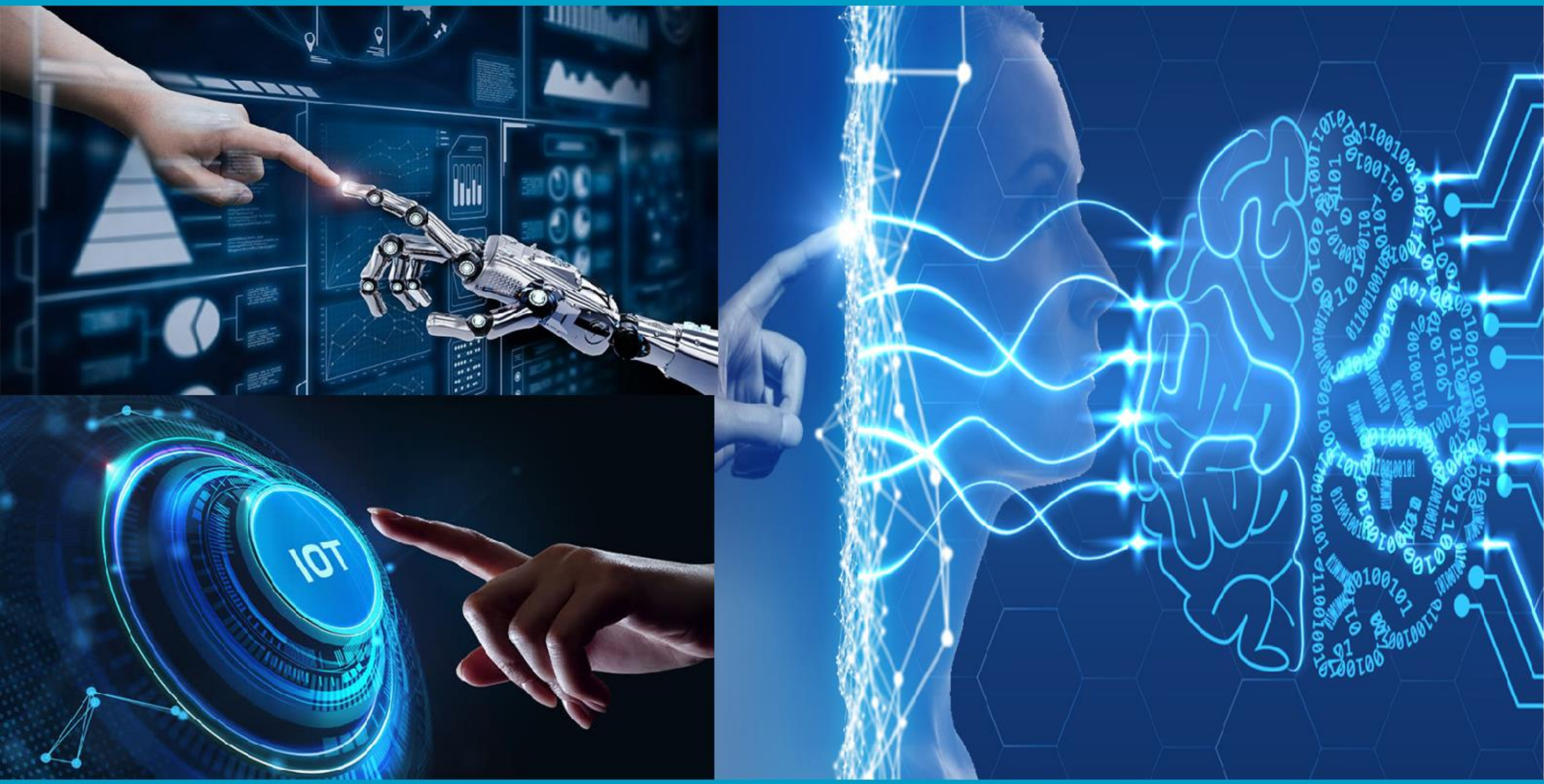
IJMRSETM

e-ISSN: 2395 - 7639



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

Volume 10, Issue 6, June 2023

ISSN
INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.580

☎ +91 99405 72462 +9163819 07438 ✉ ijmrsetm@gmail.com @ www.ijmrsetm.com

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 10, Issue 6, June 2023 |

Multi-Document Summarization: Techniques for Generating Coherent Summaries from Multiple Text Sources

[1] Mr. Mangesh J. Bhandurge, [2]Dr. Prakash Kene

[1]MCA Student, Department of MCA, P.E.S. Modern College of Engineering, Pune, Maharashtra, India

[2]Professor, Department of MCA, P.E.S. Modern College of Engineering, Pune, Maharashtra, India

ABSTRACT: Multi-document summarization is a complex task that involves condensing relevant information from multiple text sources into concise and coherent summaries. This research paper explores various techniques and methodologies for generating high-quality summaries from diverse document collections. The study investigates the use of clustering algorithms to group similar documents, topic modeling approaches to identify key themes, and sentence fusion methods to create summaries that maintain coherence and readability. The proposed techniques are evaluated using established metrics such as ROUGE scores and human assessments. The findings contribute to the advancement of multi-document summarization by providing insights into effective strategies for generating coherent summaries and their potential applications in information retrieval, document analysis, news aggregation, and other text processing domains.

KEYWORDS: Multi-document summarization, coherence, clustering algorithms, topic modeling, sentence fusion, evaluation metrics, information retrieval, document analysis, news aggregation.

I. INTRODUCTION

The exponential surge in digital information has posed a significant challenge for individuals in navigating through vast quantities of textual data to extract the most relevant and important information. Multi-document summarization, which falls under the umbrella of natural language processing (NLP), addresses this issue by condensing multiple text sources into concise and coherent summaries. The primary objective of multi-document summarization is to distill the essence of the source documents, capturing key ideas and salient information while preserving overall coherence and informativeness.

The importance of multi-document summarization is evident across various domains, including information retrieval, document analysis, news aggregation, and decision-making processes. In information retrieval, summarization techniques enable users to quickly grasp the main points of multiple documents, saving valuable time and effort. Within document analysis, summarization facilitates the extraction of crucial content for further examination or knowledge extraction. News aggregation platforms heavily rely on summarization algorithms to deliver users concise summaries of news articles from diverse sources. Additionally, summarization plays a vital role in decision-making processes by presenting key information to aid informed choices.

This research paper aims to explore and investigate diverse techniques and methodologies for generating high-quality summaries from multiple text sources. The challenge lies not only in identifying the most relevant sentences but also in ensuring coherence, capturing important themes, and avoiding redundancy in the summaries. Recent years have witnessed the emergence of various approaches for multi-document summarization, including clustering algorithms, topic modeling techniques, and sentence fusion methods. These approaches aim to group related documents, identify dominant themes, and seamlessly merge sentences from multiple sources to maintain coherence.

The objective of this research paper is to contribute to the field of multi-document summarization by proposing novel techniques and methodologies that produce coherent and informative summaries. The effectiveness of these proposed techniques will be evaluated using established metrics, such as ROUGE scores and human assessments, to measure the quality and informativeness of the generated summaries. The outcomes of this research hold practical implications for enhancing information access, facilitating decision-making processes, and improving the efficiency of knowledge extraction from extensive textual data sources.

In the subsequent sections of this paper, we will delve into the specific techniques explored, elaborate on the experimental setup, present the results and analysis, and discuss the implications of the research findings across various domains where multi-document summarization is crucial. Through this research, our aim is to advance the field of multi-document summarization and contribute to the development of more efficient and effective techniques for condensing large volumes of textual information into coherent and informative summaries.

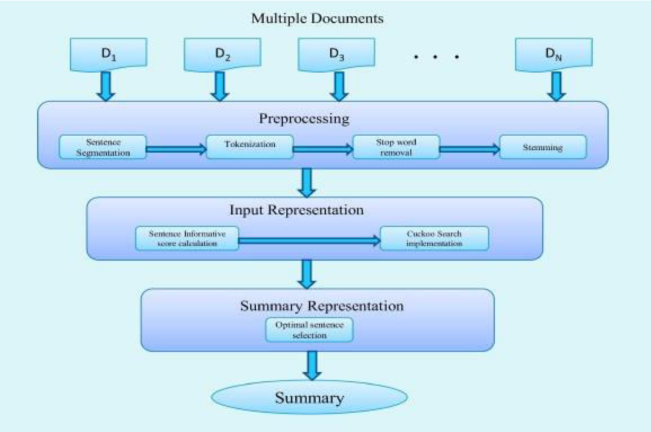


Fig 1. Multiple Documents Summarization

II. LITERATURE REVIEW

The field of multi-document summarization has received considerable attention from researchers aiming to tackle the challenges of condensing multiple text sources into concise and coherent summaries. This section provides an overview of the existing literature and research conducted in multi-document summarization, outlining the diverse approaches, techniques, and methodologies employed by scholars.

One prevalent approach in multi-document summarization is the extraction-based method, which involves selecting relevant sentences or passages from the source documents to form the summary. Early studies focused on using statistical and linguistic features to rank sentences based on their importance. Subsequent research explored advanced techniques such as graph-based algorithms and optimization models to enhance the sentence selection process.

Another approach is the abstraction-based method, which aims to generate summaries by paraphrasing and rephrasing the content of the source documents. This technique involves comprehending the meaning and context of the text and generating new sentences that convey essential information. Abstraction-based methods often employ natural language generation techniques like syntactic parsing, semantic analysis, and language modeling.

Recently, there has been a growing interest in hybrid approaches that combine extraction and abstraction techniques. These methods leverage the strengths of both approaches by extracting important sentences from source documents and modifying or reorganizing them to create coherent and readable summaries. Sentence fusion and sentence compression are among the various methods proposed to achieve this objective.

Topic modeling has also been explored in the context of multi-document summarization. Techniques like Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) have been used to uncover latent topics in document collections. These topics guide the summarization process, ensuring that the generated summaries capture the primary themes and key information from the source documents.

The evaluation of multi-document summarization systems is vital for assessing their effectiveness. Several evaluation metrics, including ROUGE scores, measure the overlap between system-generated summaries and human-created or reference summaries. Human assessments, where judges rate the quality of the summaries, are also commonly employed to evaluate readability, coherence, and informativeness.

Despite significant progress in multi-document summarization, challenges persist, including redundancy handling, coherence across multiple documents, and adaptation to different genres and domains. Future research directions may involve leveraging deep learning models, exploring novel evaluation metrics that better capture coherence and readability, and investigating cross-lingual and multi-modal approaches to handle diverse data sources.

In conclusion, the existing body of work in multi-document summarization encompasses extraction-based, abstraction-based, hybrid, and topic modeling approaches. Evaluation metrics such as ROUGE scores and human assessments are commonly used to evaluate the quality of generated summaries. While notable advancements have been made, there are still ample opportunities for further research and innovation in addressing the challenges associated with multi-document summarization.

III. METHODOLOGY

Multi-document summarization refers to the automatic condensation of a set of documents into a concise and coherent summary. Various techniques and methodologies are employed to address this task. Below is an overview of commonly used approaches.

Extractive Summarization:

In extractive summarization, methods are used to select and assemble sentences or passages from the input documents to form a summary. The process typically involves the following steps:

a. Sentence Scoring: Each sentence is assigned a score based on features like word frequency, position, importance of the containing document, etc.

b. Sentence Selection: Sentences with the highest scores are chosen to construct the summary. Different strategies, such as greedy algorithms or integer linear programming formulations, can be employed for sentence selection.

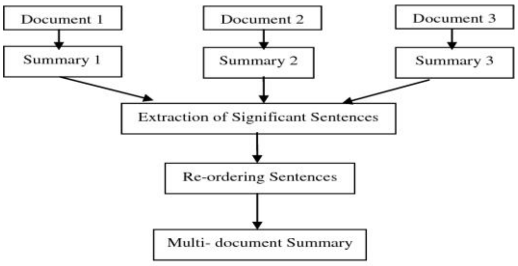


Fig 2. Extractive Summarization

Abstractive Summarization:

Abstractive summarization methods aim to generate a summary that may include new sentences not present in the source documents. These approaches utilize advanced techniques such as natural language generation and deep learning. Some popular methodologies include:

a. Sequence-to-Sequence Models: Models like Recurrent Neural Networks (RNNs) or Transformer models are trained on large datasets to learn the mapping between input documents and their corresponding summaries. They generate summaries by decoding from an encoded representation of the input.

b. Reinforcement Learning: This approach combines sequence-to-sequence models with reinforcement learning, formulating summarization as a reinforcement learning problem. The model receives rewards based on the quality of the generated summaries.

c. Transformer-Based Language Models: Models such as GPT (Generative Pre-trained Transformer) have been applied to abstractive summarization tasks. They are fine-tuned on specific summarization datasets, leveraging the transformer architecture to capture contextual information and generate coherent summaries.

Clustering algorithms play a crucial role in grouping related documents for multi-document summarization. They help identify the underlying structure and relationships within the document collection. Some commonly used clustering algorithms are:

K-means Clustering: This algorithm partitions documents into k clusters based on the similarity of their features. It iteratively assigns documents to the cluster with the closest centroid and updates the centroids accordingly.

Hierarchical Clustering: This algorithm builds a hierarchy of clusters by iteratively merging or splitting existing clusters based on their similarity. It can be either agglomerative (bottom-up) or divisive (top-down).

Density-Based Clustering:

These algorithms identify clusters based on the density of the data points. They group together documents that are close to each other in the feature space while considering the density of neighboring points.

Topic modeling approaches can be employed to identify key themes within the document collection, which can aid in the summarization process. Two widely used topic modeling techniques are:

Latent Dirichlet Allocation (LDA): LDA is a probabilistic generative model that represents documents as mixtures of topics, where each topic is characterized by a distribution over words. It automatically identifies the underlying topics and their prevalence in the document collection.

Non-negative Matrix Factorization (NMF): NMF factorizes the term-document matrix into two matrices, one representing topics and the other representing document-topic weights. It extracts a set of non-negative basis vectors (topics) and assigns a weight to each topic for each document.

These topic modeling approaches provide a higher-level representation of the document collection and enable the identification of key themes, which can be utilized to guide the summarization process.

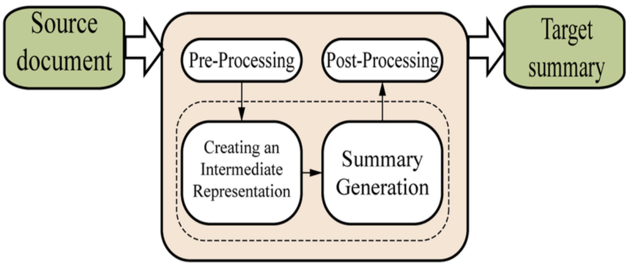


Fig 3. Abstractive Summarization

IV. EXPERIMENTAL SETUP

Dataset:

The selection of an appropriate dataset for evaluation relies on the specific domain and task of multi-document summarization. Typical datasets encompass news articles, scientific papers, legal documents, or collections of online articles. It is essential for the dataset to include multiple documents pertaining to the same topic or event, accompanied by reference summaries serving as ground truth.

Evaluation Metrics:

Various evaluation metrics are employed to gauge the quality of generated summaries. Commonly used metrics for multi-document summarization comprise:

a. ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE quantifies the degree of overlap between the generated summary and the reference summaries, measuring n-gram matches, recall, and precision. Variants such as ROUGE-1, ROUGE-2, and ROUGE-L (Longest Common Subsequence) are frequently utilized.

b. Coherence Measures: Coherence measures evaluate the logical and semantic coherence of the generated summaries. They can encompass lexical similarity measures, sentence-level coherence metrics, or even human evaluation based on coherence judgments.

c. Content Coverage: This metric assesses the inclusion of significant information from the source documents in the generated summary. It may involve evaluating the presence of key entities, events, or important facts.

Pre-processing Steps:

Pre-processing steps are applied to the dataset and input documents to enhance the performance of summarization models. Common pre-processing steps involve:

a. Tokenization: Dividing the text into tokens, such as words, phrases, or sentences, for further analysis.

b. Stop word Removal: Eliminating frequently occurring words (e.g., "the," "is," "and") that contribute little to the overall meaning of the text.

c. Lemmatization/Stemming: Reducing words to their base or root form to minimize redundancy and variation in the text.

d. Sentence Segmentation: Segmenting the documents into individual sentences to facilitate sentence-level analysis and selection.

e. Named Entity Recognition (NER): Identifying and labelling named entities (such as person names, locations, organizations) to enhance the understanding of the text.

Parameter Settings:

The parameter settings depend on the specific algorithms and models utilized for multi-document summarization. These parameters may involve the number of clusters in clustering algorithms, the desired summary length, learning rate and batch size for neural network-based models, or the number of topics for topic modeling approaches. Typically, these settings are fine-tuned through experimentation or cross-validation to achieve optimal performance.

It is important to note that the experimental setup can vary depending on the particular research or implementation context. Researchers and practitioners often tailor their methodology and parameter settings based on the dataset's characteristics, evaluation objectives, and available resources.

V. APPLICATIONS

1. Multi-document summarization plays a vital role in different domains, including information retrieval, document analysis, news aggregation, decision-making processes, data mining and knowledge extraction, legal and patent analysis, and social media analysis.

2. In information retrieval, it helps users quickly grasp the main points of multiple documents by condensing them into concise summaries, saving time and effort in information extraction.

3. In document analysis, it aids in extracting important content from a collection of documents, facilitating efficient knowledge extraction for researchers and analysts.

4. News aggregation platforms utilize multi-document summarization techniques to provide users with coherent summaries of news articles, enabling them to stay informed without reading the full articles.

5. In decision-making processes, summarization techniques present key information in concise and coherent summaries, helping decision-makers understand main points, consider perspectives, and make informed choices.

VI. LIMITATIONS AND DISADVANTAGES

1.Loss of Information: Summarization involves condensing lengthy documents into shorter summaries, leading to some inevitable information loss. Selecting the most relevant information while omitting less significant details can result in the exclusion of nuances, context, or specific examples that are crucial for a comprehensive understanding of the original documents.

2. Ambiguity and Subjectivity: Summarization becomes challenging when dealing with complex or ambiguous content. The process of summarization involves subjective judgments and interpretations, which can vary across different

systems or human summarizers. This subjectivity can introduce biases and influence the perception and understanding of the presented information.

3. Domain and Genre Dependency: Multi-document summarization techniques often rely on domain-specific knowledge and linguistic patterns, making their performance vary depending on the domain or genre of the source documents. Adapting and fine-tuning these techniques for different domains or genres can be time-consuming, limiting their applicability across diverse datasets.

VII. THE FUTURE OF MULTI-DOCUMENTS TEXT SUMMARIZATION

The future of multi-document text summarization presents exciting possibilities for progress and transformative applications. Here are several crucial domains being investigated by researchers and practitioners:

1. Advanced Neural Models: The integration of advanced neural models, like GPT (Generative Pre-trained Transformer), has demonstrated promising outcomes in various natural language processing tasks. These models excel at capturing intricate relationships, contextual information, and semantic understanding, resulting in more accurate and coherent summaries.

2. Contextual Understanding: Future research aims to enhance the contextual understanding of multi-document summarization systems. This involves developing models that can effectively interpret and leverage the context of source documents, including temporal, spatial, and domain-specific information. By considering the broader context, summaries can offer a more comprehensive and nuanced representation of the original texts.

3. Detailed Summarization: While traditional techniques provide high-level summaries, there is increasing interest in generating more detailed and specific summaries. Fine-grained summarization allows users to extract precise facts, arguments, or evidence from multiple documents, facilitating focused and targeted information retrieval.

4. User-Centric Summarization: The future of multi-document summarization focuses on user-centric approaches. This entails incorporating user preferences, needs, and feedback into the summarization process. Customizable summarization systems that enable users to specify desired level of detail, domain-specific requirements, or preferred summary structures can enhance the usefulness and relevance of generated summaries.

5. Multimodal Summarization: As multimedia data becomes more abundant, future research will concentrate on multimodal summarization, which combines textual information with other modalities like images, videos, or audio. Integrating and summarizing information from multiple modalities can provide a more comprehensive and enriched understanding of complex topics.

VIII. CONCLUSION

In conclusion, multi-document text summarization is a dynamic and rapidly progressing field with vast potential across various applications. Extensive research and development have led to the exploration of a wide range of techniques, methodologies, and models for generating concise and informative summaries from multiple text sources. The advancements in natural language processing, deep learning, and linguistic analysis have significantly enhanced the accuracy and effectiveness of summarization.

However, there are still notable challenges and limitations that need to be addressed. Balancing the trade-off between information loss and summary length, addressing the subjectivity inherent in summarization, and ensuring coherence and readability remain important areas for future investigation. Adapting summarization approaches to different domains and genres, tackling scalability concerns, and establishing robust evaluation metrics are also key for the continued progress of the field.

Looking ahead, the future of multi-document text summarization is promising. Advanced neural models, such as transformer-based architectures, along with improved contextual understanding, fine-grained summarization, and user-centric approaches, are expected to elevate the quality and relevance of generated summaries. The integration of multimodal information, explain ability, domain adaptation, and real-time summarization will expand the capabilities and practicality of summarization systems. Ethical considerations and bias-aware summarization will be pivotal in ensuring fairness and inclusivity throughout the summarization process.

REFERENCES

1.Multi-Topic Multi-Document Summarizer Fatma El-Ghannam1 and Tarek El-Shishtawy2 1 Electronics Research Institute, Cairo, Egypt 2 Faculty of Computers and Information, Benha University, Benha, Egypt

2. Information Fusion in the Context of Multi-Document Summarization Regina BarzUay and Kathleen R. McKeown Dept. of Computer Science Columbia University New York, NY 10027, USA

3. Rhetorics-based multi-document summarization q John Atkinson ⇑, Ricardo Munoz Department of Computer Sciences, Universidad de Concepcion, Chile

4. Multi-Document Summarization of Evaluative Text Giuseppe Carenini, Raymond Ng, and Adam Pauls Department of Computer Science University of British Columbia Vancouver, Canada

5. Generating Single and Multi-Document Summaries with GISTEXTER Sanda M. Harabagiu

6. Empirical Analysis of Single and Multi-Document Summarization using Clustering Algorithms Mrunal S. Bewoor Department of Computer Engineering Bharati Vidyapeeth (Deemed to be) University College of Engineering Pune, India

7. From Single to Multi-document Summarization: A Prototype System and its Evaluation Chin-Yew Lin and Eduard Hovy University of Southern California / Information Sciences Institute 4676 Admiralty Way Marina del Rey, CA 90292

8. Comparing Redundancy Removal Techniques for Multi–Document Summarisation Eamonn Newman, William Doran, Nicola Stokes, Joe Carthy, John Dunnion Intelligent Information Retrieval Group, Department of Computer Science, University College Dublin, Ireland

9. A Proposed Textual Graph Based Model for Arabic Multi-document Summarization Muneer A. Alwan1, Hoda M. Onsi2 Information Technology Department Faculty of Computers and information, Cairo University Cairo, Egypt

10. Using Topic Themes for Multi-Document Summarization SANDA HARABAGIU and FINLEY LACATUSU University of Texas at Dallas

11. A Hybrid Approach to Multi-document Summarization of Opinions in Reviews Giuseppe Di Fabbrizio Amazon.com∗ Cambridge, MA - USA

INTERNATIONAL JOURNAL
OF MULTIDISCIPLINARY RESEARCH
IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

+91 99405 72462 +91 63819 07438 ijmrsetm@gmail.com

www.ijmrsetm.com