International Journal of Scientific Research and Engineering Development-– Volume 5 Issue 3, May-June 2022

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Survey of Extractive Text Summarization Techniques

Prajakta Mane, Snehal Sarangi
(Department of Computer Engineering
D. Y. Patil College of Engineering
Pune, India
Email:prajakta916mane1@gmail.com, ssarangi@dypcoeakurdi.ac.in)

----------**********----------

Abstract:

The amount of text data on the internet and in archives of news articles, scientific studies, legal documents, and even online product reviews has increased dramatically in recent years. The process of extracting important information from a text document is known as text summarization. In this method, the user is given a succinct overview of the extracted data in the form of a summarised report. Abstractive and extractive summarization are two types of text summarization techniques.The major extractive text summarization approaches have been reviewed in this work. Extractive text summarization algorithms are interpreted in this paper with a less redundant summary,highly sticky, coherent, and depth information.

Keywords — **Text Summarization, Extractive Summary, NLP Methods.**

----------**********----------

I. INTRODUCTION

In the current fast-paced emergent information age, text summarization has evolved into a critical and appropriate engine for supporting and illustrating text content. It is quite difficult for humans to physically summarise large amounts of text. On the internet, there is a multitude of textual content. However, the internet frequently provides more data than is required. As a result, a two-fold challenge arises: finding acceptable papers in a sea of them, as well as fascinating a large volume of important data. The goal of automatic text summarising is to condense the original text into a precise version that retains the report's content and overall meaning. The fundamental benefit of text summarising is that the user's reading time is minimised. A wonderful text summary system should reproduce the document's many themes while minimising repetition. Abstractive and extractive summarising are the only two approaches of text summarization that are publicly available [1]. After interpreting and examining the text, abstractive summarization generates a new shorter text that delivers the most significant information from the original one, utilising advanced natural language algorithms,the end summary generated does not include exact sentences as mentioned in text. Whereas Extractive summaries provide a summary after interpreting and examining the text, and include parts of the sentence from the original text. This document highlights the approaches used for extractive summarization.

This paper is organized as follows. Section 2 depicts about the steps for extractive text summarization, Section 3 describes extractive text summarization methods,Section 4 is the conclusion.

II. TASKS IN EXTRACTIVE SUMMARIZATION

A. Intermediate Representation Of The Input Text

Topic representation and indicator representation are the two sorts of representation-based techniques.The text is converted into anintermediate representation, and the text's topic is interpreted in topic representation. The techniques utilised for this are separated into frequency-driven approaches, topic word approaches, latent semantic analysis, and Bayesian topic models based on their complexity. Whereas in,Indicator representation every sentence is described as a collection of formal properties (indicators) of relevance, such as sentence length,position in the document, the presence of specific phrases, and so on.

B. Sentences Scoring Based On The Representation

When the intermediate representation is constructed, each sentence is assigned an importance score. In topic representation approaches, a sentence's score measures how well
it communicates some of the text's most essential themes.The score is calculated using evidence from various weighted indicators in indicator representation.

C. Choosing of a summary made up of several sentences

To create a summary, the summarizer algorithm selects the top k most important sentences. To pick crucial sentences,some systems use greedy algorithms, while others turn sentence selection into an optimization problem in which a set of sentences is chosen with the constraint of maximising overall relevance and coherency while minimising redundancy.

III. EXTRACTIVE SUMMARIZATION APPROACHES

Topic Representation Approaches

A. Topic Words

This frequent technique seeks to find terms in the input document that describe the topic. The use of the log-likelihood ratio test to find explanatory words, referred to as the "subject signature," was an advancement of Luhn's original proposal[2]. In general, a sentence's relevance can be calculated in two ways: as a function of the number of subject signatures it contains, or as a proportion of the subject signatures in the sentence. The first strategy favors longer sentences with more words, whereas the second assesses the density of the topic words. All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

B. Frequency-Based Methods

The frequency of words is used as a measure of importance in this method. Word probability and TFIDF are the two most frequent strategies in this category (Term Frequency Inverse Document Frequency). The probability of a word w
is calculated by dividing its number of occurrences, f (w), by the total number of words in the input (which can be a single document or multiple documents) [3]. The summary includes words that are most likely to describe the document's topic.

$P(w) = f(w)/N$

TFIDF, a more advanced technique, evaluates the relevance of terms in the document(s) and detects very common words(that should be eliminated from consideration) by assigning low weights to words that exist in most texts [4]. TFIDF has given way to centroid-based techniques, which order sentences based on the prominence of a set of attributes.
Following the construction of TFIDF vector representations of documents, the documents that explain the same topic are clustered together and centroids — pseudo-documents made up of terms with TFIDF scores greater than a particular
threshold — are computed. The centroids are then utilised to determine central to the topic sentences in each cluster.

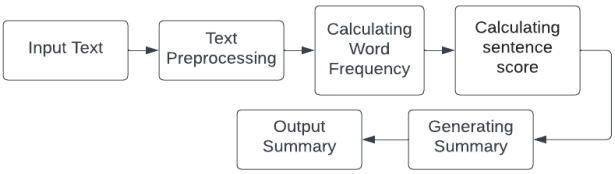

Fig. 1. TFIDF Method

C. Latent Semantic Analysis

LSA is an unsupervised technique for extracting a representation of text semantics from observed words [5]. The first stage is to construct a term-sentence matrix, in which each row represents a word from the input (n words) and each column represents a sentence. The weight of the word i in sentence j, computed using the TFIDF approach, is represented by each element in the matrix. The matrix is then transformed into three matrices using singular value decomposition (SVD):a term-topic matrix with word weights, a diagonal matrix with topic weights in each row, and a topic-sentence matrix.

The outcome of multiplying the diagonal matrix with weights with the topic-sentence matrix is the weight of the topic i in sentence j, which describes how much a phrase represents a topic.

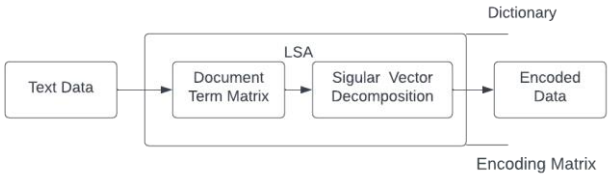

Fig. 2. LSA Processing

It is an effective approach for abstracting out the document's buried context.

D. Discourse Based Method

Perform discourse analysis, discovering the semantic linkages between textual units, to generate a summary, is a logical progression of analysing semantics. Radev began researching cross-document relationships and developed the

Cross-Document Structure Theory (CST) model [6]. If words,phrases, or sentences are semantically connected, they can be linked in his model. CST was definitely useful for document summarising, as well as for treating repetition, complementarity, and inconsistency across the many data sources.

Nonetheless, this strategy has a severe drawback in that the CST relations must be explicitly specified by humans.

E. Bayesian Topic Models

While other approaches lack unambiguous probabilistic interpretations, Bayesian topic models are probabilistic models that can convey information that is lost in other approaches because they describe topics in greater detail.

The purpose of topic modelling of text documents is to infer terms connected to a specific topic and the topics mentioned in a specific document based on a corpus of documents. It's feasible because to Bayesian inference, which uses a combination of common sense assumptions and the results of previous similar events to compute the likelihood of an event. The model is improved over time by going through numerous iterations in which a prior probability updated with observable evidence to obtain a new posterior probability. [7]proposed a Bayesian sentence-based topic model that employed both term-document and term-sentence associations for summarization. Their methodology exceeded several other summarising methods in terms of significance.

Indicator representation approaches

F. Graph Methods

The graph-based method to text summarization is an unsupervised technique in which we use a graph to rank the required sentences or words. The most essential sentences from a single document are extracted using the graphical method.

The relevance of a vertex in a graph is basically determined. To implement text-based ranking, we use unidirected and weighted graphs. Documents or sentences can be represented as nodes in this manner. Edges connect any two nodes that share the same data. Initializing weightage to the nodes of the graph is how sentence scoring is done. These methods, which are influenced by the PageRank algorithm, describe documents as a connected graph, with sentences serving as vertices and edges indicating how similar two sentences are. When the similarity of two phrases is greater than a particular threshold,they are connected using cosine similarity with TFIDF weights for words. The sub-graphs in the graph produce subjects that are covered in the texts, and the key sentences are selected as a result

of this graph representation. Sentences in a sub-graph that are linked to a large number of other sentences are likely to constitute the graph's centre and will be included in the summary. This method can be applied to any language because it does not require language-specific linguistic processing. At the same time, the method's application is limited because it only measures the formal side of the sentence structure and ignores syntactic and semantic information. Graph-based methods are considerably easier to perceive, grasp, and use than other summarising techniques. However, this method can only generate a summary for a single document.

Fig. 3. Graph Reduction

G. Machine Learning

To create a true-to-life summary, machine learning algorithms that treat summarising as a classification problem are increasingly commonly employed, attempting to apply Naive Bayes, decision trees, support vector machines, Hidden Markov models, and Conditional Random Fields.After all,strategies that explicitly assume set dependencies, such as hidden Markov models [8] and conditional random fields[9], are often superior to other methods. However, using supervised learning methods for summarization requires a set of labelled documents to train the classifier, which necessitates.The creation of a corpus. Semi-supervised techniques, which combine a small quantity of labelled data with a big amount of unlabeled data in training, may be a viable option.

IV. CONCLUSION

The rapid expansion of the Internet has resulted in a massive influx of data. Summary of big amounts of text is challenging for humans. In this age of information overload, there is a huge demand for automatic summarising technologies. Various extraction methodologies for single and multi-document summarization were highlighted in this research. Topic representation approaches, frequency-driven methods, graph-based and machine learning techniques were highlighted as some of the most often utilised methodologies. Although it is hard to comprehend all of the many algorithms and approaches in this paper, we believe it gives a solid overview of recent trends and advances in automatic extractive summarising methods and explains the current state of the art in this field.

R EFERENCES

[1] Pooja Raundale, Himanshu Shekhar, "Analytical study of Text Summarization Techniques," Asian Conference on Innovation in Technology (ASIANCON).2021.

[2] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," Computational linguistics, vol. 19, no. 1, pp. 61–74, 1993.

[3] Manoj Kumar, "Frequent Term Summarization Using Word Probability Based Semantic Similarity," IJLTET.2013

[4] Sarika Zaware, Deep Patadiya, Abhishek Gaikwad, Sanket Gulhane, Akash Thakare, "Text Summarization using TF-IDF and Textrank algorithm," 5th International Conference on Trends in Electronics and Informatics (ICOEI) 2021.

[5] Shuchu Xiong, Yihui Luo, "A New Approach for Multi-document Summarization Based on Latent Semantic Analysis," Seventh International Symposium on Computational Intelligence and Design.2014.

[6] Erick Galani Maziero,Maria Lucía del Rosario Castro Jorge, "Revisiting Cross-document Structure Theory for multi-document discourse parsing,"Information Processing and Management.2014.

[7] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong, "Multidocument summarization using sentence-based topic models,"Conference Short Papers. Association for Computational.2009

[8] John M Conroy,Dianne P. O'leary, "Text summarization via hidden Markov models,"Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.2001.

[9] Dou Shen,Jian-Tao Sun,Hua Li, Qiang Yang, "Document Summarization Using Conditional Random Fields.,"Conference: IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence.2007.

