JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# Harnessing the Power of NLP for Text Summarization: A Comprehensive Overview

**[1]Richa Sharma, [2]Dr. Meenakshi Nawal, [3]Nidhi Srivastav, [4]Shalini Singhal**
[1]Depatment of CS(AI), [2,3]Department of Computer Science & Engineering , [4] Department of Information Technology

[1] Swami Keshwanand Institute of Technology, Management & Gramothan , Jaipur, Raj

***Abstract:*** This comprehensive overview delves into the transformative capabilities of Natural Language Processing (NLP) in the context of text summarization. As information continues to proliferate across various domains, the need for efficient and accurate summarization tools becomes paramount. This paper explores the key techniques and methodologies employed in harnessing the power of NLP for text summarization, encompassing extractive and abstractive approaches. Additionally, it examines the challenges and opportunities associated with these techniques, providing insights into the current state of the field and potential future advancements. By unraveling the intricacies of NLP-based summarization, this paper aims to guide researchers, practitioners, and enthusiasts toward a deeper understanding of this evolving landscape.

***Index Terms*** **- NLP, Text Summarization, Natural Language Processing, Extractive Summarization, Abstractive Summarization.**

## I. Introduction

In today's world, where information inundates us from diverse sources, the challenge of distilling and comprehending this vast sea of data has become a crucial task. Our surroundings are saturated with an unprecedented volume of textual content, spanning from news articles and research papers to social media updates and corporate reports. The sheer magnitude of available information poses a formidable challenge for both individuals and organizations—how to extract the most relevant and valuable insights from this overwhelming abundance of words [1]. The core of the issue lies in the impracticality of manually sorting through this vast amount of text. Traditional methods of consumption and analysis are no longer adequate in the face of this information overload. There is a palpable need for automated mechanisms that can rapidly and intelligently distill the essence of textual content, providing users with concise and meaningful summaries. This is precisely where text summarization emerges as a pivotal solution [1].

At its core, text summarization aims to condense lengthy passages of text while retaining essential information and context. In addressing the intricacies of this task, Natural Language Processing (NLP) takes the spotlight. NLP, a subset of artificial intelligence, is devoted to empowering machines with the ability to comprehend, interpret, and generate human-like language. By integrating NLP into the text summarization process, we not only enhance the efficiency of information extraction but also introduce a layer of understanding that transcends mere word counting [2].

The objective of this paper is to thoroughly explore the interplay between NLP and text summarization. We aim to unravel the techniques and methodologies that underscore this symbiotic relationship. Delving into both extractive and abstractive approaches to text summarization, we will examine their strengths, weaknesses, and real-world applications. Moreover, the paper will scrutinize the challenges inherent in utilizing NLP for summarization, addressing issues related to accuracy, coherence, and adaptability to diverse content types [2]. The scope of this comprehensive overview goes beyond a simple examination of the current landscape; it encompasses a forward-looking perspective. We will explore the current state of the field, identify opportunities for innovation, and propose future directions for research and development. By providing insights into the intricacies of NLP-based text summarization, this paper seeks to serve as a guiding beacon for researchers, practitioners, and enthusiasts. Our aspiration is not only to elucidate the existing landscape but also to inspire the ongoing evolution of this dynamic field, ensuring its continued alignment with the evolving needs of information consumers in the digital age [2].

## II. Natural Language processing

Exploring the foundational concepts of Natural Language Processing (NLP) that are relevant to summarization necessitates an understanding of the essential components that empower machines to comprehend and manipulate human language. NLP stands as the cornerstone upon which automated summarization is built, providing the tools to extract meaning, context, and subtle nuances from textual data [3]. At its essence, NLP covers a spectrum of linguistic phenomena, spanning from syntactic and semantic analysis to discourse and pragmatic considerations. Syntactic analysis involves dissecting sentences to grasp their grammatical structure, while semantic analysis delves into the meanings of words and their contextual relationships. Discourse analysis focuses on comprehending how sentences interconnect to form cohesive passages, and pragmatic considerations involve capturing the implied meanings and intentions behind language use [3].

Machines endowed with these NLP capabilities possess the agility to navigate the complexities of human language, rendering them indispensable in the realm of text summarization. Their ability to discern key entities, identify relationships between words, and interpret the underlying meaning of sentences contributes to a more nuanced and contextually accurate approach to summarization [4]. The importance of Natural Language Processing (NLP) in the context of text summarization is monumental and cannot be overstressed. NLP essentially serves as the cognitive bridge that connects raw textual data to meaningful and coherent summaries. Its absence would confine automated systems to simplistic word frequency analyses, limiting their capacity to grasp the intricacies inherent in language [4].

NLP empowers machines to grasp the context, sentiment, and nuances vital for crafting effective summaries. It goes beyond mere word extraction, facilitating the identification of key themes, extraction of essential information, and preservation of intended meanings. This elevates the summarization process from a mechanical extraction of words to a more sophisticated and human-like comprehension of content. Furthermore, NLP's adaptability to diverse linguistic styles and domains renders it an incredibly versatile tool for handling various types of textual data. Its dynamic language processing capabilities ensure that summarization algorithms can efficiently distill information from diverse sources such as scientific articles, social media updates, or legal documents [5].

In essence, NLP stands as the linchpin that transforms summarization from a basic task into a nuanced and context-aware process. Its role in enhancing the efficiency, accuracy, and contextual understanding of automated summarization systems is pivotal. NLP becomes a cornerstone in the relentless pursuit of effective information distillation, ensuring that summaries not only capture the essence of the text but also resonate with the subtleties and nuances embedded within language [5].

Navigating the expansive landscape of Natural Language Processing (NLP) for text processing requires a comprehensive understanding of the pivotal algorithms, models, and techniques that underpin machine language comprehension. Here's a detailed exploration:

### 2.1 Tokenization:

Tokenization stands as a fundamental process within the realm of Natural Language Processing (NLP), designed to deconstruct a given text into smaller units, commonly known as tokens. The primary objective is to establish a structured representation of textual content, laying the groundwork for subsequent analysis and processing. Tokens, serving as elemental building blocks, empower machines to engage with and manipulate language at a finely detailed level [6].

*Key Steps in Tokenization:*

1. **Word Tokenization:**
    - In this phase, the text undergoes segmentation into individual words. Punctuation marks and special characters are often treated as distinct tokens. For instance, the sentence "Tokenization is crucial!" would be tokenized into ["Tokenization", "is", "crucial", "!"].
2. **Sentence Tokenization:**
    - Certain applications necessitate the identification of individual sentences within a larger text. Sentence tokenization accomplishes this by segmenting the text into coherent sentences. For example, the paragraph "NLP is fascinating [6]. It enables machines to understand human language." would be tokenized into ["NLP is fascinating.", "It enables machines to understand human language."].

*Importance:*

- Text Processing: Tokenization serves as a foundational step in the processing of textual data, offering a structured representation that facilitates subsequent analysis.
- Statistical Analysis: In tasks such as sentiment analysis or word frequency analysis, tokenization provides the means to calculate statistics at the word level.
- Feature Extraction: In the domain of machine learning, tokenization plays a pivotal role in extracting features from text, enabling models to comprehend and learn underlying patterns.
- Search Engines: Essential in search engines, tokenization aids in matching user queries with relevant tokens within documents, enhancing search accuracy.
- Language Understanding: By breaking down text into tokens, machines acquire a nuanced understanding of the language's intricate structure, facilitating interpretation and effective content analysis [6].

*Challenges:*

- Ambiguity: Some words possess multiple meanings, introducing challenges for tokenization to accurately capture the intended sense without additional context.
- Languages with No Spaces: Languages like Chinese or Thai, lacking clear word separations, pose difficulties in determining appropriate token boundaries.
- Handling Contractions and Hyphenated Words: Tokenizing contractions (e.g., "can't") and hyphenated words demands careful consideration to preserve meaningful units during the process.
- Tokenization emerges as a pivotal step in the NLP journey, providing the foundational framework for an array of language processing tasks. By enabling machines to comprehend and analyze textual content at a granular level, tokenization unlocks the potential for sophisticated language understanding and processing [6].

### 2.2 Part-of-speech tagging (POS)

Part-of-speech tagging (POS) stands as a pivotal component in the realm of natural language processing (NLP), encompassing the assignment of grammatical categories to individual words within a sentence. These categories span nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, and interjections. At its core, POS tagging seeks to furnish a comprehensive and organized comprehension of the syntactic structure inherent in a given text. In the intricate choreography of language, each word assumes a distinct role, and POS tagging serves as the guiding force for machines to discern and allocate these roles. For example, the identification of nouns illuminates the entities or objects within a sentence, verbs delineate actions or processes, adjectives articulate attributes, and adverbs furnish additional details about actions. This categorization of words into grammatical roles serves as the foundation for syntactic analysis, empowering machines to apprehend the relationships and structure embedded within a sentence [7].

The POS tagging process involves harnessing pre-existing lexical databases, statistical models, or machine learning algorithms trained on annotated corpora. Every word is then tagged with its corresponding part of speech, guided by its contextual placement within the sentence. This contextual understanding assumes paramount importance, as words may undertake varying roles contingent upon their positioning and the words surrounding them [7]. Beyond its role in syntactic analysis, the significance of POS tagging resonates across diverse downstream NLP applications, spanning information extraction, sentiment analysis, and machine translation. In information extraction, the identification of nouns and verbs proves instrumental in capturing pivotal entities and actions, while in sentiment analysis, adjectives emerge as key players in discerning the emotional tone embedded within the text [7]. Despite its inherent importance, POS tagging grapples with challenges such as word ambiguity and context-dependent meanings. Homonyms, denoting words with identical spelling but differing meanings, introduce complexities in accurately assigning POS tags. Furthermore, context assumes a pivotal role in unraveling the roles of words, as a word can morph into a different part of speech contingent upon its contextual surroundings [7].

In summation, Part-of-Speech Tagging serves as a foundational cornerstone in NLP, furnishing machines with the capacity to dissect and comprehend the grammatical roles assumed by words within a sentence. By aiding syntactic analysis, POS tagging not only enriches our understanding of language structure but also equips machines to navigate and interpret textual content with enhanced efficacy.

### 2.3 Semantic analysis

Semantic analysis, a pivotal aspect of natural language processing (NLP), engages in the intricate task of grasping the embedded meaning within words and sentences, all within a specific context. In contrast to syntactic analysis, which focuses on the structural relationships between words, semantic analysis delves into the profound layers of meaning by considering the nuances and intricacies of language usage [8].At its essence, semantic analysis entails unraveling the meaning of individual words and their collective impact within a sentence. This process surpasses mere word definitions, incorporating contextual variations and connotations that words adopt based on their surroundings. The significance of understanding the context in which words are employed becomes paramount, given that words can assume diverse meanings contingent upon the specific context they inhabit [8].

A significant facet of semantic analysis is sentiment analysis, where the goal is to discern the emotional tone or sentiment conveyed by a piece of text. This involves determining whether the expressed sentiment is positive, negative, or neutral, offering valuable insights into the emotional undertones of a given document or conversation. This application proves particularly crucial in evaluating public opinion, customer feedback, or social media sentiment [9]. Another dimension of semantic analysis involves word sense disambiguation. Words frequently carry multiple meanings, necessitating the determination of the intended sense in a given context for accurate language understanding. For instance, the word "bank" could refer to a financial institution or the side of a river. Semantic analysis employs various techniques, including machine learning models and context-based algorithms, to discern the intended meaning of words within specific instances [9].

Practically, semantic analysis assumes a pivotal role in applications spanning from chatbots and virtual assistants to search engines and recommendation systems. By deciphering the nuanced meaning behind user queries or content, these systems can furnish more contextually relevant and accurate responses. Nevertheless, challenges persist in semantic analysis, primarily stemming from the inherent ambiguity and dynamic nature of language. Words can adopt different meanings based on evolving societal norms, cultural shifts, or individual perspectives. Consequently, semantic analysis models must grapple with the complexity of language, endeavoring to adapt to the ever-changing landscape of human expression [9].

In summary, semantic analysis in NLP unfolds as a multifaceted journey into the depths of language meaning. Through unraveling contextual intricacies, deciphering sentiment, and disambiguating word senses, semantic analysis augments the machine's capacity to comprehend and interpret language with a depth that surpasses surface-level understanding.
Syntactic Parsing: Analyzing the grammatical structure of sentences to identify relationships between words.

### 2.4 Topic modeling

Topic modeling is a method within natural language processing (NLP) designed to unveil underlying themes or topics present in a collection of documents. It employs mathematical algorithms to analyze the distribution of words across documents and identifies clusters of words that frequently co-occur. By doing so, it enables the extraction of key themes, helping users understand the dominant subjects or ideas within a diverse set of texts. Common algorithms for topic modeling include Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) [10].

**Machine Translation Models:** Utilizing models like Transformer-based architectures for language translation tasks, which can also be adapted for summarization. Machine translation models leverage advanced architectures such as Transformers to translate text from one language to another. The Transformer model, known for its attention mechanisms, has significantly improved translation accuracy. Interestingly, these models can be adapted for text summarization tasks. By understanding contextual relationships and capturing essential information during translation, they can be repurposed to generate concise and contextually meaningful summaries.

**Attention Mechanisms:** Enhancing models' ability to focus on specific parts of a text, crucial for understanding context and relationships. Attention mechanisms are pivotal in NLP, allowing models to assign varying degrees of importance to different parts of a text. This dynamic focus is especially crucial for understanding context and relationships within the text. Instead of treating each word equally, attention mechanisms enable models to concentrate on specific words or phrases based on their relevance to the overall meaning. This enhances the model's contextual understanding and improves its ability to capture intricate relationships within the text [10].

**BERT (Bidirectional Encoder Representations from Transformers):** A pre-trained model capable of capturing contextual information, widely used in various NLP tasks. BERT represents a significant advancement in NLP, being a pre-trained model that excels in capturing contextual information. Unlike earlier models that processed text in a unidirectional manner, BERT considers the entire context of a word within a sentence. This bidirectional approach results in a more nuanced understanding of language and has proven effective in various NLP tasks, including text summarization. By grasping the intricacies of context, BERT enhances the accuracy and context-awareness of NLP models.

Understanding these algorithms, models, and techniques is crucial as they collectively empower NLP systems to process and understand text. They lay the groundwork for advanced text summarization methods by enabling machines to discern key topics, translate and summarize text effectively, focus on relevant parts of a document, and capture nuanced contextual information. Delving into the nuances of each contributes to a comprehensive understanding of how NLP systems distill valuable information from diverse textual data [10].

## III. Text Summarization

Text summarization serves as a crucial component of natural language processing (NLP), tackling the formidable task of distilling information from extensive textual datasets. This comprehensive exploration delves into diverse approaches for text summarization, unraveling the intricacies inherent in both extractive and abstractive methods. Moreover, the integration of NLP techniques assumes a pivotal role in augmenting the efficiency and contextual comprehension of summarization processes [11].

- Extractive summarization methods: The process of extractive summarization involves the meticulous selection and arrangement of existing sentences or phrases from the source text to craft a concise summary. This approach hinges on the identification of the most informative and relevant sentences, guided by criteria such as sentence importance or keyword presence. Techniques like sentence scoring, graph-based algorithms, and machine learning models are frequently employed to ascertain the significance of sentences. While extractive summarization maintains the original wording, it grapples with challenges in generating coherent and contextually rich summaries [11].

- Abstractive summarization techniques: In contrast, abstractive summarization seeks to generate a summary not explicitly present in the source text. This approach entails a profound understanding of the text's meaning, culminating in the creation of concise and coherent summaries in a more human-like manner. Techniques employed include natural language generation, deep learning models like transformers, and attention mechanisms. While abstractive summarization holds promise for producing contextually nuanced summaries, it encounters challenges in maintaining accuracy and ensuring that the generated content aligns faithfully with the source text [11].

### 3.1 NLP for summarization:

The pivotal role of Natural Language Processing (NLP) in text summarization cannot be overstated. NLP techniques empower machines to comprehend, interpret, and generate human-like language, contributing significantly to both extractive and abstractive summarization methods. The application of NLP encompasses syntactic and semantic analysis, part-of-speech tagging, and sentiment analysis. Leveraging NLP not only enhances the efficiency of information extraction but also introduces a layer of understanding that transcends mere word counting, contributing to more nuanced and contextually aware summarization. Natural Language Processing (NLP) serves as the cornerstone for proficient text summarization, providing an array of techniques and tools that empower machines to comprehend, interpret, and generate language akin to human expression. Within the realm of summarization, NLP assumes a pivotal role in elevating efficiency and depth of understanding, contributing significantly to both extractive and abstractive methodologies. Let's delve into the fundamental components and applications of NLP in the domain of text summarization [12]:

In essence, NLP equips summarization systems with the cognitive prowess required to process and interpret language effectively. Whether disassembling grammatical structures, deciphering sentiment, identifying key entities, or generating human-like summaries, NLP serves as the propelling force behind the intelligent extraction and synthesis of information in the field of text summarization.

### 3.2 Consideration of issues related to accuracy, coherence, and handling diverse content:

Text summarization grapples with challenges encompassing accuracy, coherence, and adaptability to diverse content types. Ensuring accuracy is paramount to guarantee that the summary accurately reflects the key information from the source text. Coherence involves presenting information in a logical and flowing manner, avoiding abrupt transitions. Handling diverse content necessitates addressing variations in writing styles, domain-specific terminology, and ensuring the summarization model is adaptable across different types of textual data. Effectively addressing these issues is crucial for enhancing the overall effectiveness of summarization methods [12].

### 3.3 Opportunities and Innovations:

The landscape of text summarization unfolds numerous opportunities for innovation. Researchers and practitioners can delve into novel algorithms, models, and techniques aimed at improving the accuracy and coherence of summaries. Innovations in machine learning and NLP, including advancements in pre-trained models like BERT and GPT, open new avenues for context-aware summarization. Exploring multi-modal approaches that incorporate visual information along with text represents a frontier for innovation. Additionally, considering the ethical implications of automated summarization and developing methods to handle bias in summarization outputs present opportunities for responsible and inclusive advancements in the field.

In summary, text summarization approaches encompass both extractive and abstractive methods, with NLP playing a crucial role in enhancing understanding. Addressing challenges related to accuracy, coherence, and diverse content types is imperative for refining summarization techniques. Opportunities for innovation lie in exploring advanced algorithms, incorporating multi-modal information, and addressing ethical considerations for responsible advancements in the field [13].

## IV. RELATED PAPERS

Alquliti, W. H., & Ghani, N. B. A. (2019) [14] Recently, natural language processing apps have emerged, employing intelligent and soft computing methods. These methods enable computer systems to emulate human text processing tasks such as plagiarism detection, pattern determination, and machine translation. The study introduces a novel Automatic Text Summarization (ATS) system, CNN-ATS, utilizing a convolutional neural network and a text matrix representation. CNN-ATS, a deep learning system, evaluates the impact of increased depth on CNN configurations, sentence assessment, and extraction of informative sentences for document summarization. The experiment, based on 26 different configurations, shows improved summaries compared to other methods, using DUC 2002 as the data source.

Al Oudah et al. (2019) [15] highlight the challenge of accessing accurate information amidst the rapid growth of Arabic content. They discuss the early stage of summarization systems for Arabic and introduce Wajeez, an automatic Arabic text summarization system with a new scoring formula. Wajeez's performance on different datasets shows promising results, especially when titles support summarization.

Gupta and Kaur (2019) [16] address the bias in Outdegree Centrality (OC) towards introductory sentences in graph-based summarization. They propose using Laplacian Centrality (LC) for more specific and informative summaries, emphasizing its ability to highlight central nodes contributing non-uniformly. The modified method surpasses existing results, enhancing informativeness and coherence in summaries.

Zaman, F., et al. (2020) [17] The study focuses on the relationship between text simplification and text summarization in Natural Language Generation. It introduces a novel hybrid architecture called HTSS, combining abstractive and extractive summarization. Results indicate that the proposed HTSS model surpasses NTS and ATS in the joint task of simplification and summarization by 38.94% and 53.40%, respectively.
A new metric, CSS1, combining SARI and ROUGE, is introduced to demonstrate the model's superior performance in the combined task.

Fang, W., et al. (2020) [ 18] The rapid development of deep learning in Natural Language Processing (NLP) has yielded impressive results. Automatic text summarization, achieved by computer programs without altering the original document's intent, is crucial in various applications. The study presents a model based on the convolutional neural network, specifically CNN-ATS, for Chinese article summarization. The model demonstrates good summarization performance but requires further research for improving sentence organization. Text Simplification and Summarization are distinct tasks in Natural Language Generation. The Hybrid Task-Specific Summarization (HTSS) model combines both abstractive and extractive summarization, outperforming existing models in simplification and summarization tasks. The model, utilizing a parallel corpus from EurekaAlert, achieves superior SARI and ROUGE scores.

Mollá et al. (2020) [19] discuss Macquarie University and the Australian National University's participation in the 2020 BioASQ Challenge. They implement a framework for query-focused multi-document extractive summarization, experimenting with BERT, BioBERT, Siamese architectures, and reinforcement learning. The best results are observed when using BERT for word embeddings followed by an LSTM layer for sentence embeddings. Siamese architectures and BioBERT variants did not enhance results.

Xu et al. (2020) [20] address the difficulty of evaluating abstractive summarization and introduce a novel metric based on fact-level content weighting. Their metric relates the facts of the document to those in the summary, aiming to reflect all relevant facts from the ground truth. The weightings correlate well with human perception, outperforming a recent manual highlight-based metric by Hardy et al. (2019).

Lal, N. M., et al. (2021) [21] The growth of textual data on the web poses challenges in retrieving vital information. Automatic Text Summarization (ATS) addresses this by extracting essential content without altering semantics. The study proposes a novel approach in Extractive text summarization, introducing a new sentence scoring parameter that improves summarizer performance compared to other models.

Varab and Schluter (2021, November) [22] address the anglo-centric nature of current automatic summarization research and the challenge of creating high-quality summarization datasets for various languages. They present a groundbreaking multilingual summarization dataset containing articles in 92 languages, making it the largest and most inclusive dataset for any NLP task. The paper investigates resource building from news platforms in low-resource language settings and explores how such settings impact automatic summarization system performance.

Liu et al. (2021) [23] focus on judicial case summarization in China, highlighting the time-consuming manual process undertaken by legal professionals. They propose an automatic generation approach using the GPT-2 pre-trained model, which excels in text generation. The paper divides datasets using extractive algorithms and compresses and integrates them using abstractive algorithms, achieving promising results in summarizing court verdicts efficiently.

Lin, N., et al. (2022) [24] Automatic Text Summarization (ATS) involves abstractive and extractive methods. Due to a lack of corpus, abstractive summarization performs poorly in low-resource language ATS tasks. The study introduces an extractive method for Indonesian documents using the Light Gradient Boosting Machine regression model. The method, considering features like PositionScore and semantic representation similarity, outperforms other models in F1 scores of ROUGE-1, ROUGE-2, and ROUGE-3. Future research is suggested to explore deep learning for automatic feature extraction in Indonesian summarization tasks.

Hellesoe (2022) [25] acknowledges the critical role of text summarization in Natural Language Processing, adopted across various domains such as news, articles, and legal documents. Despite its wide use, controlling domain-specific summaries poses a challenge, often neglecting crucial domain-specific information. The thesis proposes a novel approach for domain-specific automatic text summarization, introducing a hybrid model with domain, focus, and context embeddings. The model, evaluated on MeQSum and LegalCosts datasets, outperforms state-of-the-art algorithms, enhancing automation and summary quality.

Alahmadi et al. (2022) [26] emphasize the importance of abstractive text summarization for creating natural language summaries. While English models excel, Arabic summarization faces challenges like syntax inconsistency, leading to low-accuracy summaries. The paper introduces a novel approach, adding topic awareness to a summarizer, resulting in the Topic-Aware Abstractive Arabic Summarization Model (TAAM). Quantitative experiments show TAAM achieves 10.8% higher accuracy than baseline models based on ROUGE matrices. Qualitatively, TAAM produces coherent and reader-friendly Arabic summaries, capturing the main ideas effectively.

Agrawal et al. (2023, February) [27] discuss the significant growth of artificial intelligence over the past decade, with a focus on AI-related text summarization. The paper presents a method utilizing BERT for embeddings to create a succinct and accurate extractive summary. The approach involves classifying each sentence in a document to determine its inclusion in the summary. The sentences are ranked based on prediction scores, and the best-scoring sentences form the final summary. CNN/Daily Mail news articles are used for validation, emphasizing factors like order of appearance and grammatical correctness in summary generation.

Mishra et al. (2023) [28] address the challenge of efficiently storing and retrieving useful information from large text documents. They propose summarizing massive text documents using text mining principles to extract important information. The paper explores various extractive approaches for text summarization and introduces metrics for evaluating the resulting summaries, emphasizing the reduction in size from the original text.

Table 1. Literature Survey

| Author Name | Year | Major Concept | Findings |
|---|---|---|---|
| Alquliti, W. H., & Ghani, N. B. A. | 2019 | CNN-ATS for Automatic Text Summarization | Introduced CNN-ATS, a convolutional neural network for text summarization. Evaluated depth variations, determining improved configurations for informative sentence extraction. Experimented with 26 configurations, achieving potential for better summaries compared to other methods using DUC 2002 data. |
| Al Oudah, A., et al. (2019) | 2019 | Arabic Text Summarization, Wajeez, Scoring Formula | Addresses the challenge of accessing accurate Arabic information, introduces Wajeez with a new scoring formula. Wajeez performs well, especially with title support. |
| Gupta, A., & Kaur, M. (2019) | 2019 | Outdegree Centrality (OC), Laplacian Centrality (LC) | Discusses bias in OC for generic summaries, proposes using LC for specific summaries. Modified method shows significant improvement in informativeness and coherence. |
| Zaman, F., et al. | 2020 | Hybrid Architecture for Simplification and Summarization | Explores the relationship between text simplification and summarization. Introduces the HTSS model, a hybrid of abstractive and extractive summarization. Outperforms NTS and ATS, showcasing superior performance in joint tasks. Introduces CSS1 metric. Utilizes a parallel corpus from EurekaAlert. |
| Fang, W., et al. | 2020 | LSTM-based Model for Chinese Article Summarization | Proposed an LSTM-based model for Chinese article summarization. Demonstrated good summarization performance but highlighted the need for further research to improve sentence organization. |
| Mollá, D., et al. (2020) | 2020 | BioASQ Challenge, Extractive Summarization, BERT, BioBERT | Macquarie University and ANU participation, Framework for query-focused extractive summarization. Best results with BERT and LSTM. Siamese architectures and BioBERT variants did not improve results. |
| Xu, X., et al. (2020, July) | 2020 | Abstractive Summarization, Evaluation Metric | Introduces a new metric based on fact-level content weighting for evaluating abstractive summarization. Correlates well with human perception, outperforming a recent highlight-based metric. |
| Lal, N. M., et al. K. | 2021 | Novel Approach in Extractive Text Summarization | Proposed a novel approach in Extractive text summarization using a new sentence scoring parameter. Experimental results showed improved performance compared to other summarization models, validated through ROUGE-1 and F1 scores. |
| Varab, D., & Schluter, N. | 2021 | Multilingual Summarization Dataset | Presents the largest and most inclusive multilingual summarization dataset, addressing the anglo-centric bias. Investigates resource building in low-resource language settings and assesses the impact on summarization system performance. |
| Liu, J., et al. | 2021 | Judicial Case Summarization in China | Explores automatic generation of judicial case summaries using the GPT-2 pre-trained model. Proposes a method involving text compression and integration, achieving effective summarization of court verdicts. |
| Lin, N., et al. | 2022 | Extractive Summarization for Indonesian Documents | Proposed an extractive method based on the Light Gradient Boosting Machine regression model for Indonesian documents. Utilized features like PositionScore, TitleScore, and semantic representation similarity. Outperformed other models in F1 scores of ROUGE-1, ROUGE-2, and ROUGE-3. Suggested future research for automatic feature extraction using deep learning. |
| Hellesoe, L. J. | 2022 | Domain-Specific Automatic Text Summarisation | Proposes a hybrid model using domain, focus, and context embeddings for domain-specific summarisation. Outperforms state-of-the-art algorithms in MeQSum and LegalCosts datasets, enhancing automation and summary quality. |
| Alahmadi, D., et al. | 2022 | Topic-Aware Abstractive Arabic Summarisation (TAAM) | Introduces TAAM, a novel topic-aware abstractive Arabic summarisation model, addressing syntax inconsistency issues. Achieves 10.8% higher accuracy than baseline models based on ROUGE matrices. Produces coherent and readable Arabic summaries capturing the main idea. |
| Agrawal, A., et al. | 2023 | AI-related Text Summarization, BERT | Introduces an approach using BERT embeddings for extractive summarization. Classifies sentences to generate summaries, with validation on CNN/Daily Mail news articles. |
| Mishra, A. R., et al. | 2023 | Text Summarization, Text Mining Principles | Addresses efficient storage and retrieval of information from large text documents through summarization. Explores various extractive approaches and introduces evaluation metrics. |

## Research Gaps

- Integration of Advanced NLP Techniques: While the previous researches acknowledge the pivotal role of NLP in previous researches summarization, there is a lack of specific exploration into how advanced NLP techniques, such as deep learning-based language models (e.g., GPT, BERT), can be effectively integrated into summarization systems to enhance their performance.

- Optimization of NLP Techniques for Summarization: There is a gap in research concerning the optimization of NLP techniques specifically for previous researches summarization tasks. This includes investigating how syntactic and semantic analysis, part-of-speech tagging, and sentiment analysis can be tailored or combined to better suit the unique requirements of summarization.

- Contextual Understanding in Summarization: While the previous researches mention that leveraging NLP introduces a layer of understanding beyond word counting, there is a research gap in exploring how NLP can be utilized to improve contextual understanding in summarization. This could involve research into contextual representation learning, discourse analysis, or document-level understanding.

- Nuanced Summarization Techniques: The previous researches suggests that NLP techniques contribute to more nuanced and contextually aware summarization. However, there is a gap in understanding how to effectively incorporate this nuanced understanding into summarization systems to generate summaries that are not only concise but also accurately capture the key information and nuances of the original previous researches.

- Evaluation and Benchmarking: There is a need for further research into the development of robust evaluation metrics and benchmark datasets specifically tailored for assessing the performance of NLP-based summarization systems. Current evaluation metrics may not fully capture the quality and nuances of summaries generated using advanced NLP techniques.

## V. Conclusion

In conclusion, this comprehensive overview has delved into the transformative capabilities of Natural Language Processing (NLP) for text summarization. The growing volume of information across diverse domains necessitates efficient and accurate summarization tools, making NLP a pivotal technology in this realm. The paper has explored key techniques, including both extractive and abstractive approaches, shedding light on their methodologies. Despite the advancements, challenges and opportunities persist in the field.

This overview not only provides insights into the current state of NLP-based summarization but also offers a glimpse into potential future advancements. The evolving landscape of NLP in text summarization holds promise for addressing complex information processing needs. By unraveling the intricacies of these techniques, this paper aspires to guide researchers, practitioners, and enthusiasts toward a deeper understanding, fostering further exploration and innovation in this dynamic field..

## References

1. Yang, B., Luo, X., Sun, K., & Luo, M. Y. (2023, August). Recent progress on text summarisation based on bert and gpt. In *International Conference on Knowledge Science, Engineering and Management* (pp. 225-241). Cham: Springer Nature Switzerland.
2. Rathi, K., Raj, S., Mohan, S., & Singh, Y. V. (2022). A Review of state-of-the-art Automatic Text Summarisation. *International Journal of Creative Research Thoughts (2022)*.
3. Barrantes, M., Herudek, B., & Wang, R. (2020). Adversarial nli for factual correctness in text summarisation models. *arXiv preprint arXiv:2005.11739*.
4. Suleiman, D., & Awajan, A. (2020). Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges. *Mathematical problems in engineering*, *2020*, 1-29.
5. Awasthi, I., Gupta, K., Bhogal, P. S., Anand, S. S., & Soni, P. K. (2021, January). Natural language processing (NLP) based text summarization-a survey. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)* (pp. 1310-1317). IEEE.
6. Rezazadegan, D., Berkovsky, S., Quiroz, J. C., Kocaballi, A. B., Wang, Y., Laranjo, L., & Coiera, E. (2020). Automatic speech summarisation: A scoping review. *arXiv preprint arXiv:2008.11897*.
7. Gao, Y., Meyer, C. M., & Gurevych, I. (2020). Preference-based interactive multi-document summarisation. *Information Retrieval Journal*, *23*, 555-585.
8. Golec, J., Hachaj, T., & Sokal, G. (2021). TIPS: A Framework for Text Summarising with Illustrative Pictures. *Entropy*, *23*(12), 1614.
9. Mridha, M. F., Lima, A. A., Nur, K., Das, S. C., Hasan, M., & Kabir, M. M. (2021). A survey of automatic text summarization: Progress, process and challenges. *IEEE Access*, *9*, 156043-156070.
10. Egger, R., & Gokce, E. (2022). Natural Language Processing (NLP): An Introduction: Making Sense of Textual Data. In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications* (pp. 307-334). Cham: Springer International Publishing.
11. Zerva, C., Nghiem, M. Q., Nguyen, N. T., & Ananiadou, S. (2020). Cited text span identification for scientific summarisation using pre-trained encoders. *Scientometrics*, *125*, 3109-3137.
12. Ghanbari Haez, S., & Shamsfakhr, F. (2022). Extractive text summarisation using Bayesian state estimation of sentences: A Markovian framework. *Journal of Information Science*, 01655515221112842.
13. Al Abdulwahid, A. (2023). Software solution for text summarisation using machine learning based Bidirectional Encoder Representations from Transformers algorithm. *IET SOFTWARE*.
14. Alquliti, W. H., & Ghani, N. B. A. (2019). Convolutional neural network based for automatic text summarization. *International Journal of Advanced Computer Science and Applications*, *10*(4).

15. Al Oudah, A., Al Bassam, K., Kurdi, H., & Al-Megren, S. (2019). Wajeez: An Extractive Automatic Arabic Text Summarisation System. In *Social Computing and Social Media. Design, Human Behavior and Analytics: 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part I 21* (pp. 3-14). Springer International Publishing.
16. Gupta, A., & Kaur, M. (2019). Text Summarisation Using Laplacian Centrality-Based Minimum Vertex Cover. *Journal of Information & Knowledge Management*, *18*(04), 1950050.
17. Zaman, F., Shardlow, M., Hassan, S. U., Aljohani, N. R., & Nawaz, R. (2020). HTSS: A novel hybrid text summarisation and simplification architecture. *Information Processing & Management*, *57*(6), 102351.
18. Fang, W., Jiang, T., Jiang, K., Zhang, F., Ding, Y., & Sheng, J. (2020). A method of automatic text summarisation based on long short-term memory. *International Journal of Computational Science and Engineering*, *22*(1), 39-49.
19. Mollá, D., Jones, C., & Nguyen, V. (2020). Query focused multi-document summarisation of biomedical texts. *arXiv preprint arXiv:2008.11986*.
20. Xu, X., Dušek, O., Li, J., Rieser, V., & Konstas, I. (2020, July). Fact-based content weighting for evaluating abstractive summarisation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5071-5081).
21. Lal, N. M., Krishnanunni, S., Vijayakumar, V., Vaishnavi, N., Siji Rani, S., & Deepa Raj, K. (2021). A novel approach to text summarisation using topic modelling and noun phrase extraction. In *Advances in Computing and Network Communications: Proceedings of CoCoNet 2020, Volume 2* (pp. 285-298). Springer Singapore.
22. Varab, D., & Schluter, N. (2021, November). MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10150-10161).
23. Liu, J., Wu, J., & Luo, X. (2021). Chinese judicial summarising based on short sentence extraction and GPT-2. In *Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Tokyo, Japan, August 14–16, 2021, Proceedings, Part II 14* (pp. 376-393). Springer International Publishing.
24. Lin, N., Li, J., & Jiang, S. (2022). A simple but effective method for Indonesian automatic text summarisation. *Connection Science*, *34*(1), 29-43.
25. Hellesoe, L. J. (2022). *Automatic Domain-Specific Text Summarisation With Deep Learning Approaches* (Doctoral dissertation, Auckland University of Technology).
26. Alahmadi, D., Wali, A., & Alzahrani, S. (2022). TAAM: Topic-aware abstractive arabic text summarisation using deep recurrent neural networks. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 2651-2665.
27. Agrawal, A., Jain, R., Divanshi, & Seeja, K. R. (2023, February). Text Summarisation Using BERT. In *International Conference On Innovative Computing And Communication* (pp. 229-242). Singapore: Springer Nature Singapore.
28. Mishra, A. R., Naruka, M. S., & Tiwari, S. (2023). Extraction Techniques and Evaluation Measures for Extractive Text Summarisation. In *Sustainable Computing: Transforming Industry 4.0 to Society 5.0* (pp. 279-290). Cham: Springer International Publishing.