. K. Agarwal, "Unsupervised Machine Learning Approach for Extractive Punjabi Text Summarization," 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2021, pp. 750-754, doi: 10.1109/SPIN52536.2021.9566038.

[11] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.