Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.755224

ADVANCING MULTI-DOCUMENT TEXT SUMMARIZATION THROUGH DEEP LEARNING AND PERSONALIZATION TECHNIQUES

Veena R
Research Scholar, Sri Siddhartha Academy of higher Education, Tumkur-572105
ORCID ID: 0009-0001-6879-1211
Corresponding mail id:Veena.channig@gmail.com

Dr. Ramesh
Professor &Head, Department of Master of Computer Applications, Sri Siddhartha Institute of Technology, Tumkur-572105

Dr. Hanumanthappa M
Professor &Chairman, Department of Computer Science and Applications, Bangalore University, Bangalore 560056.
ORCID ID: 0000-0002-5240-5021

INRTODUCTION

In today's information age, the vast amount of textual data available on the internet presents both opportunities and challenges. While access to information has never been easier, the sheer volume of data can overwhelm users seeking to extract relevant insights efficiently. Multi-document text summarization emerges as a critical solution to this challenge, offering a means to distill large collections of documents into concise and informative summaries

The primary objective of multi-document text summarization is to produce condensed representations that capture the essential information across multiple documents while retaining the key insights and context present in the original texts. By condensing the content into a more manageable form, summaries enable users to grasp the core ideas and themes without the need to sift through extensive volumes of text. Traditional approaches to multi-document summarization have often relied on rule-based or statistical methods, which may struggle to capture the nuances and complexities of natural language. However, recent advancements in deep learning techniques have opened new avenues for improving the quality and effectiveness of text summarization. By leveraging neural networks and large-scale language models, researchers have achieved remarkable progress in generating summaries that exhibit greater coherence, relevance, and fluency

In this paper, we introduce an innovative method for multi-document text summarization that extends the capabilities of deep learning techniques while integrating personalized elements to customize the summaries according to individual user preferences and requirements. Inspired by recent progress in personalized recommendation systems and natural language processing, our approach endeavors to provide summaries that not only encapsulate the essential information from multiple documents but also align with the user's interests and goals. By integrating deep learning models, personalized profiling, and content analysis, our methodology aims to advance the frontier of multi-document text summarization, providing users with a customized and user-friendly approach to traverse the extensive realm of online textual information. Through the

utilization of machine learning and natural language understanding, our goal is to equip users with the capabilities necessary to comprehend the continuously expanding digital knowledge ecosystem. In the following sections of this paper, we will explore the specifics of our proposed methodology, covering aspects such as the architecture of our deep learning models, the methods utilized for personalized profiling, and the evaluation framework employed to gauge the effectiveness of our approach. Furthermore, we will provide insights into experimental outcomes and examine the implications of our findings, shedding light on potential applications and avenues for future research in personalized multi-document text summarization

2.RELATED WORKS

[1] In recent years, multi-document text summarization has garnered attention due to the proliferation of textual data. Veningston et al. (2023) introduce a personalized approach using deep learning, building on prior NLP and machine learning research. Extractive and abstractive methods have been explored, with deep learning, particularly RNNs and transformer models, showing promise. Challenges include information fusion and coherence, addressed through user feedback and domain-specific knowledge integration. The field continues to evolve with ongoing research aimed at developing innovative techniques for summarizing diverse textual data sources.

[2] Recent research in multi-document summarization has focused on developing innovative approaches to generate concise summaries from multiple documents. Puduppully et al. (2022) propose centroid-based pretraining, where a model learns document embeddings representing the centroid of a document set to guide summary generation. This method aims to capture essential information shared across documents. Challenges in multi-document summarization include effectively capturing the essence of multiple documents and generating coherent summaries. Prior research has explored various techniques, including extractive and abstractive methods, as well as the integration of external knowledge sources. Despite advancements, there are ongoing efforts to improve the summarization of diverse textual sources.

[3] Recent studies in multi-document summarization have delved into inventive strategies to augment summarization models. Ketineni and Sheela (2023) introduce a hybrid optimization model, which integrates metaheuristic algorithms with LSTM networks. Despite advancements, challenges persist in optimizing models to generate informative and coherent summaries. Their approach utilizes metaheuristic algorithms such as genetic algorithms to optimize LSTM parameters. This hybrid model endeavors to enhance LSTM-based summarization by improving information capture. Current research efforts in the field are concentrated on devising models for concise and informative summaries..

[4] Extractive multi-document automatic text summarization continues to pose challenges in natural language processing. Wahab et al. (2024) introduce a decomposition-based multi-objective differential evolution approach to tackle this challenge. Their methodology is designed to concurrently optimize multiple objectives, striving to achieve a balance between summary informativeness and diversity. While previous studies have investigated different techniques such as graph-based methods and clustering algorithms, persistent challenges remain. The approach proposed by Wahab et al. utilizes differential evolution to optimize multiple objectives,

thereby enhancing the quality of summaries. Current endeavors are directed towards the development of innovative approaches for summarizing extensive volumes of textual data.

[5] Extractive multi-document summarization remains a formidable challenge within the realm of natural language processing .Ghadimi and Beigy (2023) present SGCSumm, a novel approach that incorporates pre-trained language models, sub modularity, graph convolutional neural networks. Their methodology is geared towards producing informative summaries by harnessing a variety of techniques. Previous research has explored various methods, but challenges persist in capturing document semantics and structure. SGCSumm addresses these challenges by combining complementary approaches. This integration allows SGCSumm to produce high-quality summaries from multiple documents. Ongoing efforts in the field focus on developing innovative techniques to improve summary quality.

[6] Query-focused multi-document summarization (QMDS) is a significant research area in natural language processing. Roy and Kundu (2023) conduct a comprehensive review of QMDS techniques, providing insights through comparative analysis. Previous research has explored diverse approaches, including retrieval-based and generation-based methods. Challenges persist in summarizing multiple documents while considering specific user queries. The review categorizes existing techniques and evaluates their performance using standard metrics. This analysis guides the development of more effective QMDS approaches. Ongoing efforts aim to address the complexities of QMDS and improve summary quality.

[7] Automatic multi-document text summarization poses a significant challenge within the field natural language processing. Abo-Bakr and Mohamed (2023) propose a large-scale sparse multi-objective optimization algorithm to address this challenge. Their approach aims to generate high-quality summaries by optimizing multiple objectives simultaneously. Previous research has explored various techniques, including graph-based methods and neural network architectures, but challenges persist in efficiently summarizing large volumes of textual data while ensuring quality. The algorithm introduced by Abo-Bakr and Mohamed (2023) leverages sparse representation techniques and multi-objective optimization to efficiently optimize summary informativeness and diversity. This approach enables the algorithm to generate high-quality summaries from multiple documents. In summary, recent research continues to advance with the development of innovative algorithms for automatic multi-document text summarization.

[8] Multi-document summarization for learning materials is crucial in educational technology research. Sakkaravarthy Iyyappan and Balasundaram (2023) propose a novel approach combining concept-based Inductive Logic Programming (ILP) and clustering methods to address this challenge. Their methodology aims to generate informative summaries by augmenting document elements with concepts and relationships extracted from the text. Previous research has explored various techniques, but challenges persist in effectively summarizing learning materials while maintaining content relevance. The approach integrates ILP techniques and clustering methods to enhance summarization. This enables the generation of informative and concise summaries tailored to learning materials. In summary, recent research continues to advance with the development of innovative approaches for multi-document summarization in educational contexts

3. Methodology Description

3.1 Research Focus

Advancing machine comprehension presents a critical challenge in artificial intelligence research. Despite advancements, machines still encounter difficulties in reading and comprehending text with human-like understanding. Hermann et al. (2015) introduced innovative methods in "Teaching Machines to Read and Comprehend," yet obstacles persist in grasping context and reasoning. This study endeavors to tackle these challenges by developing original techniques to augment machine comprehension capabilities. Through the utilization of neural network architectures and training methodologies, our aim is to expand the frontiers of machine comprehension. Our goal is to narrow the disparity between human and machine comprehension of textual information, laying the groundwork for more sophisticated natural language understanding systems.

3.2. Text Processing

Prior to model training, it is essential to preprocess the raw data to ensure optimal learning outcomes. Disorganized data can hinder learning despite its volume, necessitating preprocessing techniques to enhance model performance. Several preprocessing steps are employed to prepare the data before feeding it into the model.

Whitespace removal : Unnecessary whitespace is removed from the text to ensure uniformity and facilitate subsequent processing steps.

Stop word removal: Common stop words, which carry little semantic meaning, are eliminated from the text to reduce noise and focus on relevant content.

Special symbol removal: Unreadable special symbols are filtered out from the text to improve readability and avoid potential encoding issues.

Contraction Mapping: Contraction mapping is employed to address contractions within the text, enhancing consistency and facilitating better comprehension in the summary text.

Lowercasing: Documentary information is converted to lowercase to standardize the text and avoid redundancy in model training. However, some standard acronyms are preserved in their original case to maintain their semantic significance.

3.3 Designing a Customized Recursive Neural Network Structure for Summarization.

A new method for multi-document text summarization is introduced, utilizing a Recursive Neural Network (ReNN) architecture enhanced with Convolutional Neural Networks (CNNs) and Siamese Networks (SNs). This architecture is designed to tackle the challenge of summarizing extensive textual data while integrating user preferences to generate personalized summaries. The ReNN initially processes each sentence individually, extracting Word-to-Vec embeddings using Global Vectors for Word Representation (GloVe) embedding, and then passing them through a CNN to acquire feature representations. These representations serve as the foundation of a hierarchical structure, where each layer captures progressively abstract features of the input text. At each level, a ReNN is constructed to recursively process the input text and capture both local and global dependencies. The output representations from each layer of the hierarchical structure are merged to produce the final summary, which is tailored to the user's preferences using SNs to evaluate the similarity between user preferences and summary candidates.

Incorporating user preferences into the summarization process is crucial for producing contextually relevant summaries. To accomplish this, a Siamese Network (SN) is utilized to evaluate the similarity between user preferences and potential summaries. The SN comprises two

identical sub-neural networks with shared weights, enabling the comparison of feature vectors between user representation and summary candidates. The proximity or similarity between user preferences and the output summary is estimated using an objective function, facilitating the generation of personalized and contextually relevant summaries. In summary, the proposed ReNN architecture, in conjunction with CNNs and SNs, allows for the integration of user preferences into the summarization process, resulting in summaries that effectively capture the essential information from multiple documents while closely aligning with the user's interests and objectives.

$$\mathbf{L = (U_{preference}, C_{summary}, y)} \quad \mathbf{(1)}$$

The proximity or similarity between user preferences *Up* and the output summary *Cs* is approximated using Equation (1), where *y* represents a Binary signal denoting if the two text segments are from the same category

In this approach, the temporal aspect of user preference is integrated, a feature absent in many related works. The sequence in which a user reads articles is deemed crucial; thus, the order of article consumption influences subsequent article choices. The user's article reading history is conceptualized as a sequential data set, capturing the order in which articles are consumed. The objective is to summarize a collection of news articles according to the preferences of user U, leveraging the user's historical data. A ReNN or LSTM-based encoder is employed to encode information from the user's historical data, which is then combined with the representation of candidate news articles. Subsequently, the representation with the highest similarity to the user's reading history is selected to generate the summary. The user's reading history serves as the basis for constructing the user preference model, encompassing articles read up to time period tUtilizing both traditional LSTMs and Attentional LSTMs, the outputs of the user model and candidate articles are combined and processed through additional layers or a Multilayered Perceptron. The resulting summary of multiple news articles is refined through backpropagation, facilitating weight updates across all layers, from the Multilayered Perceptron to the RNN layers, if the summary fails to meet expectations [11, 12].

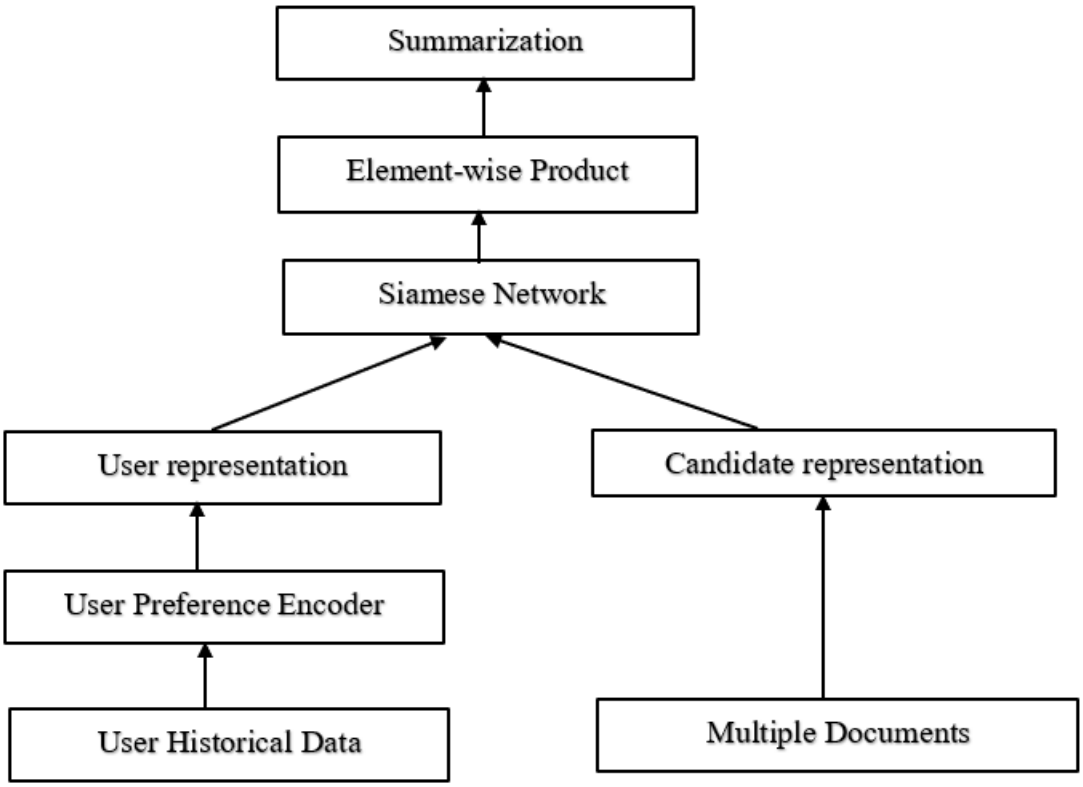

Fig.1, User Preference Integration in Text Summarization.

4. Experimental Assessment

4.1 DATA DISCRIPTION

Evaluation of the Proposed Work Using Two Distinct Datasets

1. **Daily mail dataset**

SIZE: This dataset contains approximately 200,000 text documents

Description: The documents in this repository are sourced from news articles published by the Daily Mail, covering a wide range of topics and domains

2. **Multinews dataset**

Source : The MultiNews dataset is composed of news articles accompanied by human-transcribed precise summaries

Discription : This dataset comprises news articles collected from a diverse array of over 1,500 news websites. Each article is accompanied by a human-generated summary that encapsulates the key points of the article's content

4.2 Baseline model

1 TextRank

In our comparative analysis, the first baseline model we consider is TextRank. Operating on graph-based principles inspired by algorithms like Hyperlink Induced Topic Search (HITS) and PageRank, TextRank autonomously extracts keywords and sentences for summarization by assessing the significance of vertices within the text graph.

2 Standard Convolutional neural network

The second baseline model, Standard Convolutional Neural Network (CNN), presents a deep learning approach to text summarization. It involves passing document embeddings through

convolutional layers with Rectified Linear Unit (ReLU) activation, followed by max-pooling and dropout layers. Subsequently, fully connected layers with sigmoid activation are utilized, and error estimation is performed using the cross-entropy loss function .

3 Hierarchical Network

We examine the Hierarchical Network (HNet) model. HNet combines architectures of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to process each word in a sentence and derive Word2Vec embeddings, resulting in a matrix representation for each sentence. These matrices are treated as images and passed through CNN layers to generate representations for each sentence. These representations form nodes in a hierarchical tree structure imposed on an RNN, enabling comprehensive semantic comprehension in text summarization tasks [15].

4.3. Evaluation metrics used

In evaluating the efficacy of our proposed summarization technique, We utilize the 'Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score', a widely accepted metric in text summarization research. We specifically employ multiple variants of the ROUGE metric [13, 14] to ensure a comprehensive assessment of our deep learning-based text summarization approach.

ROUGE-1 score : This evaluation metric assesses the agreement between the unigrams present in the reference Summary of the test instance and the summary generated by the model.

ROUGE-2 score: The ROUGE-2 score assesses the agreement between the bigrams in the reference Summary of the test instance and the summary generated by the model.

ROUGE-L score : This metric quantifies the length of the longest common subsequence present in both the reference Summary of the test instance and the summary generated by the model. Through the identification of the longest overlapping sequence of n-grams shared between them, the ROUGE-L score evaluates the similarity between the reference summary and the generated output.

4.4. Assessment of LSTM and Recursive neural network Model Performance

Over time, numerous foundational models have been developed for text summarization, each aimed at enhancing summarization quality. In our evaluation, we compare the performance of various summarization models, including traditional techniques and deep learning approaches. The experimental findings are consolidated in Table 1 and Table 2, highlighting the effectiveness of our proposed LSTM and Recursive Neural Network-based technique

Table 1: Performance Metrics on Daily Mail Dataset

Summary Models	ROUGE-1	ROUGE-2
LexRank	32.430	5.880
Standard Convolutional neural network	34.190	7.230
Hierarchical network	37.210	8.960
LSTM and Recursive neural network	41.340	9.510

Table 2: Performance Metrics on Multi News Dataset.

Summary Models	ROUGE-1	ROUGE-2
LexRank	35.290	7.540
Standard Convolutional neural network	35.730	8.690
Hierarchical network	39.170	9.610
LSTM and Recursive neural network	43.940	11.290

Based on the data presented in the tables, it becomes clear that deep learning methodologies, encompassing standard Convolutional Neural Network (CNN), Hierarchical Network, and LSTM-based summarization techniques, demonstrate superior accuracy as indicated by ROUGE scores in comparison to the traditional LexRank-based model. Among the array of deep learning models evaluated, our proposed LSTM and Recursive Neural Network (RNN)-based model consistently exhibits better performance across both datasets.

These results highlight the efficacy of our proposed approach based on LSTM and Recursive Neural Network in producing high-quality summaries, underscoring its potential for practical applications in text summarization tasks.

4.5. Assessment of Personalized Model

Data Collection

To evaluate the effectiveness of personalized multi-document summarization, we collected news articles from a commonly used search engine, specifically Google. Using the headlines of news articles as search queries, we retrieved 15 results per query, typically selecting from the top results. We then extracted only the text content from these news articles to serve as input data for our summarization task. Additionally, in our evaluation process, we incorporated the search histories of 100 users to provide further validation for the proposed model. This data collection methodology was chosen to ensure a reasonable degree of diversity, as there is currently no established benchmark dataset available for evaluating personalized summarization models."

Assessment by Human Judges:

Alongside objective evaluation, we subjected the proposed approach to subjective human assessment. Three skilled annotators proficient in English script writing evaluated fifty randomly chosen test examples from the MultiNews test set. These annotators ranked the summaries based on criteria such as in formativeness, fluency, and succinctness. Their evaluation focused on determining if the final summary effectively conveyed important facts from the input articles, maintained fluency and grammatical correctness, and avoided redundant information.

Reviewers assessed the summaries produced by each technique on a scale ranging from 1 to 5, where 1 represented the highest quality and 5 the lowest. If different techniques demonstrated similar quality, they could receive the same ranking position. The rating for

each technique was calculated as the average score across all test examples used in the experiment

Table 3. Ranking Results of Summaries According to Human Assessment:

Summary models	Reviewer 1	Reviewer 2	Reviewer 3	Average rating
LexRank	27.00	25.00	28.00	26.670
Standard Convolutional neural network	31.00	30.00	30.00	30.340
Hierarchical network	35.00	37.00	38.00	36.670
LSTM and Recursive neural network	39.00	41.00	41.00	40.340

These rankings provide insight into the perceived quality of summaries generated by each technique, as assessed by human annotators. The LSTM and ReNN-based model consistently achieved the highest average score across all annotators, indicating its effectiveness in producing high-quality summaries tailored to individual preferences

5. CONCLUSION

In summary, this paper has examined the importance of personalization in text summarization and utilized deep learning techniques to improve the quality of summaries. Our approach, which involves incorporating pre-trained LSTMs and RNNs into our model, has resulted in notable improvements in multi-document summarization tasks. Integrating user preferences into the encoding process of documents has enabled us to capture more nuanced relationships, resulting in summaries that are both informative and concise. Empirical evidence indicates that our model, which utilizes personalized LSTM and RNN architectures, surpasses several established benchmarks by a significant margin. Moving forward, future research will explore the integration of additional graph representation models, such as knowledge graphs, to further enhance the quality of generated summaries.

REFERENCE

[1] Veningston, K., Rao, P. V., & Ronalda, M. (2023). Personalized Multi-document Text Summarization using Deep Learning Techniques. *Procedia Computer Science*, *218*, 1220-1228.
[2] Puduppully, R., Jain, P., Chen, N. F., & Steedman, M. (2022). Multi-document summarization with centroid-based pretraining. *arXiv preprint arXiv:2208.01006*.
[3] Ketineni, S., & Sheela, J. (2023). Metaheuristic Aided Improved LSTM for Multi-document Summarization: A Hybrid Optimization Model. *Journal of Web Engineering*, *22*(4), 701-730.
[4] Wahab, M. H. H., Hamid, N. A. W. A., Subramaniam, S., Latip, R., & Othman, M. (2024). Decomposition–based multi-objective differential evolution for extractive multi-document automatic text summarization. Applied Soft Computing, 151, 110994.
[5] Ghadimi, A., & Beigy, H. (2023). SGCSumm: An extractive multi-document summarization method based on pre-trained language model, submodularity, and graph convolutional neural networks. *Expert Systems with Applications*, *215*, 119308.

[6] Roy, P., & Kundu, S. (2023). Review on Query-focused Multi-document Summarization (QMDS) with Comparative Analysis. *ACM Computing Surveys*, *56*(1), 1-38.
[7] Abo-Bakr, H., & Mohamed, S. A. (2023). Automatic multi-documents text summarization by a large-scale sparse multi-objective optimization algorithm. *Complex & Intelligent Systems*, 1-16.
[8] Sakkaravarthy Iyyappan, K., & Balasundaram, S. R. (2023). A novel multi document summarization with document-elements augmentation for learning materials using concept based ILP and clustering methods. *International Journal of Computers and Applications*, 1-12.
[9] Christensen, J., Soderland, S., & Etzioni, O. (2013, June). Towards coherent multi-document summarization. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 1163-1173)
[10] Mascarell, L., Chalumattu, R., & Heitmann, J. (2023). Entropy-based Sampling for Abstractive Multi-document Summarization in Low-resource Settings. In *16th International Natural Language Generation Conference (INGL 2023)*..
[11] Liu, S., Zhou, M. X., Pan, S., Qian, W., Cai, W., & Lian, X. (2009, November). Interactive, topic-based visual text summarization and analysis. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 543-552).
[12] Chen, K. Y., Liu, S. H., Chen, B., Wang, H. M., Jan, E. E., Hsu, W. L., & Chen, H. H. (2015). Extractive broadcast news summarization leveraging recurrent neural network language modeling techniques. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(8), 1322-1334.
[13] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).
[14] Lin, C. Y., & Och, F. J. (2004, July). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04) (pp. 605-612).
[15] Singh, A. K., Varma, V., & Gupta, M. (2018). Neural approaches towards text summarization. *International Institute of Information Technology Hyderabad*.