A Survey of Text Summarization Using NLP

Bhuvan Shingade,
Computer Engineering Department,
Terna Engineering College, Nerul,
Navi Mumbai,
(Subject Area – Machine Learning)

Ved Kolambkar,
Computer Engineering Department,
Terna Engineering College, Nerul,
Navi Mumbai,
(Subject Area – Machine Learning)

Yash Matha,
Computer Engineering Department,
Terna Engineering College, Nerul,
Navi Mumbai,
(Subject Area – Machine Learning)

Suyash Kasar,
Computer Engineering Department,
Terna Engineering College, Nerul,
Navi Mumbai,
(Subject Area – Machine Learning)

Prof. Rohini Palve
Computer Engineering Department,
Terna Engineering College, Nerul,
Navi Mumbai,
(Subject Area – Machine Learning)

***Abstract*- The information explosion necessitates innovative approaches for processing and comprehending vast amounts of text data. Text summarization, a subfield of Natural Language Processing (NLP), automatically condenses lengthy documents into concise summaries, preserving the core meaning. This research explores the potential of cloud-based NLP for improved text summarization across various domains. Cloud computing offers the scalability and resources required to handle large datasets and complex NLP models.[1] By leveraging cloud platforms, we aim to develop a robust text summarization system that effectively extracts key information from diverse textual sources. This system will be applicable to a wide range of use cases, including scientific literature reviews, news article summaries, and condensing legal documents.[3] The effectiveness of the generated summaries will be evaluated based on metrics like ROUGE score, human evaluation for coherence and readability, and task-specific measures depending on the use case. This research strives to establish a cloud-based NLP framework for versatile text summarization, aiding knowledge extraction and information retrieval across diverse domains.[2]**

Keywords— Machine Learning, Natural Language Processing(NLP), Long term short memory(LSTM), Abstractive Summarization, Extractive Summarization.

I. INTRODUCTION

The human desire to condense and synthesize information has a long history, dating back to the invention of writing itself. Early forms of text summarization can be seen in practices like ancient scribes creating compendia of knowledge or medieval monks summarizing religious texts. However, the advent of the digital age and the exponential growth of textual data necessitate more sophisticated approaches. The field of automatic text summarization emerged in the mid-20th century alongside the development of computational linguistics and artificial intelligence. Early work focused on statistical methods like keyword extraction and sentence scoring.[12] Pioneering research by Hans Peter Luhn (1958) utilized word frequency to identify key terms in scientific documents, laying the foundation for extractive summarization techniques.[9]

The late 20th and early 21st centuries witnessed a surge in Natural Language Processing (NLP) advancements. Techniques like rule-based systems and statistical machine translation began to be applied to text summarization. These methods incorporated an understanding of syntax and semantics, allowing for more nuanced summarization beyond simple keyword extraction.

The emergence of cloud computing in the early 21st century transformed the landscape of NLP and text summarization. Cloud platforms offer unparalleled scalability and access to vast

computational resources. This empowers researchers to train complex deep learning models on massive datasets, leading to significant improvements in summarization accuracy and quality.

Today, NLP offers a rich toolkit for text summarization. Techniques like sentence embedding, where sentences are represented as vectors in a high-dimensional space, allow models to capture semantic relationships between sentences. Additionally, recurrent neural networks (RNNs) and transformers excel at understanding the context and flow of text, enabling the generation of abstractive summaries that paraphrase the original content while preserving the core meaning.[6]

The integration of NLP with cloud computing promises a future of even more effective and versatile text summarization. As cloud platforms become more powerful and NLP techniques continue to evolve, we can expect summaries that are not only accurate but also tailored to specific user needs and domains. This will have a transformative impact on various fields, from scientific research and journalism to education and business intelligence.

II. NATURAL LANGUAGE PROCESSING

The ever-growing volume of textual data demands efficient methods for extracting key information. Text summarization, a cornerstone of Natural Language Processing (NLP), tackles this challenge by automatically condensing lengthy documents into concise, informative summaries.[7] This emerging field leverages the power of deep learning to achieve unparalleled summarization capabilities.

Understanding Language: NLP empowers computers to process and analyze human language. It employs techniques like:

- Part-of-Speech Tagging: Identifying the grammatical function of words (nouns, verbs, adjectives, etc.).
- Named Entity Recognition: Recognizing and classifying named entities like people, organizations, and locations.
- Dependency Parsing: Understanding the relationships between words in a sentence

Media	Text Images Video Speech Hypertext
Input	Single-document Multi-document
Output	Extract Abstract Headline
Purpose	Generic Personalised Query-focused Update Sentiment-based Indicative Informative Critical
Language	Mono-lingual Multi-lingual Cross-lingual

Table 1: Summarization types according to several factors[2]

A. *ABSTRACTIVE TEXT SUMMARIZATION*

Abstractive summarization, aiming to capture the essence of a text and generate a new, concise summary, relies heavily on effective preprocessing to prepare the text data for deep learning models. This stage transforms raw text into a format suitable for training and generating summaries.[4]

When summarizing a text using abstraction, new sentences are created to represent specific textual information from the original text. By using natural language generation, we achieve this. This produces information that is more human-like in which sentences, as opposed to using specific sentences or words from the original text in the final summary, as in extraction-based summarization techniques, best capture the primary idea of the material.

Summarizations using abstractive techniques are broadly classified into two categories: Structured based approach and Semantic based approach. [5]

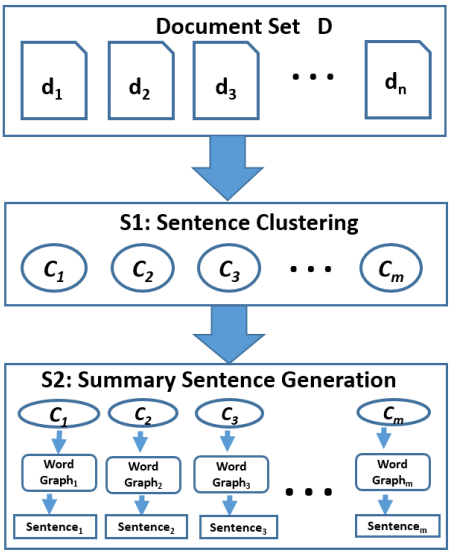

Figure 1: Abstractive Text Summarization

Structured-based approaches focus on the organization and relationships between sentences within the text. Here, techniques like tree-based methods construct hierarchical representations of the document, identifying important sentences based on their position or relationships within the hierarchy.[8] Template-based methods rely on predefined templates or rules to identify key information and generate summaries based on a fixed structure.

Semantic-based approaches delve deeper into the meaning of the text. Multimodal semantic models, for example, incorporate additional information beyond pure text, such as images or videos associated with the document. By analyzing these different modalities together, the model can gain a richer understanding of the content and generate more comprehensive summaries. This approach is particularly beneficial for summarizing complex documents that rely heavily on visual or multimedia elements.[10]

B. EXTRACTIVE TEXT SUMMARIZATION

Identifying the most important sentences or phrases in a text document is the first step in the extraction-based summarization process, which culminates in the creation of a summary. The lines or phrases that are chosen are usually chosen based on how well they fit together, how important they are, and how relevant they are to the main topic.[5]

Selected phrases and keywords for the summary are included in extraction-based summarization. The phrases and terms come from the major body of work. This divides the primary text into sentences or words, which are then chosen and rejected based on a relevance score. The final summary includes the chosen phrases and words.[1]

Extractive summarization pinpoints key sentences by combining TF-IDF, which prioritizes informative words within a document, with clustering, which groups similar sentences thematically.[3] This approach ensures summaries capture the most relevant aspects of the document by selecting representative sentences from each thematic cluster, weighted by their informative keywords.

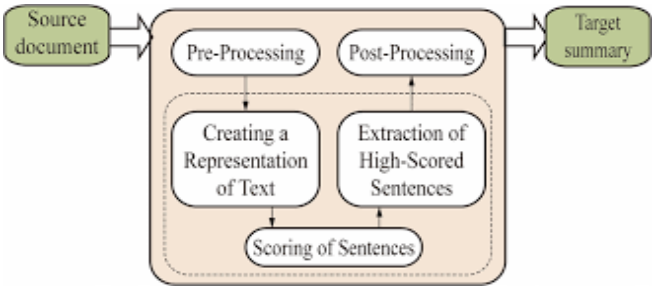

Figure 2: Extractive Text Summarization [3]

Ref	Category	Model Type	Techniques/Algorithm	Strength	Weakness
[1]	Extractive Summarization	Feature-based	Sentence Scoring (TF-IDF, keyword frequency, named entity recognition) Sentence Ranking (TextRank, LexRank) Positional features (sentence location)	Easy to implement , Fast and efficient High factual accuracy	Limited creativity Relies heavily on feature engineering May not capture the overall flow of the text
[3]	Extractive Summarization	Graph-based	Sentence similarity measures (cosine similarity) Sentence centrality in a graph representation	Captures relationships between sentences Identifies thematic clusters	Computationally expensive for large documents. May not be suitable for complex texts.
[4]	Abstractive Summarization	Statistical Machine Translation (SMT)	Phrase-based translation models Word alignment techniques	Can handle diverse text formats. Leverages existing translation technology	Limited ability to paraphrase and capture complex meaning. Prone to errors in factual accuracy
[5]	Abstractive Summarization	Neural Network-based	Recurrent Neural Networks (RNNs) (LSTM, GRU) Attention mechanisms Sequence-to-sequence learning	Captures long-term dependencies in text. Generates grammatically correct summaries. Handles complex sentence structures	Computationally expensive. Requires large amounts of training data. May introduce hallucinations (sentences not present in the original text)
[6]	Abstractive Summarization	Transformers	Encoder-decoder architecture, Self-attention mechanism, Pre-trained models (BERT, BART)	State-of-the-art performance in text summarization tasks. Captures complex relationships between words and sentences. Handles long documents efficiently	Requires significant computational resources, Pre-trained models can be biased, Interpretability can be challenging
[7]	Sparse Summarization	Focuses on Key Information Extraction	Extractive techniques with emphasis on factual content. May utilize topic modelling or entity recognition	Efficient for extracting core factual information. Less emphasis on overall flow or style	Useful for summarizing factual documents like news articles or scientific reports. Can be combined with other summarization techniques for a more comprehensive summary.
[8]	Hybrid Models	Combining Extractive & Abstractive Techniques	Leverage strengths of both approaches Can be tailored to specific summarization goals	Increased complexity compared to individual approaches * Requires careful integration of techniques	Requires careful design of the hybrid architecture and training strategy. * Can be particularly effective for complex or domain-specific text.
[9]	Multi-document Summarization	Summarizing Multiple Related Documents	Techniques for document clustering and topic modelling. Combines summaries of individual documents of generates a single overview.	Useful for summarizing collections of related documents. Can be computationally expensive for large document sets	Requires effective methods for identifying relationships between documents. Can be applied to summarizing news articles on a specific event or research papers on a particular

Table 2: Survey table on different models of Text Summarization

III. CONCLUSION

The rapidly developing topic of text summarization has been examined in this literature review, with a focus on the revolutionary effects of cloud computing and natural language processing (NLP). We've seen how the capacity to produce clear and insightful summaries has been transformed by deep learning models, which are skilled at capturing intricate textual links. Cloud platforms have further democratized access to these potent tools by providing scalability, accessibility, and collaborative environments.

The studied literature demonstrates the noteworthy improvements in efficiency and accuracy of text summarizing. But there are still issues to resolve. Skewed summaries can result from biases in the training data, and more research is needed to determine how interpretable sophisticated deep learning models are. However, more advancements in these fields are anticipated due to ongoing research efforts.

In conclusion, text summarization has a promising future thanks to the combination of NLP and cloud computing. With the use of this technology, people could be better able to traverse the enormous ocean of knowledge, overcome linguistic barriers to communication, and advance in a variety of fields. We anticipate the development of ever more advanced and adaptable summarizing methods as research moves forward, which will finally unlock the full potential of text data.

REFERENCES

[1] Kanithi Purna Chandu "Text Summarization Using Natural Language Processing", International Journal of Research Publication and Reviews, Vol 3, no 11, pp 649-655,.

[2] Elena Lloret, Manual Palomar "Text summarisation in progress: A literature review", Research Gate

[3] Dr. Rashmi Sharma, Shivam Chaudhary, Sejal Tyagi "TEXT SUMMARIZER USING NLP NATURAL LANGUAGE PROCESSING", *International Research Journal of Modernization in Engineering Technology and Science*

[4] Reeta Rani, Sawal Tandon, "LITERATURE REVIEW ON AUTOMATIC TEXT SUMMARIZATION", 2019.

[5] Sheetal Patil, Avinash Pawar, Siddhi Khanna, Anurag Tiwari, Somay Trivedi "Text Summarizer Using NLP", 2022.

[6] Aakash Srivastava, Kamal Chauhan, Himanshu Daharwal, Nikhil Mukati, Pranoti Shrikant Kavimandan "Text Summarizer Using NLP (Natural Language Processing)" , 2020.

[7] Divakar Yadav, Jalpa Desai, Arun Kumar Yadav, "Automatic Text Summarization Methods: A Comprehensive Review" , 2023.

[8] ChetanaVaragantham, J.SrinijaReddy, UdayYelleni, MadhumithaKotha, Dr P.VenkateswaraRao, "TEXT SUMMARIZATION USING NLP", *International Conference on Industry 4.0 Technology (I4Tech)*, 2022.

[9] Adhika Pramita Widyassari, Supriada Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, De Rosal Ignatius Moses Setiadi "Review of automatic text summarization techniques & methods" , *Journal of King Saud University, 2020.*

[10] Ravali Boorugu, Dr. Gajula Ramesh, Dr. Karanam Madhavi, "Summarizing Product Reviews Using NLP Based Text Summarization", International Journal of Scientific & Technology Research Vol 8, Issue 10, 2019

[11] Waseemullah, Zainab Fatima Shehnila Zardari, Muhammad Fahim, Maria Andleeb Siddiqui, Ag. Asri Ag. Ibrahim, Kashif Nisar, Laviza Falak Naz, "A Novel Approach for Semantic Extractive Text Summarization", MDPI, 2022.

[12] Hamza Shabbir Moiyadi, Harsh Desai, Dhairya Pawar, Geet Agrawal, Nilesh M.Patil, "NLP Based Text Summarization Using Semantic Analysis", 2016