International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Multi-Document Summarization Techniques: A Comparative Study

Mr. Saish Hanuman Mahale
Masters of Engineering,
Department of Information Technology, Goa College of Engineering, Ponda, Goa 403401
saishmahale4165@gmail.com

Mr. Amogh Sanzgiri
Assistant Professor,
Department of Information Technology, Goa College of Engineering, Ponda, Goa 403401
amoghs@gec.ac.in

Keywords: Text Summarization, Multi Document Summarization, TextRank, LexRank, Centroid Based Summarization, Maximal Marginal Relevance, K-Means Clustering.

Abstract— Text summarization is the process of generating a shorter version of a longer text while retaining its most important information. The aim of text summarization is to extract the most important information from a document and present it in a concise and easy-to-read format. There are two main types of text summarization: extractive and abstractive. Multi-document summarization is a text summarization technique that aims to generate a summary from multiple documents instead of just one. There are various algorithms for implementing Multi-document summarization such as TextRank, LexRank, Centroid Based Summarization, Maximal Marginal Relevance, Cluster-based Sentence Selection (CBSS), SumBasic and K-Means Clustering. In this paper, we present an overview of Multi-document summarization and five methods: TextRank, LexRank, Centroid Based Summarization, Maximal Marginal Relevance and K-Means Clustering. Also we have discussed how these algorithms can be evaluated using several metrics such as Precision, Recall, F1 Score, Pyramid Score, Content Selection Score, Sentence Compression Ratio.

I. INTRODUCTION

Text summarization is the process of automatically creating a condensed version of a longer text while retaining its most important information. The goal of text summarization is to make it easier and faster for people to get the key insights or main points from a text without having to read through the entire document.

Multi-document summarization is a type of text summarization that involves the creation of a summary from multiple related documents. This technique is used to provide a concise and coherent summary of a large collection of documents, such as news articles, research papers, or legal documents.

Multi-document summarization can be achieved using various techniques, such as clustering, sentence scoring, and graph-based methods. The goal of multi-document summarization is to provide a summary that contains the most important and relevant information from the collection of documents, while minimizing redundancy and preserving the coherence and readability of the summary.

Multi-document summarization has many practical applications, such as in the fields of journalism, information retrieval, and data analysis. It can help to reduce information overload, save time, and provide a quick overview of large amounts of text data.

In this study we will discuss some of the multi document summarization techniques ie. TextRank, LexRank, Centroid Based Summarization, Maximal Marginal Relevance and K-Means Clustering.

There are two main approaches to text summarization: extractive summarization and abstractive summarization.

Extractive summarization: Extractive summarization involves selecting the most relevant sentences or phrases from the original text and assembling them into a summary. This approach works by analyzing the original text and selecting the most important sentences based on some predetermined criteria. The criteria used for selecting sentences may include sentence length, word frequency, and semantic relevance.

The process of extractive summarization typically involves the following steps:

- Sentence splitting: The original text is split into individual sentences.
- Sentence scoring: Each sentence is scored based on its relevance to the main ideas of the text. This can be done using various techniques such as term frequency-inverse document frequency (TF-IDF), Latent Semantic Analysis (LSA), or neural networks.
- Sentence selection: The sentences with the highest scores are selected to be included in the summary.

Extractive summarization is relatively simple and effective, and it can be used to quickly generate summaries of large volumes of text. However, it may not be able to capture the nuance or context of the original text, and it may produce summaries that are difficult to read and understand.

Abstractive Summarization: Abstractive summarization involves generating a summary that is not limited to the sentences or phrases in the original text. This approach works by analyzing the content of the original text and generating a summary that is similar to a human-written summary. The summary generated by abstractive summarization may not include the same sentences or phrases as the original text, but it will capture the main ideas and key points of the text.

The process of abstractive summarization typically involves the following steps:

- Text understanding: The original text is analyzed to identify the main ideas and concepts.
- Text generation: The summary is generated using natural language generation techniques, such as deep learning-based models, which can generate summaries that are similar to human-written text.

II. WORKING OF MULTI-DOCUMENT SUMMARIZATION

Multi-document summarization involves producing a brief and focused summary that encapsulates the primary aspects and significant details found in a set of multiple documents. The objective is to condense the content of these documents while retaining essential information and ensuring logical coherence.

The working of multi-document summarization involves the following steps:

1. **Document Collection:** The first step is to gather a set of documents related to a specific topic or domain. These documents could be from various sources such as articles, research papers, news reports, or web pages.
2. **Preprocessing:** Once the documents are collected, they undergo preprocessing steps like text cleaning, sentence splitting, tokenization, and removing stopwords. These steps help to prepare the text for further analysis.
3. **Sentence Scoring:** In this step, each sentence in the document collection is assigned a score based on various criteria. Common methods include calculating sentence relevance, frequency, position, or using more advanced techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or graph-based algorithms.
4. **Sentence Selection:** After scoring, a subset of sentences is selected based on their scores. The number of selected sentences depends on the desired length of the summary. Different techniques can be used for sentence selection, such as selecting the top-scoring sentences or using a threshold to include sentences above a certain score.
5. **Redundancy Removal:** Since multiple documents might contain overlapping or redundant information, it is essential to eliminate redundancy in the summary. This can be done by comparing the selected sentences and

removing similar or duplicate content.

6. **Summary Generation:** Finally, the selected sentences are combined to form a coherent and concise summary. Various methods can be used for summary generation, including concatenating the selected sentences, rephrasing to improve readability, or using techniques like extractive or abstractive summarization.

III. METHODS OF MULTI-DOCUMENT SUMMARIZATION

A. TextRank

[1] The research paper titled "TextRank: Bringing Order into Texts" by Rada Mihalcea and Paul Tarau introduces the TextRank algorithm, a graph-based method for keyword extraction and text summarization. The algorithm is inspired by Google's PageRank and applies it to a graph representation of a document, where nodes represent words or phrases, and edges represent relationships between them.

The algorithm iteratively calculates the importance score of each node based on the scores of its neighbors and the strength of their connections. This process identifies the most important keywords or key phrases in the text. The authors also demonstrate the application of TextRank to text summarization, where sentences are represented as nodes in the graph. By applying TextRank, important sentences can be selected to create a summary that captures the main information from the source text. The research paper has made a significant impact on natural language processing, inspiring further advancements in graph-based algorithms for text analysis and summarization.

TextRank, proposed by Mihalcea and Tarau, is a graph-based algorithm used for keyword extraction and text summarization. It represents a text document as a graph, where the nodes represent either individual words or phrases (in keyword extraction) or sentences (in text summarization). The relationships between these nodes are captured by the edges in the graph, which can be weighted based on factors like co-occurrence frequency or semantic similarity.

The algorithm assigns an initial importance score to each node and then iteratively updates these scores based on the scores of neighboring nodes and the weights of the connecting edges. This iterative calculation process continues until the scores converge, resulting in a final set of importance scores for the nodes.

In keyword extraction, nodes with the highest importance scores are considered the most significant keywords or key phrases in the text. These can be used to represent the main themes or important concepts in the document.

For text summarization, the algorithm applies the importance scores to the sentence nodes. Sentences with higher scores are considered more important and are selected to form a summary that captures the essential information from the source document.

TextRank has been widely adopted in the field of natural language processing for its ability to effectively extract keywords and generate summaries by leveraging the graph structure and importance scoring. Its versatility and effectiveness make it a valuable tool in various text analysis and information retrieval applications.

Advantages of TextRank:

Language Independence: TextRank operates solely based on the structural properties of the text, such as sentence similarity and relationships, making it language-independent. It can be applied to texts in various languages without requiring language-specific modifications.

Unsupervised Learning: TextRank is an unsupervised learning approach that does not require training data or manual annotations. It automatically ranks sentences based on their importance, relying on the inherent structure of the text.

Extractive Summarization: TextRank performs extractive summarization, which means it selects and outputs sentences directly from the source text. This approach ensures that the summary contains actual sentences from the original documents, preserving the integrity of the information.

Considers Contextual Similarity: TextRank takes into account the contextual similarity between sentences by analyzing their semantic relationships. It captures the important concepts and themes present in the text, resulting in summaries that reflect the main points of the document.

Scalability: TextRank is computationally efficient and scalable. It can handle large volumes of text and process multiple documents simultaneously. This scalability makes it suitable for summarizing large collections of documents or real-time applications.

Adaptability: The TextRank algorithm can be adapted and extended to incorporate additional features or fine-tune the importance scoring mechanism based on specific requirements. It provides flexibility for customization and experimentation with different parameters.

B. LexRank

The research paper[2] introduces a graph-based algorithm called LexRank for automatic text summarization. The paper was published in 2004 and addresses the challenge of extracting key information from a text and generating concise summaries. LexRank utilizes the concept of lexical centrality, where sentences that contain

important words are considered more salient and likely to be included in a summary.

The authors propose a method that constructs a graph representation of a document, with sentences as nodes and edges representing the similarity between sentences. The similarity between sentences is computed using a cosine measure based on the overlap of content words. The resulting graph is then ranked using an iterative algorithm that assigns importance scores to each sentence based on its centrality and the scores of its neighboring sentences.

The LexRank algorithm is evaluated using the DUC (Document Understanding Conference) 2002 and 2003 datasets, which contain news articles and their human-generated summaries. The results demonstrate that LexRank outperforms other existing summarization approaches, producing summaries that are more informative and closer to the reference summaries.

The paper concludes that LexRank provides an effective and efficient method for text summarization by leveraging graph-based centrality measures. It highlights the significance of incorporating word centrality in the process of extracting salient information from a document and generating coherent and concise summaries.

Advantages of LexRank:

Graph-Based Representation: LexRank represents the document collection as a graph, where sentences are nodes and edges denote the similarity between sentences. This graph-based representation captures the structural relationships among sentences, enabling a more nuanced analysis of their importance.

Centrality-Based Scoring: LexRank uses the concept of centrality to assign importance scores to sentences. Sentences that are similar to many other sentences in the document collection are considered important. This approach ensures that important sentences that capture the core ideas and information are given higher scores.

Redundancy Reduction: LexRank incorporates a redundancy reduction mechanism to avoid including similar or redundant sentences in the summary. By promoting diversity in the selected sentences, LexRank produces concise summaries that avoid repetitive information.

Language Independence: LexRank operates solely on the textual content and does not rely on language-specific features or resources. It can be applied to text documents in various languages without requiring language-specific adaptations.

Evaluation and Interpretability: LexRank's importance scores can be interpreted as the relative importance of sentences within the document collection. This transparency allows for easy evaluation and interpretation of the summarization results.

Adaptability: The LexRank algorithm is adaptable and can be extended to incorporate additional features or variations in the similarity computation. This flexibility allows for customization and experimentation based on specific requirements or domain-specific characteristics.

C. Centroid Based Summarization

[3] "Centroid-based summarization of multiple documents" is a research paper by Dragomir R. Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam, published in 2004. The paper introduces a centroid-based approach for summarizing multiple documents.

The authors address the challenge of summarizing a collection of related documents, where the goal is to generate a concise summary that captures the essential information while minimizing redundancy. They propose a method that constructs a centroid, or representative summary, by identifying sentences that are most similar to the overall content of the documents.

The approach involves several steps. First, the documents are preprocessed to remove stop words and perform stemming. Then, the authors compute the term frequency-inverse document frequency (TF-IDF) weight for each word in the documents, which helps determine their significance. Sentence similarity is then measured using a cosine similarity metric based on the TF-IDF weights.

Next, the centroid is constructed by selecting sentences that are most similar to the centroid itself. This process is performed iteratively, where the centroid is updated at each step. The final summary is generated by ranking and selecting the most representative sentences from the centroid.

The paper evaluates the centroid-based approach using the DUC 2002 and 2003 datasets, which contain multiple documents and corresponding human-generated summaries. The results show that the centroid-based summarization method outperforms several baseline approaches in terms of informativeness and redundancy.

The authors conclude that centroid-based summarization is a promising technique for summarizing multiple documents. It effectively captures the key information and reduces redundancy by identifying sentences that best represent the overall content. The paper highlights the significance of considering the global context and relationships between sentences when

generating summaries from a collection of documents.

Advantages of Centroid Based Summarization:

Conceptual Grouping: CBS uses clustering to group similar sentences together based on their content. This grouping allows for the identification of key themes and topics present in the document collection. By selecting representative sentences from each cluster, CBS ensures that the summary covers a diverse range of important information.

Redundancy Reduction: CBS incorporates redundancy reduction mechanisms, ensuring that similar or duplicate sentences are eliminated from the summary. This helps to create concise and focused summaries by avoiding the repetition of information.

Coherence and Consistency: By selecting sentences from each cluster, CBS ensures that the summary maintains coherence and consistency. The representative sentences are chosen in a way that they form a coherent and meaningful summary, reflecting the main ideas across the document collection.

Domain Adaptability: CBS can be adapted to different domains or document collections by adjusting the clustering parameters and similarity measures. This flexibility allows the algorithm to capture domain-specific concepts and effectively summarize documents from various subject areas.

Language Independence: CBS operates on the content of the text and does not rely on language-specific features. It can be applied to document collections in different languages without requiring language-specific modifications.

User Control: CBS allows for user control over the summary generation process. By adjusting parameters such as the number of clusters or the level of redundancy reduction, users can customize the output summary according to their preferences and requirements.

D. Maximal Marginal Relevance

[4] "Less Is More: Maximal Marginal Relevance as a Summarisation Feature" is a research paper by Jan Frederik Forst, Anastasios Tombros, and Thomas Roelleke, published in 2007. It introduces the concept of Maximal Marginal Relevance (MMR) as a feature for text summarization.

The authors address the challenge of producing informative and diverse summaries by leveraging the notion of relevance. They propose MMR, a measure that balances the informativeness of a sentence with its redundancy to ensure the generated summaries are both concise and diverse.

The MMR measure is computed using two key components: a similarity measure and a diversity measure. The similarity measure captures the relevance of a sentence to the

given query or topic by considering its similarity to previously selected sentences. The diversity measure encourages the inclusion of sentences that are dissimilar to the selected ones, promoting a more comprehensive summary.

To implement MMR, the authors employ a retrieval-based summarization approach. They experiment with different retrieval models, such as vector space models and language models, to compute the similarity between sentences and the query. The MMR score is then used to rank the sentences and select the most relevant and diverse ones for the summary.

The paper evaluates the MMR approach using the DUC 2004 and 2005 datasets, which contain news articles and their corresponding human-generated summaries. The results show that MMR outperforms baseline methods in terms of both informativeness and diversity, producing more concise and well-rounded summaries.

The authors conclude that incorporating MMR as a summarization feature improves the quality of the generated summaries. By balancing relevance and diversity, MMR helps overcome the trade-off between informativeness and redundancy, ultimately leading to more effective summarization. The paper highlights the significance of considering both relevance and diversity when creating summaries, as it enables a more comprehensive representation of the underlying information.

Advantages of Maximal Marginal Relevance:

Information Diversity: MMR promotes information diversity in the generated summary by selecting sentences that are both relevant to the main topic and dissimilar to each other. This helps to avoid redundancy and ensures that the summary covers a wider range of important information.

Reducing Redundancy: MMR includes a redundancy reduction mechanism that prevents the selection of highly similar or redundant sentences. By encouraging the inclusion of diverse and distinct information, MMR produces more concise and focused summaries.

Customizability: MMR allows for customization and parameter tuning based on specific preferences and requirements. The trade-off between relevance and diversity can be adjusted according to the desired level of emphasis on either aspect, providing flexibility in generating summaries tailored to different needs.

User Control: MMR enables user control over the summary generation process. The user can influence the selection of sentences by specifying a query or by providing feedback on the initial summary, allowing for a more interactive and personalized summarization experience.

Evaluation and Interpretability: MMR provides a clear evaluation framework for summarization tasks. By balancing relevance and diversity, it offers a comprehensive evaluation metric that measures the quality and informativeness of the generated summaries.

Extensibility: MMR can be combined with other summarization techniques or algorithms to enhance their performance. It can be integrated into existing systems or used in conjunction with other algorithms to achieve improved summarization results.

E. K-Means Clustering

[5] The research paper titled "An Experiment on Multi-Document Summarization Using K-means Clustering Algorithm" by Chandrali Sarma and Hirakjyoti Sarma addresses the significance of multi-document summarization in managing information overload and extracting concise, meaningful summaries from multiple sources. The authors emphasize the challenges associated with this task and highlight the potential advantages of leveraging clustering algorithms.

The paper introduces the K-means clustering algorithm as a viable approach to group similar sentences based on their distinctive features, forming coherent clusters. The authors elaborate on how the K-means clustering algorithm can be tailored for multi-document summarization by considering sentence similarity and selecting representative sentences from each cluster.

Sarma and Sarma conduct an experiment wherein they apply the K-means clustering algorithm to a dataset comprising multiple documents. They assess the quality of the generated summaries by comparing them against human-generated summaries using evaluation metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation).

The experimental results indicate the effectiveness of the K-means clustering algorithm for multi-document summarization. The authors discuss the merits and limitations of this approach, emphasizing its capacity to capture diverse perspectives from multiple documents while acknowledging the challenges associated with determining the optimal number of clusters.

Advantages of K-Means Clustering:

Grouping Similar Sentences: K-means clustering groups similar sentences together based on their feature similarity. This allows for the identification of clusters that capture different themes or topics present in the document collection. By selecting representative sentences from each cluster, K-means clustering helps to create a diverse and comprehensive summary.

Scalability: K-means clustering is computationally efficient and can handle large volumes of text. It is well-suited for

summarizing collections of multiple documents, making it a scalable solution for multi-document summarization tasks.

Interpretability: K-means clustering provides interpretable results as each cluster represents a specific theme or concept. This allows users to understand the content coverage and main ideas captured in the summary easily.

Language Independence: K-means clustering operates solely based on the feature similarity between sentences, making it language-independent. It can be applied to document collections in various languages without requiring language-specific modifications.

Flexibility and Customization: K-means clustering offers flexibility in terms of defining the features used for clustering, such as TF-IDF or word embeddings. It can be customized and adapted to specific requirements or domain-specific characteristics, allowing for fine-tuning and optimization.

Redundancy Reduction: K-means clustering can incorporate redundancy reduction mechanisms to avoid including similar or duplicate sentences in the summary. This helps to create concise and focused summaries by removing repetitive information.

IV. EVALUATION PARAMETERS

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE measures the overlap between the system-generated summary and one or more human-generated reference summaries. It calculates metrics such as ROUGE-N (N-gram overlap), ROUGE-L (longest common subsequence), and ROUGE-S (skip-bigram overlap) to evaluate the quality of the summary.

Pyramid Score: Pyramid Score compares the system-generated summary with multiple reference summaries in a hierarchical manner. It assigns scores based on the extent to which the summary covers the information present in the reference summaries at different levels of abstraction.

F1 Score: F1 Score is a commonly used metric that measures the balance between precision and recall. It calculates the harmonic mean of precision and recall, reflecting the overall effectiveness of the summarization system.

Content-based Metrics: These metrics evaluate the content quality of the summary by considering factors such as informativeness, relevance, and coverage. They assess the degree to which the summary captures the main points and important information from the source documents.

Readability Metrics: Readability metrics assess the linguistic quality of the summary, considering factors like grammaticality,

coherence, fluency, and readability. These metrics help evaluate how well the summary conveys information in a clear and understandable manner.

V. CONCLUSION

Text summarization is a powerful tool for summarizing large volumes of text. There are two main approaches to text summarization: extractive summarization and abstractive summarization. Extractive summarization involves selecting the most important sentences or phrases from the original text, while abstractive summarization involves generating a summary that captures the main ideas and key points of the text. Both approaches have their advantages and disadvantages, and the choice of approach depends on the specific application and the goals of the summarization process. A comprehensive evaluation of multi-document summarization systems involves utilizing various metrics. These metrics, including ROUGE, Pyramid Score, F1 Score, content-based metrics, and readability metrics, provide insights into different aspects of summary quality, such as content overlap, hierarchical coverage, overall effectiveness, relevance, coherence, and user satisfaction. By considering this diverse range of metrics, a holistic assessment of the summarization system's performance can be achieved, ensuring the generation of informative, coherent, and user-friendly summaries.

REFERENCES

[1] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

[2] Erkan, Gunes & Radev, Dragomir. (2011). LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. Journal of Artificial Intelligence Research - JAIR. 22.10.1613/jair.1523.

[3] Radev, Dragomir & Jing, Hongyan & Styś, Małgorzata & Tam, Daniel. (2004). Centroid-based summarization of multiple documents. Information Processing & Management. 40. 919-938. 10.1016/j.ipm.2003.10.006.

[4] Forst, J.F., Tombros, A., Roelleke, T. (2009). Less Is More: Maximal Marginal Relevance as a Summarisation Feature. In: , et al. Advances in Information Retrieval Theory. ICTIR 2009. Lecture Notes in Computer Science, vol 5766. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04417-5_37

[5] Sarma, Chandrali & Sarma, Hirakjyoti. (2014). An Experiment on Multi-Document Summarization Using K-means Clustering Algorithm.