AUTOMATIC MEDICAL TRANSCRIPTS SUMMARIZATION USING MACHINE LEARNING TECHNIQUES

A Thesis Submitted in partial fulfillment of the Requirements for the Award of the degree of

DOCTOR OF PHILOSOPHY

by

PARMINDER PAL SINGH BEDI

(2K18/PhD/IT/06)

Under the Supervision of

Prof. KAPIL SHARMA
Professor
Department of Information Technology
Delhi Technological University

Dr. MANJU BALA
Associate Professor
Computer Science Department
I.P. College for Women, University of Delhi

To The

Department of Information Technology

Delhi Technological University

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042 (India)

May, 2024

CANDIDATE DECLARATION

I hereby declare that the thesis entitled "Automatic Medical Transcripts Summarization using Machine Learning Techniques" submitted to Delhi Technological University, Delhi, in the partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy in the Department of Information Technology is an original work and has been done by myself under the joint supervision of Prof. Kapil Sharma, Department of Information Technology, Delhi Technological University, Delhi, India and Dr. Manju Bala, Associate Professor, I.P. College for Women, University of Delhi.

The interpretations presented are based on my study and understanding of the original texts. The work reported here has not been submitted to any other institute for the award of any other degree.

Parminder Pal Singh Bedi
2K18/Ph.D/IT/06
Department of Information Technology
Delhi Technological University
Delhi-110042, India

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
(Govt. of National Capital Territory of Delhi)
Shahbad Daulatpur, Main Bawana Road,
Delhi-110042, India

Date: 29/08/2004

CERTIFICATE

This is to certify that the work incorporated in the thesis entitled "Automatic Medical transcripts Summarization using Machine Learning Techniques" submitted by Parminder Pal Singh Bedi in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy, to the Delhi Technological University, Delhi, India is carried out by the candidate under our supervision and guidance at the Department of Computer Science and Engineering, Delhi Technological University, Delhi, India.

The results embodied in this thesis have not been presented to any other University or Institute for the award of any degree or diploma.

Supervisors:

Prof. Kapil Sharma
Professor
Dept. of Information Technology,
Delhi Technological University

Dr Manju Bala
Associate Professor
Dept. of Computer Science,
I.P. College for Women, University of Delhi

ACKNOWLEDGMENT

I extend my sincere thanks to Almighty God for giving me the inner power to complete my thesis and guide me in every step of my life.

It is an immense pleasure to have the opportunity to express my heartiest gratitude to everyone who helped me throughout this research journey. With immense joy and heartfelt gratitude, I would like to extend my indebtedness to my supervisors, Prof. Kapil Sharma (Professor, Dept. of Information Technology, DTU) and Dr. Manju Bala, (Associate Professor, I.P College for women, University of Delhi) for their invaluable guidance, mentorship, encouragement, and patience. During the research, their motivation and encouragement have made me strive to work harder to achieve my goals. I am deeply humbled and indebted to my supervisors for continually motivating me to persevere and making me believe in myself during times of hardship. Their technical expertise, precise suggestions, kind nature, and detailed timely discussions are greatly appreciated.

Also, I sincerely thank Delhi Technological University for considering my candidature for this course. I am also very thankful to Prof.. Prateek Sharma, Vice-Chancellor, Delhi Technological University, Delhi, India, for being a constant source of enthusiasm. He has always motivated young researchers like me to pursue excellence to achieve higher goals in academics and research. Also, my sincere thanks reciprocate to Prof. Dinesh Kumar Vishwakarma (HoD, Dept. of Information Technology and Chairperson DRC, Dept. of Information Technology) for insightful comments and valuable suggestions. Special thanks to my seniors and colleagues of Delhi Technological University, Delhi, India. My sincere thanks to all the professors, faculty, researchers, and nonteaching staff of the Information Technology Department.

I also wish to take this opportunity to thank all my teachers who have taught me and shaped me into the person I am, aggravated me to be an academician, and have directly indirectly made me capable of succeeding in completing this research work. I am deeply thankful to all my colleagues and friends during my journey as a Ph.D. scholar. The engaging discussions, brainstorming sessions, and collaborative teamwork significantly impacted my growth as an independent researcher.

I would also like to thank my wife who always supported me in all my endeavors and believed in me and encouraged me in all the challenging times. Finally, and most

importantly, I would like to express my deepest gratitude to my parents who stood by me like a pillar of strength and always supported me to realize my goals. I will cherish their utmost love and blessings throughout my life.

Parminder Pal Singh Bedi

2K18/Ph.D/IT/06

Department of Information Technology

Delhi Technological University,

Delhi-110042, India

ABSTRACT

The healthcare sector and biomedical domain are essential for public health and medical advancement, providing services from clinical care to research. Healthcare facilities offer crucial services like check-ups and disease management, while the biomedical domain drives medical innovation through research and experimentation. With the increasing volume of biomedical literature, automatic text summarization is vital for efficiently extracting insights. These algorithms, equipped with domain-specific knowledge, simplify complex information, facilitating knowledge dissemination and collaboration. Additionally, in the rapidly evolving field of biomedical research, automatic summarization systems ensure timely access to up-to-date information by monitoring and summarizing the latest literature and databases. There are two main approaches of Automatic Text Summarization: Extractive and Abstractive. Extractive summarization involves selecting and extracting specific sentences or phrases directly from the source text, prioritizing their frequency or relevance to compose the summary. In contrast, Abstractive summarization interprets and paraphrases the content to create new sentences conveying the essential meaning in a concise form.

In this research work, extractive text summarization techniques in biomedical domain are explored, focusing on issues such as redundancy, coherence, and the risk of overlooking crucial information. Extractive summarization techniques in the biomedical domain utilize various algorithms and approaches, including Frequency-based Methods, Graph-based Algorithms, and Machine Learning Approaches, to identify and extract key sentences or phrases from biomedical documents. Hybrid approaches combine multiple techniques to improve accuracy and coverage, effectively summarizing complex biomedical texts while addressing challenges such as redundancy and information loss.

To address the identified research gaps, numerous novel approaches have been proposed for biomedical text summarization. Firstly, a novel approach using the Methathesaurus from UMLS to extract named entity concepts is proposed which applies the BERT method to generate concise summaries from Pubmed and Mtsamples. Further, an unsupervised approach focusing on semantic similarity and keyword-phrase extraction for both single-document and multi-document summarization is proposed. Furthermore, to further improve

upon the results, a distinctive framework utilizing deep neural networks for contextually aware summarization of biomedical literature is proposed which employs a binary classifier and bidirectional long-short term memory recurrent neural network.

To validate the proposed approaches, comparisons are made with baseline methods in biomedical text summarization, including a recent graph-based approach with the FP-Growth method. The results indicate that the last proposed approach outperforms state-of-the-art methods, achieving the highest ROUGE score of 0.96, surpassing the scores of the first and second approach (0.74, 0.76).

The research concludes that the proposed methods demonstrate superior results in the medical domain compared to existing state-of-the-art techniques, highlighting the efficacy of the developed summarization approaches for biomedical literature.

Table of Contents

LIST OF ABBREVIATIONS

Abbreviation	Expansion
NLP	Natural Language Processing
AI	Artificial Intelligence
ATS	Automatic Text Summarization
PCFG	Probabilistic Context-Free Grammars
HMM	Hidden Markov Models
SVM	Support Vector Machine
LSA	Latent Semantic Analysis
SVD	Singular Value Decomposition
TF-IDF	Term Frequency-Inverse Document Frequency
MSE	Mean Squared Error
OOB	Out-of-Bag
DUC	Document Understanding Conferences
TREC	Text Retrieval Conference
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
AQUAINT	Advanced Question Answering for Intelligence
UMLS	Unified Medical Language System
CSO	Cat Swarm Optimization
MAP	Mean Average Precision
PERSIVAL	Personalized Search and Summarization in a Virtual Library
EBM	Evidence-Based Medicine
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
PoS	Part-of-Speech
MT	Medical Transcription
ENT	Ear, Nose, and Throat
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
MLM	Masked Language Modeling
UML	Unified Medical Language
CUI	Concept Unique Identifier
ST	Semantic Type
GELU	Gaussian Error Linear Unit
CBOW	Continuous Bag of Words
HITS	Hyperlink-Induced Topic Search
PPF	Prepositional Phrase Frequency

PMC	PubMed Central
RAKE	Rapid Automatic Keyword Extraction
DTM	Document Term Matrix
MRI	Magnetic Resonance Imaging
EEG	Electroencephalogram
CNS	Central Nervous System
CVA	Cerebrovascular Accident
ECG	Electrocardiogram
ICU	Intensive Care Unit
ACLS	Advanced Cardiovascular Life Support
EMS	Emergency Medical Services
CAD	Coronary Artery Disease
CKD	Chronic Kidney Disease
COPD	Chronic Obstructive Pulmonary Disease
HIV	Human Immunodeficiency Virus
HSV	Herpes Simplex Virus
MMF	Mandibulomaxillary Fixation
PACU	Post-Anesthesia Care Unit
FP-Growth	Frequent Pattern Growth Algorithm
TexLexAn	Textual Lexical Analysis
CNN	Convolutional Neural Network
ELMo	Embeddings from Language Models
BN	Batch Normalization
RNN	Recurrent Neural Networks
TSE	Total Squared Error
RAKE	Rapid Automatic Keyword Extraction
PEGASUS	Pre-training with Extracted Gap-sentences for Abstractive Summarization
Pubmed	PubMed Central

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. Introduction

Automatic Text Summarization, related to Artificial Intelligence and natural Language processing, is a computational procedure aimed at generating concise and well-organized summaries from provided texts or documents. Its primary goal is to extract critical information from the source text while preserving its core meaning and context. This technology is efficient in managing huge data, facilitating tasks such as information retrieval, document categorization, and facilitating rapid comprehension of textual content.

Automatic text summarization contributes to tasks such as information retrieval, document categorization, and rapid comprehension of textual content. The prime task is content selection, that identifies the most relevant and important information from the source text to include in the summary. This task requires algorithms to analyse the content of the document, identify key sentences or passages, and determine their significance in relation to the overall context. Content selection is essential for ensuring that the summary accurately reflects the main themes and key points of the original text. Another task is the summarization of lengthy documents. This task involves condensing the content of lengthy documents into shorter, more manageable summaries while preserving the essential meaning and context. Summarizing lengthy documents is challenging due to the volume of information involved and the need to prioritize and condense the content effectively. Techniques such as sentence extraction and abstraction are commonly used to generate concise summaries of lengthy documents [1].

Automatic summarization is utilized across various domains to effectively process and condense large volumes of textual information into concise summaries. Some of the key domains where automatic summarization is extensively applied include:

- **News and Media**: In the fast-paced world of journalism and media, automatic summarization helps to generate succinct summaries of news articles, reports, and updates. It enables readers to quickly grasp the main points of a story without having to read through lengthy articles, facilitating efficient information consumption.

- **Research and Academia**: In academic and research settings, automatic summarization aids in summarizing lengthy research papers, articles, and journals. Researchers can use automatic summarization tools to quickly extract key findings, methodologies, and conclusions from vast amounts of scholarly literature, thus facilitating literature review processes and aiding in knowledge dissemination.
- **Business and Market Intelligence**: In the business domain, automatic summarization assists in analysing market trends, competitor reports, and business intelligence data. It enables companies to extract relevant insights and actionable information from large datasets, helping decision-makers to make informed strategic decisions and stay competitive in their respective industries.
- **Legal and Compliance**: In the legal sector, automatic summarization is used to summarize legal documents, court cases, contracts, and regulatory compliance documents. It helps legal professionals to extract essential details, precedents, and key arguments from lengthy legal texts, saving time and improving productivity in legal research and case preparation.
- **Healthcare and Medical**: In the healthcare domain, automatic summarization aids in summarizing medical records, patient histories, research articles, and clinical trial reports. It assists healthcare professionals in quickly accessing relevant patient information, medical findings, and treatment outcomes, thereby improving decision-making processes and patient care delivery.
- **Social Media and Content Curation**: With the proliferation of social media platforms and user-generated content, automatic summarization is used to summarize social media posts, comments, and discussions. It enables users to quickly skim through relevant information, identify trending topics, and curate personalized content feeds based on their interests and preferences.
- **Educational Technology**: It is used to summarize educational materials, textbooks, and lecture notes. It helps students and educators to distal complex information into concise summaries, facilitating learning comprehension, revision, and knowledge retention.
- **Customer Support and Feedback Analysis**: In customer service and feedback analysis, automatic summarization assists in summarizing customer reviews, feedback surveys, and support tickets. It enables businesses to identify common

themes, sentiments, and actionable insights from customer feedback, helping them to improve products, services, and customer experiences.

- **The biomedical domain** encompasses a vast and intricate landscape of scientific research, clinical studies, patient records, and scholarly literature, generating an overwhelming volume of data and information.

These are just a few examples of the diverse domains where automatic summarization finds application. Its versatility and effectiveness in processing textual data make it a valuable tool across various industries and sectors, enabling efficient information extraction, analysis, and decision-making. Fig 1.1 shows the summarization process [1].

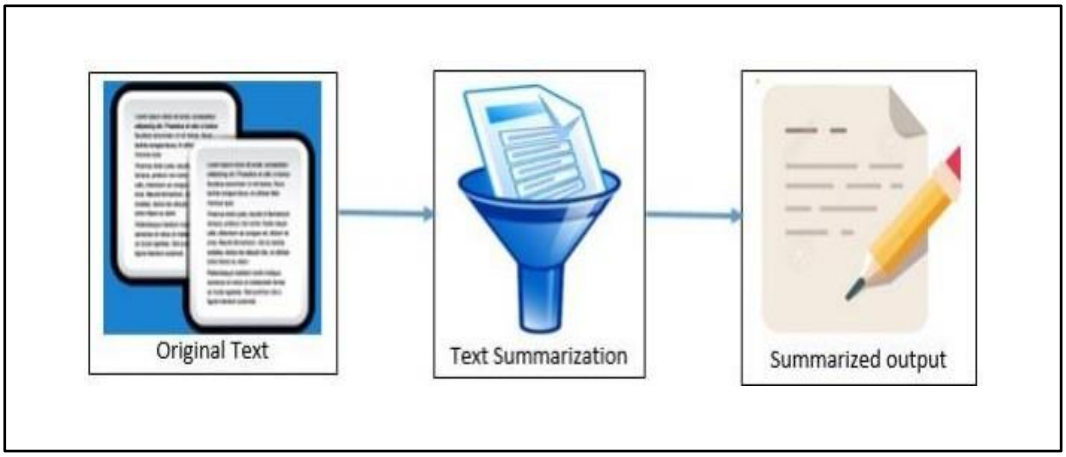


Fig 1.1 Text Summarization for large number of documents [1]

1.2. Automatic text summarization

Automatic text summarization encompasses two main approaches: **Extractive and Abstractive**.

Extractive summarization involves selecting and extracting specific sentences or phrases directly from the source text to compose the summary. These selected excerpts are typically deemed as representative of the essential information contained within the document. The primary goal is to preserve the original meaning and context of the text while condensing it into a shorter form. This approach relies on algorithms that analyse the content of the document based on various criteria such as importance, relevance, and coherence. Extractive summarization is advantageous in that it ensures that the summary accurately reflects the content of the original text. However, it may struggle with generating coherent and cohesive summaries, especially when dealing with complex or lengthy documents [2], [3], [4]. Whereas, **abstractive summarization** aims to generate

summaries by interpreting and paraphrasing the content of the source text in a more human-like manner. Instead of merely extracting existing sentences, it rephrases and synthesizes the information to create new sentences that convey the essential meaning of the text in a concise form. This approach utilizes natural language generation approaches such as neural networks and deep learning models, to generate summaries that are not limited to the exact wording of the original text. Abstractive summarization has the advantage of producing more fluent and coherent summaries compared to extractive techniques. However, it also poses challenges in accurately capturing the intended meaning of the original text and ensuring grammatical correctness in the generated summaries [5], [6].

Further, Summarization can also be classified as either mono-document or multi document.

- **Mono-document summarization-** It summarizes a single document or text at a time. The goal of is to condense the content of a single document into a shorter form while retaining its essential information and main points. This approach is commonly used when dealing with individual articles, reports, or documents where the goal is to provide a concise overview of the document's content. Mono-document summarization techniques typically involve analysing the content of the document, identifying key sentences or passages, and selecting the most relevant information to include in the summary. The resulting summary provides a brief and coherent representation of the original document's content [7].
- **Multi-document summarization**- It involves summarizing multiple documents or texts on a similar topic or theme. The objective is to synthesize information from multiple sources into a single, concise summary that captures the main points and key findings across the documents. This approach is useful when dealing with a large volume of documents, such as news articles, research papers, or online sources, where the goal is to distil information from multiple sources into a condensed form. Multi-document summarization techniques typically involve clustering related documents, identifying common themes or topics, and extracting relevant information from each document to create a comprehensive summary. The resulting summary provides an overview of the main ideas and findings across multiple documents, enabling readers to grasp the key information without having to read each document individually [8], [9]. Additionally, summarization can be **Generic or**

Query-based, each serving different purposes and employing different methodologies.

- **Generic summarization** focuses on generating a summary that provides a comprehensive overview of the set of documents, without any specific query or question guiding the summarization process. The main objective is to distil the main points, key ideas, and essential information from the source text(s) into a concise and coherent summary. This approach is commonly used in scenarios where the goal is to provide a general understanding of the content, such as summarizing news articles, research papers, or long documents. Generic summarization techniques typically involve analysing the content of the document(s), identifying important sentences or passages, and synthesizing the information to create a summary that captures the main themes and key points [10].
- On the other hand, **Query-based summarization** involves generating a summary that directly addresses a specific query or question posed by the user. The objective of query-based summarization is to extract information relevant to the query from the source text(s) and present it in a concise and informative manner. This approach is particularly useful in scenarios where the user is seeking specific information or answers to specific questions, such as summarizing search engine results, user-generated content, or FAQs. Query-based summarization techniques typically involve analysing the query, identifying relevant passages or sentences from the source text(s) that contain the information needed to answer the query, and synthesizing the information into a summary that directly addresses the user's query [11], [12].

Additionally, the evaluation of automatically generated summaries is an important aspect of automatic text summarization. Evaluating the quality and effectiveness of summaries generated by algorithms is crucial for assessing their performance and identifying areas for improvement. Various metrics and evaluation criteria, such as Recall-Oriented Understudy for Gisting Evaluation, Bilingual Evaluation Understudy and human judgment, are used to evaluate the coherence, relevance, and informativeness of automatic summaries.

Understanding the different types of tasks involved in automatic text summarization, as well as the challenges associated with each task, contributes to a comprehensive understanding of current methodologies and future directions in the field. By addressing

these challenges and developing novel techniques, researchers can continue to advance the state-of-the-art in automatic text summarization, enabling more effective and efficient processing of textual information.

1.3. Healthcare Sector and Biomedical Domain

The healthcare sector and biomedical domain are integral components of society, playing pivotal roles in safeguarding public health, advancing medical knowledge, and enhancing overall well-being. These sectors encompass a wide range of activities, including clinical care, medical research, disease prevention, and health promotion, all aimed at addressing diverse health needs and challenges. First and foremost, the healthcare sector serves as the cornerstone of public health by providing essential medical services, treatments, and preventive care to individuals and communities. From primary care clinics to specialized hospitals, healthcare facilities offer a spectrum of services, ranging from routine check-ups and vaccinations to surgical interventions and chronic disease management. Through these interventions, healthcare professionals diagnose and treat illnesses, alleviate suffering, and promote healthy behaviours, ultimately improving health outcomes and enhancing quality of life. Moreover, the biomedical domain plays a crucial role in driving medical research and innovation, leading to groundbreaking discoveries and advancements in healthcare. Biomedical researchers and scientists engage in rigorous scientific inquiry, clinical trials, and experimentation to develop new drugs, therapies, medical devices, and treatment protocols. These innovations contribute to the development of more effective and targeted treatments for various diseases and conditions, ultimately saving lives and improving patient outcomes. Furthermore, the healthcare sector and biomedical domain are instrumental in disease prevention and control efforts, addressing public health challenges on a global scale. Through epidemiological surveillance, vaccination campaigns, and health education initiatives, these sectors strive to prevent the spread of infectious diseases, reduce morbidity and mortality rates, and promote healthy lifestyles. Additionally, they play a critical role in addressing emerging health threats, such as pandemics and epidemics, by coordinating response efforts, conducting research, and disseminating vital information to the public. Equally important is the role of the healthcare sector and biomedical domain in promoting healthcare access and equity. Access to affordable, high-quality healthcare services is essential for ensuring that all individuals, regardless of socioeconomic status or background, receive timely and

appropriate medical care. By advocating for health equity, expanding insurance coverage, and reducing barriers to care, these sectors work towards addressing health disparities and achieving equitable health outcomes for diverse populations [13].

Within this context, the need for automatic summarization in the biomedical domain arises from several critical factors. Firstly, the sheer volume of biomedical literature and research outputs has escalated exponentially in recent years, making it increasingly challenging for researchers, clinicians, and healthcare professionals to navigate and extract relevant insights efficiently. With thousands of new research articles, clinical trials, and medical reports published daily, manual review and synthesis of this vast corpus of information become impractical and time-consuming. Automatic summarization offers a solution by condensing lengthy texts into concise summaries, enabling researchers to quickly grasp the key findings, methodologies, and implications of scientific studies. Fig. 1.2. Shows the summarization in biomedical domain.

Secondly, the complexity and technical nature of biomedical literature present unique challenges for information retrieval and comprehension. Biomedical texts often contain specialized terminology, complex scientific concepts, and intricate experimental methodologies that may be challenging for non-experts to decipher. Automatic summarization algorithms, equipped with domain-specific knowledge and linguistic models, can effectively distil and simplify this complex information into digestible summaries, facilitating knowledge dissemination and interdisciplinary collaboration within the biomedical community.

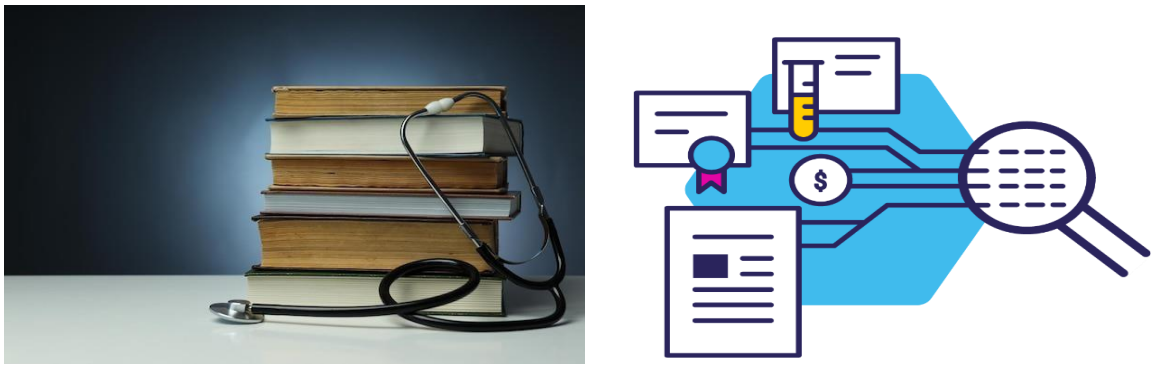


Fig 1.2. Summarization in biomedical domain [14]

Moreover, the rapid pace of biomedical research and clinical practice demands timely access to up-to-date information and evidence-based insights. With new discoveries, treatment

protocols, and medical guidelines emerging regularly, healthcare professionals and researchers require efficient mechanisms for staying abreast of the latest developments in their respective fields. Automatic summarization systems can continuously monitor and summarize the latest research literature, clinical trials, and medical databases, providing real-time updates and actionable insights to support evidence-based decision-making and clinical practice. Furthermore, the biomedical domain encompasses diverse stakeholders with varying information needs and preferences. Clinicians may seek succinct summaries of treatment guidelines and diagnostic protocols to inform patient care, while researchers may require comprehensive reviews of literature to inform experimental design and hypothesis formulation. Automatic summarization techniques can cater to these diverse needs by generating tailored summaries tailored to the specific requirements and expertise of different user groups, thereby enhancing information accessibility and usability across the biomedical community.

In this thesis, the intricacies encountered by extractive techniques in biomedical domain are explored, especially concerning issues like redundancy, repetition, coherence, and the risk of overlooking crucial information. Furthermore, it tackles hurdles related to resolving pronouns and references, as well as the intricate management of named entities. The scalability of extractive summarization for handling extensive documents and its applicability across various domains are also carefully analysed, acknowledging the constraints within existing methodologies. Moreover, the research investigates the reliance of extractive methods on sentence length and structure, shedding light on the potential biases introduced by these dependencies during the summarization process. Fig.1.3 shows the long transcripts of bio medical domain and its concise summaries [9].

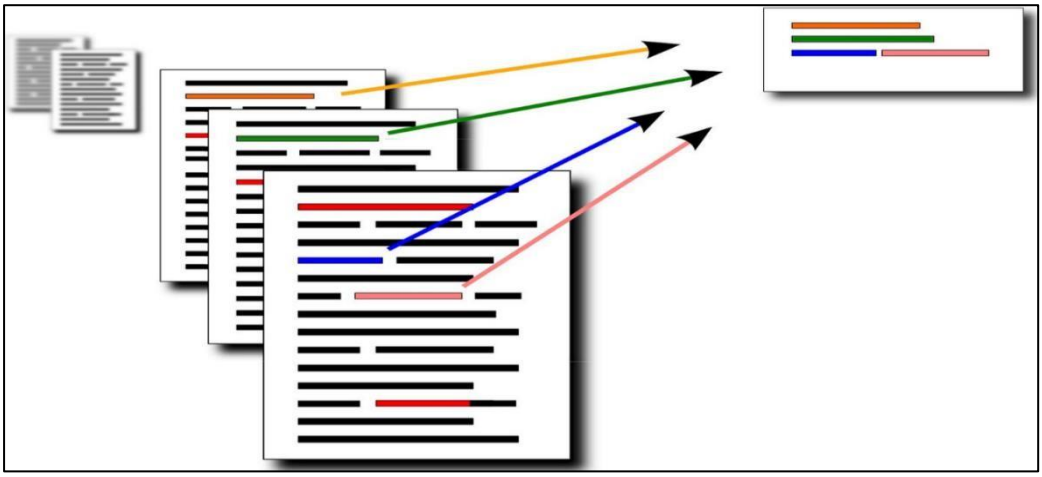


Fig. 1.3. The long transcripts of bio medical domain and its concise summaries [9]

Extractive summarization techniques in the biomedical domain utilize various approaches to identify and extract key phrases from biomedical documents. Some approaches used are Frequency-based Methods which prioritize phrases based on their frequency of occurrence in the document. Sentences containing frequently occurring keywords or terms are more important and are selected for the summary. For example, in biomedical literature, sentences discussing critical concepts or findings may appear more frequently and are thus deemed essential for summarization [5]. Another approach is Graph-based Algorithms that represent the biomedical text as a graph, where sentences or phrases are nodes, and relationships between them are edges [6-8]. Centrality metrics such as degree centrality or betweenness centrality are then used to identify important nodes, which correspond to key sentences or concepts in the document. This approach ensures that sentences with significant connections to other sentences are included in the summary. Apart from this, Machine Learning Approaches are used such as support vector machines and deep learning models. These models are trained on annotated datasets to learn the importance of sentences based on various features, such as keyword frequency, semantic similarity, and syntactic structure [2], [3], [15] - [23]. Once trained, the models can automatically identify and extract important sentences from new biomedical documents to generate summaries. At the end, Hybrid approaches combine multiple techniques, such as statistical methods, graph-based algorithms, and machine learning models, to improve the accuracy and coverage of extractive summarization. By leveraging the strengths of different methods, hybrid approaches can effectively summarize complex biomedical texts while addressing various challenges, such as redundancy and information loss.

These techniques and research contributions demonstrate the diversity and innovation in extractive summarization methods for handling the complexities of biomedical texts and advancing information retrieval in the biomedical domain.

1.4. Motivation

This work is motivated by a strong desire to delve into the complex realm of biomedical text summarization. This driving force is fundamental to our work, pushing us to explore deeper levels of comprehension within this domain.

Primarily, tapping into the extensive reservoir of biomedical data presents significant potential for revealing invaluable insights. This study is rooted in the belief that unlocking this potential depends on effectively summarizing complex biomedical information. By

consolidating intricate narratives into concise summaries, our goal is to equip researchers, practitioners, and decision-makers with actionable knowledge, empowering them to leverage the wealth of insights embedded within biomedical data.

Also, to address the expansive landscape of biomedical knowledge presents a formidable challenge, often resulting in a gap that impedes our comprehensive understanding and utilization of crucial discoveries. Therefore, this work focusses on to overcome this barrier by introducing inventive solutions that not only capture the intricate details within biomedical texts but also present them in a format that is easily accessible and comprehensible. To attain a deeper understanding of biomedical concepts among diverse audiences, bridging the knowledge gap is essential.

Thirdly, the complex nature of biomedical data characterized by intricate terminology and diverse document formats. This motivates to spearhead solutions that directly confront these obstacles. By conducting thorough research and employing novel methodologies, our aim is to develop tools and approaches that not only meet but exceed the demands of biomedical text summarization. Therefore, to address these challenges, various research gaps have been identified in the following section.

1.5. Research Gaps

Despite considerable advancements in biomedical text summarization, there are still significant research gaps and areas for further exploration.

- The limited availability of large, well-annotated datasets that are crucial for training and evaluating summarization models specific to the biomedical domain. The creation of such datasets and the establishment of benchmarks for evaluating summarization systems are areas that require focused attention.
- There is a need to capture intricate details within biomedical texts and also present them in a format that is easily accessible and comprehensible. To overcome barriers that hinder the effective utilization of biomedical information is necessary.
- To integrate domain-specific knowledge such as medical ontologies, biomedical databases, and expert annotations into extractive summarization algorithms is necessitated. Incorporating such knowledge can improve the relevance and accuracy of generated summaries by ensuring a deeper understanding of biomedical concepts and relationships between entities.

- There is a need for the development of specialized summarization techniques tailored to specific subdomains within biomedicine. Given the diverse nature of biomedical literature, which encompasses research articles, clinical notes, and genomics data, there is a demand for customized summarization solutions that can effectively handle the unique characteristics of each subdomain.

1.6. Research Objectives

Literature review is a fundamental step of every research. Based on the detailed literature review and identified research gaps, research objectives were designed particularly focussing on the need for specialized corpus and new methods of text summarization of biomedical data. The objectives are as follows:

- To extract a new corpus for the biomedical domain and to determine the significant features that can be used to generate a summary of biomedical transcripts.
- To process a novel approach for extractive summarization based on unsupervised learning. The approach is based on the concept of semantic similarity and keyword phrase extraction, generating summaries for both single documents and multi-documents.
- Design and implementation of a new approach for extractive summarization using Deep learning techniques to evaluate the performance of the proposed model.
- Comparing results with the state-of-the-art methods in the current domain and providing better solutions with a comprehensive approach.

1.7. Structure of the thesis

Chapter 1 examines the challenges faced by extractive techniques in the biomedical field, and analyses the scalability of extractive summarization for large documents. Various research gaps have been identified and objectives are specified.

Chapter 2 aims to investigate and analyze current text summarization methods applied to the extensive body of unstructured data within the biomedical domain. By examining existing approaches within this specific context, the study seeks to gain insights into their effectiveness and limitations. Through this analysis, the chapter endeavors to identify areas for improvement and potential avenues for further research in biomedical text summarization. The overarching goal is to contribute to the development of more robust and

efficient summarization techniques tailored to the unique challenges posed by biomedical data.

Chapter 3 explains the generic methodology for automatic text summarization and extraction of a novel corpus specifically tailored for the biomedical domain. This involves compiling a comprehensive collection of relevant texts to serve as the foundation for further research in summarization within this domain. Additionally, it aims to identify and analyze significant features present in biomedical transcripts that can be leveraged to generate effective summaries. By examining these features in detail, the study seeks to understand their importance and potential impact on the summarization process. A robust corpus and insights into key features for summary generation is presented.

Chapter 4 presents a novel approach to text summarization. The approach utilizes the Methathesaurus obtained from the Unified Medical Language System (UMLS) to extract concepts associated with named entities and the BERT method is applied. Subsequently, a concise summary for biomedical text data, including samples from Pubmed and Methathesaurus is generated.

Chapter 5 introduces a pioneering method for extractive summarization, utilizing unsupervised learning techniques. This innovative approach capitalizes on the notion of semantic similarity and the extraction of keyword phrases to produce summaries for both individual documents and collections of documents. By harnessing semantic relationships and identifying key phrases, the method aims to generate concise yet informative summaries that capture the essence of the original content. The chapter delves into the intricacies of this approach, detailing its implementation and demonstrating its efficacy through evaluations on both single and multi-document datasets. This research contributes to advancing the field of extractive summarization by offering a novel technique that leverages semantic understanding and keyword extraction for enhanced summarization results.

Chapter 6 presents a distinctive framework capable of intelligent and contextually aware summarization of biomedical literature. Deep neural network binary classifier is developed and bidirectional long-short term memory recurrent neural network is utilised to generate a concise summary of biomedical articles. This research contributes to the advancement of extractive summarization by showcasing the effectiveness of Deep Learning techniques in improving summarization accuracy and quality.

Chapter 7 presents the validation and results of the proposed approaches. In this chapter we present the validation of the proposed methods and comparison with the state-of-the art methods. Rough score is used for the validation of results.

Chapter 8 summarizes the conclusion and future work.

1.8. Conclusion

In this chapter an introduction to text summarization, focusing particularly on its application in the biomedical domain is discussed. It delves into the historical challenges encountered by researchers in this field and outlines the objectives of the proposed research along with potential solutions. The chapter highlights the complexities faced by extractive techniques in biomedical text summarization, including issues like redundancy, repetition, coherence, and the risk of overlooking essential information. It also addresses challenges related to pronoun and reference resolution, as well as the intricate management of named entities. Finally, it identifies research gaps in biomedical text summarization and offers a glimpse into the proposed solutions and chapter structure of the thesis.

CHAPTER 2

LITERATURE REVIEW AND DATA COLLECTION

2.1. Introduction

Automatic Text Summarization, a facet of Natural Language Processing and Artificial Intelligence, focuses on generating brief, coherent summaries from provided texts or documents. Its purpose is to highlight key information while preserving the original context and meaning. This technology plays a vital role in handling vast amounts of data, aiding in tasks like information retrieval, document classification, and rapid understanding of textual content.

There are two main types of summarization techniques: extractive and abstractive. Extractive summarization involves selecting key sentences from the source text, whereas abstractive summarization involves paraphrasing and rephrasing the content to generate more human-like summaries. Additionally, summarization can be divided into mono-document and multi-document types, as well as generic or query-based.

Automatic text summarization involves several key tasks, including content selection, summarizing lengthy documents, and evaluating the quality of generated summaries. A thorough understanding of these tasks, types, and associated challenges is essential for improving summarization methods. Extractive summarization faces issues like redundancy, coherence, and the potential loss of important information. Addressing challenges such as pronoun and reference resolution and the handling of named entities is critical. The scalability and adaptability of extractive summarization across different domains, including news, legal, scientific, and healthcare, are also important considerations.

Early extractive summarization methods focused on identifying important words through their frequency, using techniques like Probabilistic Context-Free Grammars (PCFG), Markov Models, Hidden Markov Models (HMM), Naive Bayes, Clustering, and Support Vector Machines (SVM). However, advancements were needed to tackle evolving challenges, especially with the rise of neural networks. Deep Neural Networks, such as Recurrent Neural Networks with Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), have enabled the generation of abstractive summaries using sequence-to-sequence models.

Evaluating automatically generated summaries can be done manually or automatically. Manual evaluation is time-consuming and costly, while automatic evaluation can be conducted with or without human references. Text summarization is beneficial for reducing data transfer, optimizing resource usage, and enabling quick understanding of large documents.

In the biomedical and healthcare sectors, the increasing volume of textual data, including scientific articles, medical guidelines, clinical trial reports, and health records, presents significant challenges for researchers and clinicians. Platforms like PubMed and Medline contain vast amounts of medical articles. Summarizing electronic health records helps healthcare professionals quickly access essential information. Various statistical and deep learning models, such as the BioMed summarizer developed by M. Afzal et al., have been applied to biomedical text summarization.

Medical data, which includes information about diseases and symptoms, has become more accessible over time through transcripts and other sources. Despite this, extracting precise information remains challenging. Researchers often rely on resources like PubMed, biomedical articles, and research texts. Summarizing real patient transcripts and other medical texts helps overcome these challenges, providing valuable insights for healthcare professionals and researchers.

This chapter is structured as follows: Initially, the distinctions among automatic text summarization categories is elucidated along with the generic methodology of summarization. Subsequently, datasets and corpora used are explained. Thirdly, comparison of key contributions from the state-of-the-art approaches is presented. Lastly, influential works in this field are highlighted.

2.2. Search and Selection Process

The exploration process for this study revolves around automatic extractive summarization within the biomedical domain, and it entailed a methodical examination of scholarly articles and specific conference proceedings spanning the years 1995 to 2022. A comprehensive range of online databases, encompassing reputable sources such as IEEE, ACM Digital Library, Springer, Elsevier, ScienceDirect, IGI Global, Taylor and Francis, IOS Press, Hindawi, and MDPI, were systematically interrogated to guarantee a comprehensive survey of the existing biomedical literature. The investigation was specifically focused on four primary domains: Information extraction in biomedicine, Text mining for biomedical data,

Biomedical ontology-based information retrieval, and Extractive summarization in the biomedical field. The refinement of inclusion criteria involved specifying descriptors like "Data extraction", "Text-mining" "Biomedical ontology", "healthcare", "Extractive Summarization", "Single Document" , "Multiple Document", "Clinical Reports", "Text summarization" and "Keyword extraction". Research papers utilizing alternative descriptors were deliberately excluded. From an initial corpus of approximately 500 papers, a meticulous filtration process, guided by descriptor keywords, led to a curated set of 250 papers. Subsequent scrutiny, emphasizing full-text readability and alignment with the research objectives, resulted in the final selection of 170 papers for comprehensive inclusion in this work. Fig. 2.1 shows the search and selection process of research papers.

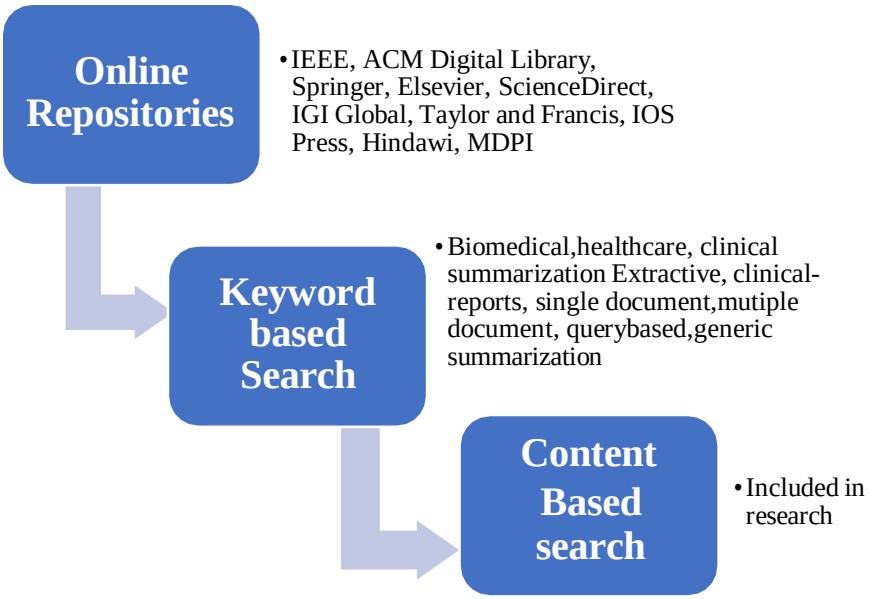


Fig. 2.1 Search and Selection process

2.3. Automatic Summarization

Over half a century has passed since the inception of initial research endeavours in automatic text summarization. During this time, the volume of data has experienced a significant surge, paralleled by an increasing demand for succinct and readily accessible summaries [34]. In the subsequent subsections, various methods employed in automatic summarization are elucidated. A brief summary of all types of automatic text summarization is shown in Fig 2.2.

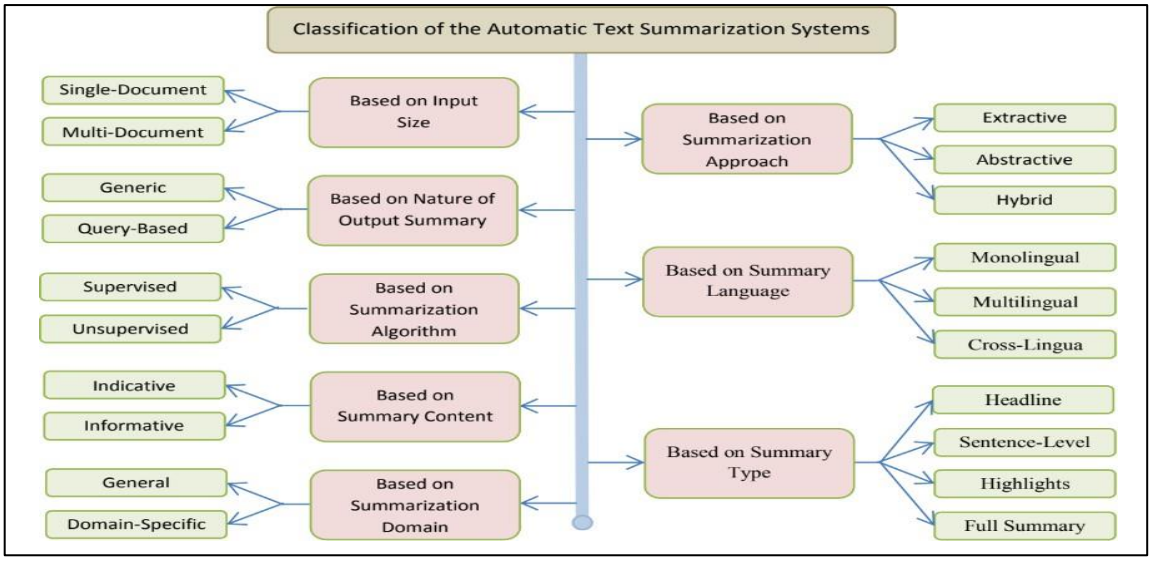


Fig. 2.2. Summary of all approaches for Automatic Text Summarization [34]

2.3.1. Extractive and Abstractive Summarization

Extractive summarization involves "cropping out and stitching together portions of the text to produce a condensed version of a text" [35], [36]. Pioneering work in this area was conducted by [24], who utilized statistical information to calculate a relative measure of significance, initially for individual words and later for sentences. Another notable contribution to automatic text summarization was made by [37], who employed three methods for determining sentence weights. Alternatively, sentences were extracted based on various weighting heuristics [38]. These approaches were employed for many years, but they often gave rise to issues in the overall coherence.

Abstractive summarization involves the generation of a summary by employing novel words to elucidate the primary idea of an article [39]. It also encompasses paraphrasing, generalizing, and introducing new words. The primary challenge encountered in abstractive summarization pertains to the text representation problem [40]. Abstractive summarization, as opposed to extractive summarization, may share more similarities with the human summarization process [41].

2.3.2. Mono and Multi document Summarization

In mono-document summarization, reliance is placed on features such as term frequency, sentence position, and stigma words. Handling the multi-document case is more intricate as challenges may arise in maintaining coherence within the summary. Nevertheless, this case

has gained increasing relevance due to the escalating volume of information and the necessity to summarize multiple documents across various domains, including medical [42], news [43], financial investments [44] and conversations [45].

2.3.3. Generic and Query based Summarization

Generic summarization, encompasses information found in a document. On the other hand, query-based summarization focused on retrieving some information from a document based on a specific information [46].

2.4. Methods and Methodology for Automatic Text Summarization

Automatic Text Summarization (ATS) employs a range of techniques to distil relevant information from extensive textual content. These techniques vary in their approaches and methodologies, catering to different summarization requirements. Various techniques are explained in further subsections.

2.4.1. Frequency-Based Methods

Significant words are identified based on their frequency of occurrence in the text. Techniques namely, Word Probability and Term Frequency-Inverse Document Frequency (TF-IDF) are common frequency-based approach that weighs the importance of terms in relation to their frequency across the entire document set. The approaches are Simple and effective for identifying key terms, but may overlook context [47].

Let's consider a document D consisting of N words $1,2, \quad , N$

The goal is to calculate the probability of each word wi being important in the document.

Term Frequency (TF):

The term frequency of a word wi in the document D is the count of how many times w_i appears in D. It is denoted as $TF(w_i, D)$.

$$TF(w_i, D) = \ Number\ of\ occurrences\ of\ w_i\ in\ D \tag{2.1}$$

Inverse Document Frequency (IDF)

The inverse document frequency of a word wi across a collection of documents is a measure of how unique or important the word is in the entire collection. It is denoted as $IDF(w_i)$.

$$IDF(w_i) = \log \frac{\text{Total number of documents}}{\text{Number of documents containing } w_i} \quad (2.2)$$

Term Frequency-Inverse Document Frequency (TF-IDF):

The TF-IDF score of a word wi in the document D is the product of its term frequency and inverse document frequency.

$$TFIDF(w_i, D) = TF(w_i, D) \times IDF(w_i) \quad (2.3)$$

Normalization:

To ensure that longer documents don't have an advantage, the TF-IDF score can be normalized. One common normalization is dividing the TF-IDF score by the Euclidean norm of the vector representing the document.

$$\text{Normalized TF} - \text{IDF}(w_i, D) = \frac{TFIDF(w_i, D)}{\sqrt{\sum_{j=1}^{n} (TFIDF(w_i, D))^2}} \quad (2.4)$$

Word Probability:

The word probability $P(w_i)$ is calculated by normalizing the TF-IDF score across all words in the document.

$$P(w_i)) = \frac{\text{Normalized TF}-\text{IDF}(w_i, D)}{\sum_{j=1}^{n} (\text{Normalized TF}-\text{IDF}(w_i, D))^2} \quad (2.5)$$

This probability represents the likelihood of each word being important in the given document based on its frequency and uniqueness across the document collection. The higher the probability, the more significant the word is considered in the context of the document.

2.4.2. Sentence Scoring Algorithms

Sentence scoring algorithms evaluate the importance of each sentence based on specific criteria. LexRank and TextRank are graph-based algorithms that evaluates sentences based on their relationships within the document [48], [49]. It is an effective approach for extractive summarization, leveraging sentence-level features.

2.4.2.1. LexRank

It is a graph-based algorithm in which sentences are nodes and edges are the similarity between sentences. The algorithm then computes a centrality score for each sentence, and

the sentences with the highest centrality scores are selected for the summary. Various steps are depicted below:

Given a set of sentences, $S = \{S1, S2, \ldots, Sn\}$

The similarity matrix M is calculated based on the cosine similarity between sentences.

$$M_{ij} = cosine_{similarity(Si,Sj)} \quad (2.6)$$

The transition probability matrix T is created by normalizing the similarity matrix.

$$T_{ij} = \frac{M_{ij}}{\sum_k M_{ik}} \quad (2.7)$$

The centrality score L_i for each sentence S_i is computed using the power iteration method.

$$L_i = (1 - d) + d.\sum_j T_{ij}\, L_j \quad (2.8)$$

Where d is a damping factor (usually set to 0.15).

Then, sentences are selected based on their centrality scores, with higher scores indicating greater importance.

2.4.2.2. Text-Rank

It is a variant of LexRank and was originally designed for keyword extraction. It extends the idea to sentence extraction for summarization. Similar to LexRank, TextRank treats sentences as nodes in a graph and determines importance through graph-based centrality scores. Various steps are as depicted.

TextRank computes a similarity matrix M based on the cosine similarity between sentences.

$$M_{ij} = cosine_{similarity(Si,Sj)} \quad (2.9)$$

Then, the importance of a sentence is determined by summing the similarity scores of sentences connected to it.

$$Importance(Si) = \sum_j \frac{M_{ij}}{\sum_k M_{jk}} \quad (2.10)$$

Finally, Sentences are selected based on their importance scores.

2.4.2.3. Latent Semantic Analysis (LSA)

LSA uncovers the underlying structure in a document by analysing relationships between terms and sentences. It utilizes singular value decomposition to identify latent semantic structures and relationships. LSA captures semantic information and relationships that may not be apparent through traditional methods [50]. The process of LSA is explained below.

Given a term-document matrix A of dimensions,$m * n$,

Where, m is the number of terms and n is the number of documents, the entry A_{ij} represents as;

$$A = \begin{matrix} a_{11} & a_{12} & a_{1n} \\ a_{21} & a_{22} & a_{2n} \\ a_{m1} & a_{m2} & a_{mm} \end{matrix} \quad (2.11)$$

Then, **Singular Value Decomposition (SVD),** decomposes the term-document matrix A into three matrices:U, Σ, and V^T

$$A = U\sum V^T \quad (2.12)$$

Where,

- U is an $m \times m$ orthogonal matrix representing the relationship between terms.
- Σ is an $m \times n$ diagonal matrix containing the singular values.
- V^T is an $n \times n$ orthogonal matrix representing the relationship between documents.

After SVD, Dimensionality reduction is performed. In this step, LSA retains the k most significant singular values and corresponding columns of U and V^T to obtain reduced matrices U_k, $\sum_k$, and V_K^T. Therefore, the term matrix is represented as

$$A = U_k\sum_k V_K^T \quad (2.13)$$

Where, k is the desired reduced dimensionality.

Further, the matrix $U_k\sum_k$ represents the term-concept matrix, capturing the relationships between terms and underlying concepts.

$$T = U_k\sum_k \quad (2.14)$$

And the matrix V_K^T represents the document-concept matrix, capturing the relationships between documents and underlying concepts.

$$D = V_K^T \tag{2.15}$$

Further, to compute the semantic similarity, each term and document is represented as a vector in the reduced-dimensional space.

$$Term\ Vector\ t_i = row_i(T) \tag{2.16}$$

$$Document\ Vector\ d_j = row_j(D) \tag{2.17}$$

Finally, Semantic similarity between terms and documents can be measured using cosine similarity between their vectors.

$$Cosine\ Similarity(t_i, d_j) = \frac{t_i . d_j}{||t_i||.||d_j||} \tag{2.18}$$

2.4.3. Machine Learning Models

Machine learning models, specifically, supervised models, can be trained to identify important sentences for summarization. Support Vector Machines (SVM) [51], Decision Trees [52], and Random Forests [53] are used for sentence classification. These algorithms are customizable and adaptable to specific datasets, enabling personalized summarization.

2.4.3.1. Decision Tree

They are used for both classification and regression tasks. They make decisions by recursively partitioning the input space based on feature values. The various methods used to construct decision tree are explained below.

a. **Entropy:**

Decision Trees use entropy as a measure of impurity in a dataset. The entropy (H) is calculated as:

$$H(s) = -(p^+ \log_2 p^+) - (p^- \log_2 p^-) \tag{2.19}$$

Where, p^+ and p^- are the probabilities of positive and negative classes, respectively.

b. **Information Gain:**

Information Gain (*IG*) is used to determine the effectiveness of a feature in reducing entropy. For a dataset *S* and a feature *A*:

$$IG(S, A) = H(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} H(S_v) \quad (2.20)$$

Where, *Sv* is the subset of *S* for which feature *A* takes value *v*, and values(*A*) are the possible values of feature *A*.

c. **Gini Impurity:**

Another impurity measure used in decision trees is Gini Impurity (*G*). For a dataset *S*:

$$G(S) = 1 - \sum_{c \in classes} (P_c)^2 \quad (2.21)$$

Where (P_c) is the proportion of instances of class *c* in *S*

d. **CART Algorithm for Binary Classification:**

The CART algorithm uses Gini Impurity to split the dataset *S* into two subsets S_{left} and S_{right} based on a feature *A* and a threshold *t*:

$$G(S, A, t) = \frac{S_{left}}{S} G(S_{left}) + \frac{S_{right}}{S} G(S_{right}) \quad (2.22)$$

The algorithm chooses the split that minimizes *G*(*S*, *A*, *t*).

e. **Regression Decision Tree:**

For regression tasks, the decision tree minimizes the Mean Squared Error (MSE) as the impurity measure. Given dataset *S* and target values y_i.

$$MSE(S) = \frac{1}{|S|} \sum_{i=1}^{|S|} (y_i - \bar{y}_s)^2 \quad (2.23)$$

Where, $\bar{y}_s$ is the mean target value of *S*

It recursively split the dataset based on features and thresholds to create a tree structure that can make predictions for unseen instances.

2.4.3.2. Random Forests

It is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and control overfitting. Each tree in the ensemble is built on a subset of

the training data, and the final prediction is obtained through a voting or averaging process [54], [55], [56]. Various steps of random forest are explained.

A. **Bootstrapped Dataset:**

Random Forest constructs multiple decision trees, and each tree is trained on a bootstrapped subset of the original training data. It involves random sampling with replacement, creating a new dataset S_i, for each tree i.

$$S_i = BootstrapSample(S) \tag{2.24}$$

B. **Feature Randomization:**

At each node, a random subset of features is used for splitting. If the original dataset has m features, a subset m_{rand} is chosen randomly.

$$m_{rand} \leq m \tag{2.25}$$

C. **Decision Tree Training:**

For each bootstrapped dataset S_i, a decision tree T_i is trained using feature randomization. The training involves recursively splitting nodes based on the selected features until a stopping criterion is met.

$$T_i = TrainDecisionTree(S_i) \tag{2.26}$$

D. **Voting (Classification) or Averaging (Regression):**

For classification, the final output is determined through a majority vote. For regression, the final output is the average of the predictions made by individual trees.

$$FinalPrediction = \frac{1}{N}\sum_{i=1}^{Ntrees} prediction(T_i) \tag{2.27}$$

E. **Out-of-Bag (OOB) Error Estimation:**

The performance of the Random Forest can be estimated using out-of-bag samples, which are instances not included in the bootstrapped dataset for each tree. The OOB error is computed by evaluating the predictions on these out-of-bag samples.

$$OOB\ Error = \frac{1}{N}\sum_{i=1}^{N} L(y_i, AveragePrediction(\{T_j | x_i \in S_j\})) \tag{2.28}$$

where N is the number of instances, y_i is the true label of instance i, and L is the loss function.

2.4.4. Deep Learning Architectures

Deep learning techniques such as Recurrent Neural Networks and transformers, excel at capturing complex relationships and patterns in text. Sequence-to-sequence models are equipped with attention mechanisms that are effective for abstractive summarization [57], [58], [59], [60], [61], [62].

2.4.5. Cluster Analysis

Cluster analysis groups similar sentences or phrases together, aiding in the identification of key themes. K-means clustering and hierarchical clustering are commonly used for grouping related content. These techniques are suitable for multi-document summarization, revealing distinct themes across the document set [8], [63], [64], [65].

2.4.6. Methodology

Prior to the development of an automatic summarization method, it is imperative to establish a formal definition outlining the nature of a summary and the expected output from the proposed method. The task of automatic text summarization is composed of the following components as depicted in Fig. 2.3.

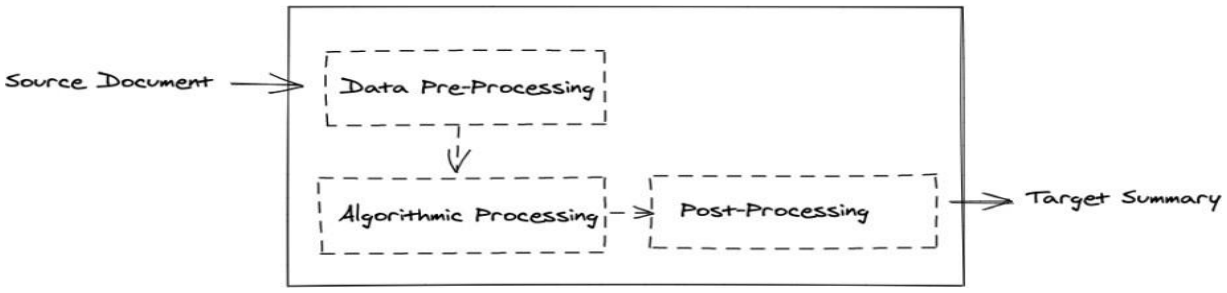


Fig 2.3. Overview of Automatic Text Summarization

Source Document- Sources of data can encompass various formats, including text, images, audio, and video. Historically, summarization methods were predominantly developed for text, with a limited focus on audio and video. However, contemporary advancements have led to the emergence of numerous methods for summarizing audio and video data. The field has evolved to accommodate the diverse nature of biomedical data, which exists in multiple forms. Figure 2.4. illustrates the various sources from which biomedical data can be generated, highlighting the expanding range of data types and modalities in this domain [11].

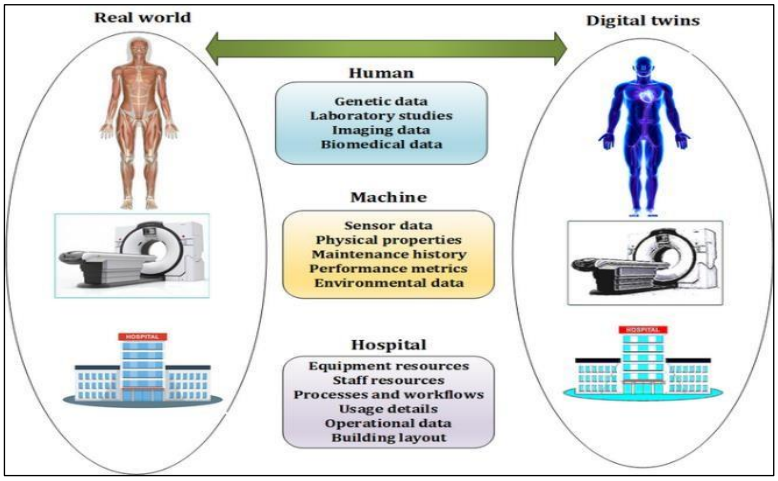


Fig 2.4. Different Sources for data generation in the Biomedical field [11]

Data Pre-Processing - In the data pre-processing phase, the raw source document undergoes cleaning and transformation. Various techniques are employed, such as:

- Noise removal: Eliminating data that does not contribute valuable information to the document or summary.
- Sentence tokenization: Dividing the data into a set of sentences.
- Removal of punctuation marks: Clearing the text of unnecessary punctuation.
- Word tokenization: Breaking down sentences into a set of words or tokens.
- Removal of stop words: Eliminating frequently occurring words like (a), (an), (the), etc.
- Word stemming: Removing suffixes and prefixes from words.
- Lemmatization: Transforming words to their base or dictionary form (e.g., converting [playing, played, plays] to [play]).
- Part-of-speech tagging (PoS tagging): Assigning grammatical categories to words or tokens.

These pre-processing techniques collectively enhance the quality and structure of the data for subsequent analysis or summarization without introducing unnecessary elements.

Algorithmic processing: The algorithmic processing phase, extensively covered in the literature review concerning various text summarization methods, involves the application of algorithms to generate a summary from the pre-processed input document. This approach can be statistical, graph-based, topic-based, machine learning-based and deep learning methods. The algorithmic processing phase encompasses both supervised and unsupervised approaches. In the current research, Extractive summarization is generally preferred over

abstractive ones due to their demonstrated superior performance and relatively simpler implementation and better in the biomedical text as no golden summaries are given for the training purpose.

Post-processing phase: The post-processing phase involves making necessary data transformations to refine the output from the algorithm processing phase into the target summary. This step is optional in some approaches; as certain summaries can be generated without the application of additional NLP techniques. Text summarization can be broadly classified into two main forms based on Information Types (Informative & Indicative):

a) **Informative Summary**: Focuses on conveying comprehensive details and facts from the source document, providing a detailed overview of the content.
b) **Indicative Summary**: Highlights key points and essential information, offering a more concise representation of the source content.

These classifications offer a framework for understanding the varied approaches and goals within the domain of text summarization.

Target Summary- It shows the output generated by the proposed algorithm based on input, algorithm and post-processing steps. The validation of a summary is not absolute; rather, it depends on the reader's interpretation of a document, shaped by their understanding. A valid summary could consist of a few crucial keywords, while another equally valid summary might encapsulate a paragraph.

In conclusion, the techniques for automatic text summarization are diverse, each offering specific advantages depending on the nature of the text and the summarization goals. The evolving landscape of ATS continues to witness innovation and integration of these techniques to enhance the accuracy and efficiency of general summarization systems.

2.5. Related work on Automatic Text Summarization

Automatic Text Summarization (ATS) has a myriad of challenges to the research community, which includes **Identification of Informative Segments** within the input text that should be incorporated into the generated summary [47], **Summarization of Lengthy Single Documents** without losing key information is a complex task, **Multi-Document Summarization** [66], requiring the synthesis of information from diverse sources into a coherent summary, Extractive summarization, **Abstractive Summary Generation** akin to those produced by humans [67] remains an ongoing challenge.

ATS is particularly employed in conjunction with information retrieval to augment the capabilities of search engines. Tuarob et al. introduced a search engine dedicated to locating algorithms and pseudo-codes. This approach involves constructing a dataset by extracting algorithms from scientific papers, followed by the utilization of ATS [68]. In a study by Yulianti et al., text summarization is utilized to extract answers for non-factoid queries. It applies to various text genres [69]. Its integration with speech recognition, medical documents, legal documents, and more has been explored by researchers such as Vodolazova et al.. Each ATS system is tailored to support one or more text genres as inputs, leading to its utilization in diverse applications such as news summarization, email summarization, and domain-specific summarization (e.g., legal or biomedical document summarization) [70]. The subsequent sections delve into the summarization of the biomedical domain.

2.5.1. Datasets for Automatic Text Summarization

Numerous datasets have been established specifically tailored for the advancement of automatic summarization. A significant milestone in this context is the Document Understanding Conferences (DUC)[1]. DUC constituted an international competition wherein the research community introduced innovative methodologies to address challenges in Natural Language Processing, particularly in the evaluation of automatic summaries. These methodologies consider reference summaries crafted by human authors. This competition spanned the years from 2001 to 2007, during which various research groups utilized distinct corpora each year.

DUC01 consisted of 147 document-summary pairs and DUC02 comprised 567 document-summary pairs. Moving forward, DUC 2003 encompassing 500 news articles sourced from the New York Times and Associated Press Wire services. Each summary in this context had four corresponding human reference summaries, resulting in a total of 624 document-summary pairs[71] . DUC 2006 introduced 50 topics, each composed of 25 relevant documents from the AQUAINT corpus[2], primarily derived from various sources. Finally, DUC 2007 presented a dataset addressing two tasks. Each of the 45 topics concerning news included 25 documents. The main task centred on question-answering-based summarization, while the second task focused on generating short summaries from multiple documents.

[1] www.kaggle.com
[2] www.paperwithcode.com

Another renowned dataset originates from the Text Retrieval Conference (TREC [72]), specifically designed for question classification. TREC is available in two versions: TREC-6, featuring six classes, and TREC-50, which incorporates fifty classes. Both versions consist of 5,452 training examples and 500 test examples each, contributing to a comprehensive set of instances for evaluating question classification models.

The Gigaword dataset comprises approximately 9.5 million news articles and a staggering four billion words. These articles are sourced from seven reputable outlets featuring over 1.8 million articles. Notably, the scientists at the library contributed over 650,000 article summaries [73]. According to [74], the average document length is 530 words, while abstracts average 38 words. Additionally, the Giga word dataset served as the inspiration for GIGA-CM, derived from the English Giga word dataset encompasses 6,626,842 documents and a voluminous 2,854 million words [75].

The CNN/DailyMail News dataset is constructed from online news articles. On average, the articles consist of 781 tokens, while the abstracts contain 56 tokens [62].

Several other notable datasets include BillSum [76], XSum [77], NEWSROOM [78], and WikiSum [79].

NEWSROOM, another dataset, includes 1.3 million articles and human-written summaries generated by authors and editors from 38 major news publications in the duration of 1998 to 2017. WikiSum, derived from Wikipedia, is designed for article generation. These datasets contribute to the diversity and comprehensiveness of resources available for the development and evaluation of text summarization models.

Throughout the years, researchers have shown keen interest in domains with characteristics differing from those of the general domain, particularly in the scientific realm. Scientific texts, exemplified by their length and inclusion of specialized terms and keywords, present distinct challenges compared to news articles. Noteworthy datasets focusing on the medical domain include Ziff–Davis and PubMed 200k RTC [77].

In the case of PubMed 200k, it serves as a dataset designed for classifying sentences within medical abstracts. Each sentence is labelled based on its role within the abstract. This dataset encompasses a total of 195,654 abstracts, providing valuable material for training and evaluating models tailored to the unique characteristics of scientific and medical text.

2.5.2. Related Work on Extractive Summarization

For extractive summarization, document is pre-processed and assigns a score to each sentence in a document based on some weighting factor. The sentence with a score above a certain threshold value is selected for generating an extractive summary. Extractive summarization has been applied to several domains such as automatic highlighting of text [1], web articles[80], multi document summarization [81] - [83] and many more. Various techniques include statistical-based approaches [70], genetic algorithms [43], graph-based [84], [85], neural networks [3], [36], [86], optimization-based [87], [88], conditional random fields [89], [90], semantic similarity-based [91], fuzzy-logic based [92] and centroid-based techniques [93]. A query-oriented approach for multi-document summarization has been proposed [94], which learns hierarchical concepts using Deep Restricted Boltzmann Machines. Other semantic similarity approaches such as latent semantic analysis [95], [96], [97] are widely used. Latent Semantic Analysis (LSA) has been previously used to summarize text [95], [98], [99], [100]. To compute the importance of sentences using textual similarity of text, graph-based approaches TextRank and LexRank have been used. With supervised techniques, unsupervised methods such as fuzzy logic [101] - [104] and k-means clustering [105] have also been used. Multi-document summarization using fuzzy logic is proposed by D. Patel et.al. Fuzzy rules were created to generate a summary of documents and cosine similarity is used to remove redundancy [101]. Another approach based on fuzzy logic with evolutionary algorithm and cellular learning automata was proposed by R. Abassi-ghaletaki et.al [102] to produce a summary. The approach was evaluated and results depict that evolutionary algorithms combined with fuzzy logic method outperform other techniques [103]. Another method that used fuzzy metrics was proposed by F.B. Goularte et.al. In fuzzy analysis, 27 rules were produced and relevance was computed. The results were improved in terms of the informativeness of the generated summary [102]. E.V. Valdes et.al. generated a semantic graph between the concepts, which were merged and a concept clustering algorithm was used to identify the relevant topics in a combination of fuzzy aggregation functions [106]. J.M. Sanchez-Gomez et.al. performs a comparative study of disparate criterions applicable to multi-document summarization. MOABC algorithm has been used as an objective function [107]. A novel method called the Karci Summarization approach was introduced, which quantifies the degree to which each sentence captures the essence of the entire text using numerical values. To prepare the data, a tool named KUSH was created, facilitating the translation of sentence relations into graphical representations. The

performance was evaluated through ROUGE [108]. In successive study [109], two concepts, textual graph and maximum independent sets, were employed. The maximum independent sets were identified from the textual graph and subsequently eliminated. The remaining nodes, representing the main concepts, were then incorporated into the document summary. The experiment utilized datasets from DUC 2002 and DUC 2004, achieving a Rouge score of 0.38072 for 100-word summaries Cat Swarm Optimization approach was given by R. Rautray et.al [110]. The similarity between sentences was computed and on selected sentences, CSO algorithm was applied to DUC data. The performance was evaluated using several metrics. Gupta et.al. made significant contributions related to statistical-based methods for extractive summarization such as favourable positioning or frequency. Additionally, detailed steps in a statistical-based extractive summarizer's sentence scoring process is specified which includes the selection and calculation of statistical and/or linguistic features, the assignment of weights to these features, and the final scoring of sentences based on a feature-weight equation [111]. Further, Gupta et.al. proposed a statistical-based extractive summarization method to automatically generate summaries from a given set of documents which is based on the selection and calculation of statistical and linguistic features [112].

A query-oriented summarization approach based to Ensemble-Noise-Auto-Encoders is proposed. Similarly, SummCoder summarizer is proposed based on deep auto-encoders is developed for single document summarization by Joshi et.al [113]. The above approach differs by an approach proposed by as the later one used sentence embedding models. Recent research has been performed towards word-embedding and achieves significant results compared to other approaches. Mohd et.al. combines word-embedding with k-means clustering for extractive summarization. The author used word2vec model and statistical features to select relevant sentences for summarization [114]. M.A. Mohammed et.al proposed several works on natural language processing techniques such as convolution neural network, adaptive intelligent learning approaches, agent-based multi natural language and other supervised learning methods for image, email classification [115]. Word embedding models are based on lexical similarity and semantic measures are not focused. Therefore, in this research work, semantic similarity has been focused in place of lexical similarity. The concepts are identified based on semantic measures using Latent Semantic Analysis. Extractive summarization is also done in microblog and tweet summarization. Social media platforms such as Facebook and Twitter encompass an immense volume of

messages. This work highlights the extensive and dynamic nature of communication on these platforms, where millions of messages are exchanged regularly. The prevalence of user-generated content underscores the significance of these platforms as vibrant spaces for information dissemination, interaction, and engagement.

Navigating the challenges of extractive summarization within the biomedical domain represents a crucial focus in this research endeavour. In this context, where precise information extraction is paramount, the challenge of redundancy and repetition in biomedical texts is addressed by developing algorithms that identify and eliminate duplicated content efficiently. The coherence of extracted sentences is enhanced through specialized linguistic models that consider the unique structure and terminology inherent in biomedical literature. To mitigate the potential loss of critical information, the research explores techniques for recognizing interconnected concepts within and across sentences, ensuring that the extracted summary retains the essential relationships and context crucial for biomedical understanding.

2.5.3. Related work on Summarization in Biomedical domain

Given the abundance of electronic health records, there arises a need to condense these records to assist clinicians and researchers in efficiently accessing comprehensive information within the biomedical field. Addressing this need, M. Afzal et al. devised a BioMed summarizer utilizing deep neural networks. This summarizer offers PICO-based intelligent summarization of biomedical articles, enhancing content accessibility and comprehension. For this purpose, Keras tokenizer was used which was integrated with a bidirectional long-short term memory classification model. 95.41% accuracy was achieved in article recognition and 93% accuracy in classification of text into five categories: aim, population, intervention, result, and outcome [116]. Further, M.S. Azadani et.al. developed a summarizer that extracts concepts and their correlation based on the Unified Medical Language System and frequent itemset mining technique, FPGrowth, to generate a graph between concepts as graph nodes and similarity between concepts as edges. The approach was evaluated on 400 articles from BioMed Central using the ROUGE evaluation metric [117]. The detailed survey of text summarization in the biomedical domain can be studied and referred to from R.Mishra et.al. [118]. The work done in literature faces two challenges: i) only lexical similarity is focused based on word-embedding approaches ii) domain-dependent knowledge is incorporated to generate summary of biomedical articles. C.

Mallick et al. introduced an innovative approach to address the large volume issue. The abstracts were utilized as base summaries. A multi-objective evolutionary algorithm is then used to generate the summary. Each sentence was changed into a concept vector of medical terms using the Unified Modeling Language System tool. These concept vectors capture essential information, facilitating the analysis of semantic similarity among sentences clustering coefficient and sparsity index were used as fitness function. After algorithm convergence, the best solution from the final population yields the ensemble summary. The approach is evaluated on articles from the PubMed MEDLINE database [119]. In another work, C. Gulden et al. focused on the extractive summarization of clinical trial descriptions to enhance efficiency to condense lengthy and detailed clinical trial descriptions into concise yet meaning-preserving synopses. A unique dataset is curated from detailed descriptions of trials registered on clinicaltrials.gov. Multiple text summarization algorithms were applied to these descriptions, and standard ROUGE metrics were computed using the brief summaries as references. To gauge the relationship between metrics, four reviewers were assessed through a Likert scale questionnaire. Results indicate that the dataset, initially consisting of 277,228 trials, was filtered down to 101,016 records. The generated summaries were 25% the length of detailed descriptions. The TextRank algorithm demonstrated the best performance with ROUGE-1/2/L F1 scores which aligned with human reviewers' assessments. The study concludes that the ROUGE-L F1 score serves as a valuable automated metric for rating the general quality of generated clinical trial summaries [120].

L. Li. et.al. addresses the challenges of document summarization concerning diversity, coverage, and balance. The authors focused on extract-based summarization and emphasized three critical requirements: diversity, aiming to minimize redundancy; sufficient coverage, to retain the document's main information; and balance, ensuring equal importance to different aspects of the document in the summary. The proposed approach explored the graph structure of output variables and utilizes structural Support Vector Machines to solve the resulting optimization problem [121]. Further, Y. Ouyang et.al. investigated the application of regression models in query-focused summarization. Support Vector Regression was applied to compute sentence importance based on predefined features. To train the regression models, "pseudo" training data was constructed. The proposed approaches are evaluated using DUC datasets, focusing on efficiency and robustness [94]. M. Moradi et.al. introduced a novel approach called CIBS. The goal was to extract essential information from lengthy documents, the challenge was to create a summary covering main

topics from multiple related texts, reducing redundant information. CIBS operates by extracting biomedical concepts and utilizing an itemset mining algorithm to identify primary topics. Subsequently, a clustering algorithm forms clusters and summarizer then selects sentences from various clusters to create a comprehensive summary encompassing a broad range of topics present in the input text. The approach was evaluated using the ROUGE toolkit to compare CIBS against four summarizers. Results demonstrate that proposed method enhances the performance [122]. In their consecutive study, M. Moradi et.al. proposed a novel method by combining itemset mining and domain knowledge. The objective was to enhance access to information from vast scientific literature. The proposed summarization method constructs a concept-based model by mapping the document to biomedical concepts using the UMLS. Subsequently, essential subtopics were identified using itemset mining, and the summarization model is created [123]. In another work M Moradi et.al. proposed a novel summarization method that leverages contextualized embeddings generated by various versions of BERT. These embeddings were combined with a clustering method. The approach was evaluated using ROUGE toolkit and demonstrates that the proposed summarizer achieves state-of-the-art results [124]. C. Yongkiatpanich et.al. introduced a novel approach that combined graph building rules with the Word Mover's Distance, a distance function between text documents. To prioritize core sentences Google's PageRank algorithm was used and is evaluated against other text summarization software using a corpus of 400 biological review papers randomly sampled from PubMed Central. The results demonstrated that the proposed method surpasses baseline comparators based on ROUGE scores. Y. Du,et.al. proposed a novel model named BioBERTSum that employed a domain-aware bidirectional language model, pre-trained on extensive biomedical corpora, as the encoder.. Sentence position embedding mechanism was used to enable to capture position information and incorporate the structural features of the document [14]. E. K. Lee et.al. developed an interactive content extraction, recognition, and construction tool for clinical and biomedical text, named CERC. A novel sentence-ranking framework was proposed based on random forest. The approach attains an 87.5% accuracy and outperforms methods based on single indicators in terms of ROUGE-1/2/SU4 scores [125].

Y.P. Chen et.al. proposed an approach that involves BERT-based structure with a two-stage training method. The model is trained on 258,050 discharge diagnoses and experienced doctors provide labelled extractive summaries. The proposed model, AlphaBERT, is fine-tuned using summary labels and addresses character-level issues by averaging probabilities

for entire words. Results indicate that AlphaBERT outperforms other models [61]. M. Moradi et. al. proposed a graph-based summarization approach that leverages domain-specific word embeddings and graph ranking techniques. The approach is evaluated using ROUGE metrics [126].

E. Davoodijam et.al. proposed a novel approach MultiGBS that incorporated features such as word similarity, semantic similarity, and co-reference similarity. The approach was evaluated based on ROUGE and BERTScore metrics [127]. D. P. Purbawa et.al. proposed an approach that utilizes cosine similarity along with MMR and TextRank to generate document summaries [128].

Rai et.al. proposed a query-specific framework to generate focused summaries from biomedical journal articles, particularly during public health emergencies like the COVID-19 pandemic. The evaluation of the approach is conducted on the CORD-19 dataset [129]. Similar work was done by P. Chen et.al. A novel query-based text summarization approach was proposed to enable efficient retrieval of concise and relevant information in response to user queries. Further, ontology-based and a keyword-only approach are compared, with the ontology-based method demonstrating superior performance [130]. N. Elhadad et.al. proposed the summarization system that utilizes a unified user model, leveraging the structure and content regularities. The results demonstrate that the generated summaries incorporate both machine-generated text and extracted information from multiple input documents [131].

M. Fiszman et.al. proposed a methodology to extract drug information from Medline citations and present it in a user-friendly format. The evaluation involves citations discussing ten drugs, demonstrating that automatic summarization can complement curated drug databases, thereby enhancing the support for quality patient care [132]. In their successive study, addresses the challenge of information retrieval for physicians in the context of increasing electronic biomedical resources. The approach was evaluated on various metrics and results depict that MAP gain was 0.17 [133].

K. R. McKeown et.al. focused on personalizing search results and summarization in the PERSIVAL medical digital library. The approach was based on re-ranking search results that highlight information relevant to the patient under the physician's care [131]. D. Molla et.al. focused on addressing NLP-related challenges in Evidence-Based Medicine. The corpus

creation process incorporates automated text extraction, manual annotation, and crowdsourcing to identify reference IDs [134].

L. Plaza et.al. enhance the performance of a concept-based summarization system for biomedical documents. The approach utilizes graphs, incorporating concepts and relations from the UMLS. The MetaMap program is employed to map the text onto concepts in the UMLS Metathesaurus for graph creation. The result demonstrates that integrating a graph-based Word Sense Disambiguation algorithm into the MetaMap output leads to improved quality in the generated summaries [135].

L. Reeve et.al. introduced a novel approach, BioChain, for biomedical text summarization using lexical chaining methods. The approach when evaluated against human summaries, a precision of 0.90 and recall of 0.92 is attained [136]. A. Sarker et.al. proposed a query-focused approach that selects informative sentences from medical documents to assist practitioners in finding relevant information efficiently. The researchers utilized a specialized corpus for EBM summarization, deriving important statistics related to extractive summaries. Their approach outperforms all baseline approaches. The study contributes to the limited research in automatically summarizing medical text and holds promise for enhancing EBM practitioners' efficiency [134] . The comparative analysis of state-of-the-art approaches is presented in Table 2.1. All the sub-domains of the biomedical domain that can be used for summarization are illustrated in Fig. 2.5.

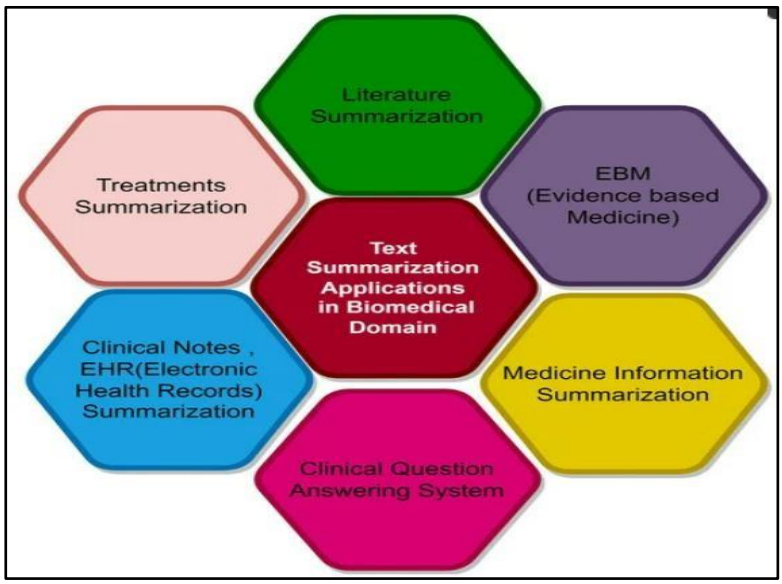


Fig.2.5. Various sub-domains of the Biomedical domain used for summarization [137]

Table 2.1. Comparative analysis of the advanced approaches for biomedical summarization

S. No	Author	Methodology	Algorithms	Datasets	Evaluation Metrics	Results
1.	A. Afzal et.al. [116]	BioMed Summarizer	Deep Neural Network	BIOSSES	Accuracy, Pearson correlation coefficient	95.41% accuracy was achieved in article recognition and 93% accuracy in classification of text into five categories
2.	M.S. Azadani et.al. [117]	Summarizer	FPGrowth	400 articles of Biomed central	ROUGE	Domain-specific knowledge and frequent itemset mining summarizes with more informativeness measurement.
3.	C. Mallick et.al. [119]	Evolutionary based summarization	Multi-objective algorithm, semantic similarity	PubMed MEDLINE	ROUGE-1, ROUGE-2, ROUGE-SU	Approach extracts the medical information from recently published articles.
4.	C. Gulden et.al. [120]	NA	Multiple machine learning algorithms	101,016 records on clinicaltrials.gov	ROUGE	ROUGE-L F1 score is valuable for rating the quality of generated clinical trial summaries
5.	L. Li. et.al.[94]	NA	SVM	DUC2001	F1 and ROUGE	Significant improvements were attained
6.	Y. Ouyang et.al. [71]	Query focussed	Support Vector Regression	DUC	ROUGE	Regression models are better than classification models to estimate the importance of the sentences
7.	M. Moradi et.al. [122]	CIBS	Itemset mining and clustering	UMLS	ROUGE	CIBS enhances the performance.
8.	M. Moradi et.al. [123]	NA	Itemset mining and domain knowledge	UMLS	ROUGE	Approach attains best scores.
9.	M. Moradi et.al. [124]	BioBERT	BERT	UMLS	ROUGE	BioBert enhances the performance as compared to domain specific and domain-independent approaches

10.	C. Yongkiatpanich et.al. [178]	NA	Graph building rules with the Word Mover's Distance	400 articles from PubMed Central	ROUGE	The proposed approach attains best results.
11.	Y. Du et.al. [14]	BioBERTSum	BERT	PubMed	ROUGE-1/2/L	The approach outperforms existing models
12.	K. Lee et.al. [125]	CERC, MINTS	Random Forest	32 full-text CRAFT articles	ROUGE-1/2/SU	The approach attains an 87.5% accuracy and outperforms methods based on single indicators.
13.	Y.P. Chen et.al. [130]	AlphaBERT	BERT	258,050 discharge diagnoses	ROUGE. AUC-ROC	AlphaBERT outperforms other models achieving an AUC-ROC of 0.947
14.	M. Moradi et.al.[126]	Graph-based	Domain specific word embeddings	1.8 million biomedical articles	ROUGE, UWR, UCR	The approach increases the informative content
15.	E. Davoodijam et.al. [127]	MultiGBS	word similarity, semantic similarity, co-reference similarity	450 articles from BioMed Central	ROUGE, BERTScore	Approach attains improved results.
16.	D.P. Purbawa et.al. [128]	NA	MMR, TextRank	Health research ethics protocols Documents	F- score	Improved F-score values were attained.
17.	Rai et.al. [129]	Query-specific framework	Named entity extraction	CORD-19	ROUGE-1/2/L	Generates a uniformly-structured summary.
18.	P. Chen et.al. [138]	Query-based	UMLS ontology	Metathesaurus	Precision, Recall	Ontology knowledge is an effective way than keyword-based information retrieval methods.
19.	N. Elhadad et.al. [131]	PERSIVAL summarizer	User modelling	Medical articles in Persival database	Precision, Recall	Precision of 90% and recall of 65% was attained.

20.	M. Fiszman et.al. [132]	NA	UMLS ontology	Medline citations	Saliency	Useful supplement was proposed.
21.	L. Plaza et.al. [139]	Concept-based	Word Sense Disambiguation , MetaMap	MetaThesaurus	Precision, Recall	Integration of graph-based Word Sense Disambiguation algorithm into the MetaMap output leads to improved quality in the generated summaries.
22.	L. Reeve et.al. [136]	BioChain	Lexical chaining methods	MetaThesaurus	Precision, Recall	The approach when evaluated against human summaries, a precision of 0.90 and recall of 0.92 is attained.
23.	A. Sarker et.al. [134]	Query-focused	Sentence classifier tailored for EBM domain	Real life Clinical queries	ROUGE L	The study contributes to the limited research in automatically summarizing medical text and holds promise for enhancing EBM practitioners' efficiency

2.6. Data Collection

Data collection is a crucial element in medical and life sciences research. The rapid growth in health-related research, especially due to the global efforts to combat the Covid-19 pandemic, has significantly increased the volume of medical articles. Many institutions and researchers worldwide are actively addressing Covid-19 challenges. The statistics highlighted the substantial rise in digital medical data:

- In 2020, submissions to Elsevier's journals increased by 58% from February to May compared to the same period in 2019.
- Health-related articles saw a 92% surge in 2020, with over 100,000 Covid-19 articles published.
- Global healthcare data grew substantially from 2013 to 2020, with a notable increase in data volume in exabytes.

- The US National Cancer Institute received over 4.5 petabytes of data from research institutions in its first year (2016-2017).

Platforms like PubMed and Dimensions now host millions of medical texts from diverse sources, emphasizing the expansion of medical data and its critical role in research and healthcare. Extracting information from these articles is essential to keep pace with medical advancements, aiding researchers and ultimately saving lives.

There are two main data collection methods in statistics: primary and secondary.

Primary data collection involves directly obtaining information from various sources, yielding raw, firsthand data, and enhancing accuracy and reliability. Methods include surveys, interviews, and observations.

Secondary data collection involves gathering data from published sources like scientific journals, government reports, or databases, offering valuable insights. Methods include literature reviews, data mining, and historical data analysis.

The MTSamples transcripts have been utilized for our text summarization research as it provides a diverse and representative corpus of real-world medical transcriptions. This dataset, with its specialized medical terminology, allows to develop and evaluate the algorithms effectively, In this research, data was collected from MTSamples, a platform providing a diverse collection of sample transcription reports across various medical specialties. These reports are regularly updated and contributed by different transcriptionists and users for reference purposes. MTSamples includes 4,996 real summaries of transcripts in 40 domains such as Neurology, Allergy, ENT, Urology, Autopsy, Bariatrics, Cardiology, Cosmetic, Diet and Nutrition, Discharge summaries, and General medicine as shown in Fig 2.6. This dataset is crucial for text summarization research, helping to overcome challenges in discovering accurate medical information and providing valuable resources for medical professionals. The MTSamples corpus was considered for its exceptional value in research on automatic text summarization within the biomedical domain. One key reason is the diversity of medical specialties it covers, with summaries in 40 fields with its inclusion of 4,996 real summaries of patient transcripts which ensures that the dataset represents a wide array of medical knowledge, making it suitable for developing summarization techniques that can be applied across various medical contexts. The use of authentic, real-world data enhances the relevance and practical applicability of the research findings, ensuring that the developed summarization models are grounded in real medical scenarios. Moreover,

MTSamples is regularly updated with new reports, reflecting the latest trends and developments in medical transcription. This continuous enrichment ensures that the data remains current and relevant, which is crucial for developing robust summarization models capable of handling the evolving nature of medical terminology and practices. The dynamic nature of the dataset allows researchers to stay abreast of contemporary medical discourse.

In addition to its practical applications, the MTSamples corpus serves as an invaluable reference for both aspiring and practicing medical transcriptionists. Furthermore, medical data can be challenging to obtain and work with due to privacy concerns and the inherent complexity of medical information. The MTSamples corpus offers a structured and accessible dataset that helps overcome these challenges, facilitating research and development in the field of automatic text summarization. Its structured format and comprehensive content make it easier for researchers to analyze and process the data effectively.

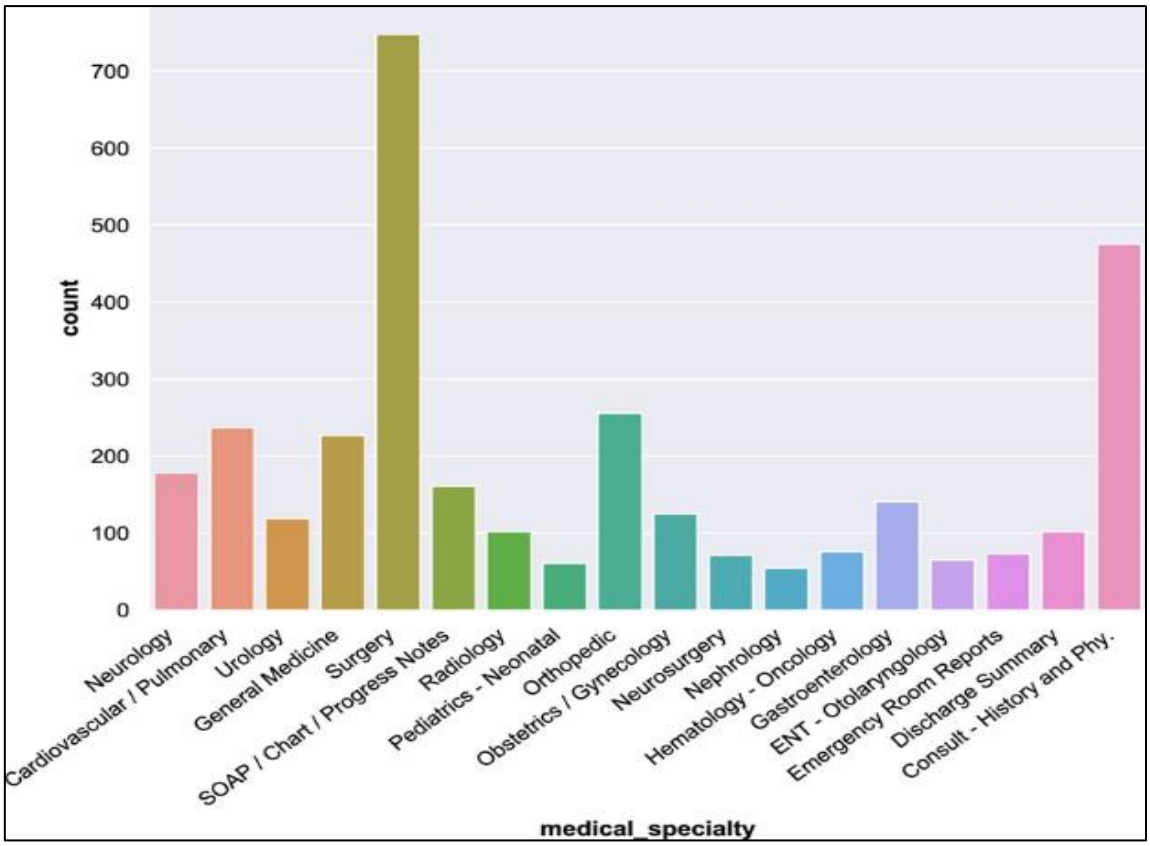
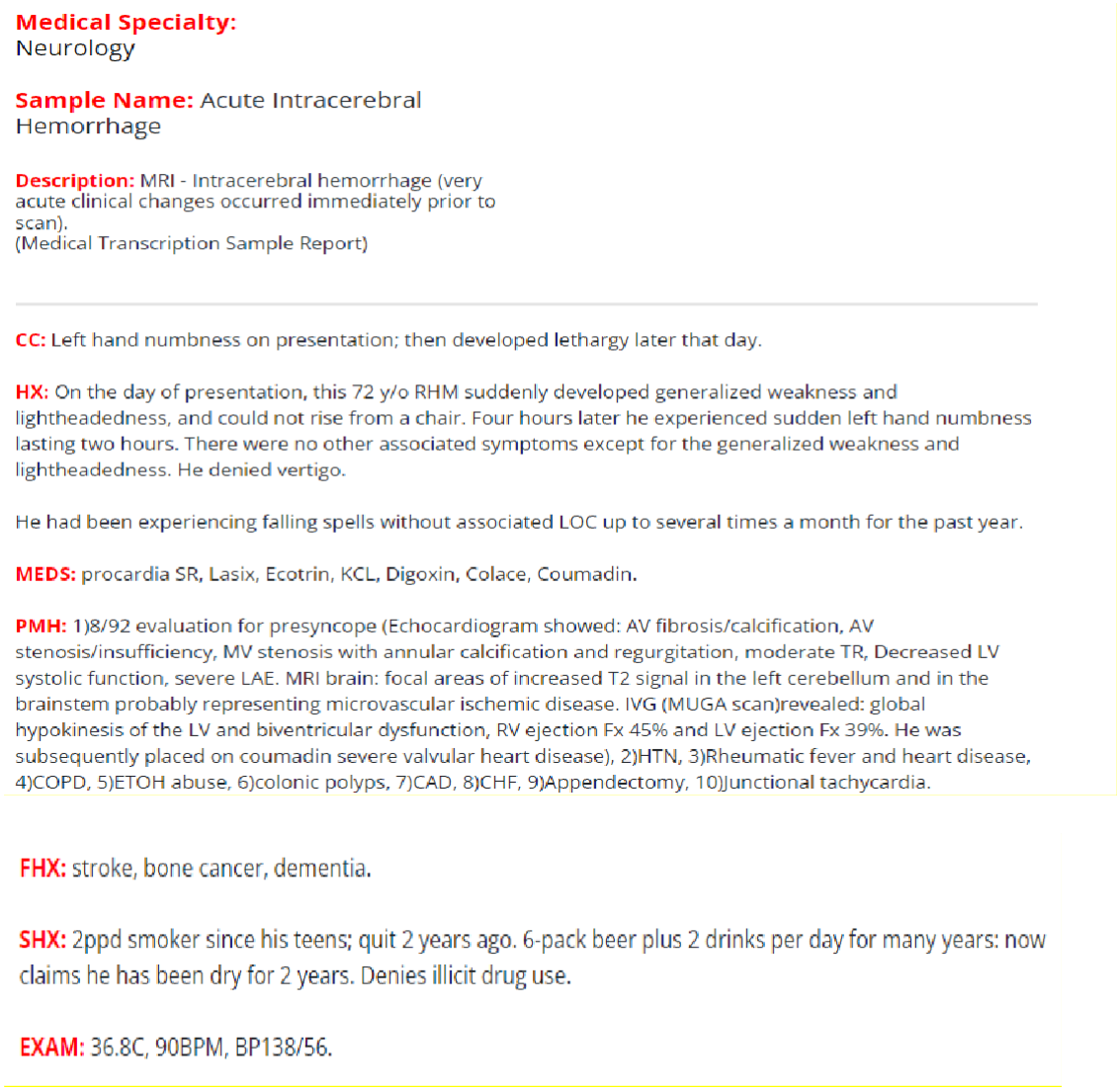


Fig 2.6. Transcripts available on MTSamples

It is aimed to identify and analyze significant features present in biomedical transcripts namely: description, medical specialty, sample name, transcription, and keywords, that can

be leveraged to generate effective summaries. Table 2.2. illustrates samples of transcripts from major domains, including Neurology, General Medicine, Gynaecology, Dental, and Cardiovascular, containing 224, 260, 154, 28, and 372 transcripts, respectively. These samples collectively form a corpus named MT Corpus, consisting of a total of 1,040 transcripts. Instead of using the entire corpus, a subset of transcripts from major domains is selected to form a more manageable and targeted dataset named MT Corpus. Unlike previous state-of-the-art techniques that focused on PubMed and BioMed articles, this research introduces a novel approach by centering on medical transcripts. The creation of the MTCorpus aims to delve into the realm of biomedical transcripts, streamlining the process of reading, comprehending, and providing diagnoses to patients. Sample transcript is depicted in Fig. 2.7. for the Neurology specialty.



Medical Specialty:
Neurology

Sample Name: Acute Intracerebral Hemorrhage

Description: MRI - Intracerebral hemorrhage (very acute clinical changes occurred immediately prior to scan).
(Medical Transcription Sample Report)

CC: Left hand numbness on presentation; then developed lethargy later that day.

HX: On the day of presentation, this 72 y/o RHM suddenly developed generalized weakness and lightheadedness, and could not rise from a chair. Four hours later he experienced sudden left hand numbness lasting two hours. There were no other associated symptoms except for the generalized weakness and lightheadedness. He denied vertigo.

He had been experiencing falling spells without associated LOC up to several times a month for the past year.

MEDS: procardia SR, Lasix, Ecotrin, KCL, Digoxin, Colace, Coumadin.

PMH: 1)8/92 evaluation for presyncope (Echocardiogram showed: AV fibrosis/calcification, AV stenosis/insufficiency, MV stenosis with annular calcification and regurgitation, moderate TR, Decreased LV systolic function, severe LAE. MRI brain: focal areas of increased T2 signal in the left cerebellum and in the brainstem probably representing microvascular ischemic disease. IVG (MUGA scan)revealed: global hypokinesis of the LV and biventricular dysfunction, RV ejection Fx 45% and LV ejection Fx 39%. He was subsequently placed on coumadin severe valvular heart disease), 2)HTN, 3)Rheumatic fever and heart disease, 4)COPD, 5)ETOH abuse, 6)colonic polyps, 7)CAD, 8)CHF, 9)Appendectomy, 10)Junctional tachycardia.

FHX: stroke, bone cancer, dementia.

SHX: 2ppd smoker since his teens; quit 2 years ago. 6-pack beer plus 2 drinks per day for many years: now claims he has been dry for 2 years. Denies illicit drug use.

EXAM: 36.8C, 90BPM, BP138/56.

MS: Alert and oriented to person, place, but not date. Hypophonic and dysarthric speech. 2/3 recall. Followed commands.

CN: Left homonymous hemianopia and left CN7 nerve palsy (old).

MOTOR: full strength throughout.

SENSORY: unremarkable.

COORDINATION: dysmetric FNF and HKS movements (left worse than right).

STATION: RUE pronator drift and Romberg sign present.

GAIT: shuffling and bradykinetic.

REFLEXES: 1+/1+ to 2+/2+ and symmetric throughout. Plantar responses were flexor bilaterally.

HEENT: Neck supple and no carotid bruits.

CV: RRR with 3/6 SEM and diastolic murmurs throughout the precordium.

Lungs: bibasilar crackles.

Fig. 2.7. Sample transcript of Neurology Domain

Table 2.2. Sample transcripts of all five medical domains

DESCRIPTION	MEDICAL_ SPECIALTY	SAMPLE_NAME	TRANSCRIPTION	KEYWORDS
EEG during wakefulness and light sleep is abnormal with independent, positive sharp wave activity seen in both frontotemporal head regions, more predominant in the right frontotemporal region.	Neurology	Video EEG - 3	PROCEDURE:EEG during wakefulness demonstrates background activity consisting of moderate-amplitude beta activity seen bilaterally. The EEG background is symmetric. Independent, small, positive, sharp wave activity is seen in the frontotemporal regions bilaterally with sharp-slow wave discharges seen more predominantly in the right frontotemporal head region. No clinical signs of involuntary movements are noted during synchronous video monitoring. right frontotemporal region. The EEG findings are consistent with potentially epileptogenic process. Clinical correlation is warranted.	neurology, epileptogenic, wakefulness, eeg, frontotemporal, activity
This is a 43-year-old female with a history of events concerning for seizures. Video EEG monitoring is performed to capture events and/or identify etiology.	Neurology	Video EEG	TIME SEEN: , 0734 hours and 1034 hours.,TOTAL RECORDING TIME: , 27 hours 4 minutes.,PATIENT HISTORY: , This is a 43-year-old female with a history of events concerning for seizures. Video EEG monitoring is performed to capture events and/or identify etiology., VIDEO EEG DIAGNOSES,1. AWAKE: Normal.,2. SLEEP: No activation.,3. CLINICAL EVENTS: None.,DESCRIPTION: , Approximately 27 hours of continuous 21-channel digital video EEG monitoring was performed. The waking background is unchanged from that previously reported.	neurology, electroencephalography, eeg monitoring, video eeg, seizures, eeg,
The patient has a history of epilepsy and has also had non-epileptic events in the past. Video EEG monitoring is performed to assess whether it is epileptic seizures or non-epileptic events.	Neurology	Video EEG - 1	DATE OF EXAMINATION: , Start: 12/29/2008 at 1859 hours. End: 12/30/2008 at 0728 hours.,TOTAL RECORDING TIME:, 12 hours, 29 minutes.,PATIENT HISTORY:, This is a 46-year-old female with a history of events concerning for seizures. The patient has a history of epilepsy and has also had non-epileptic events in the past. Video EEG monitoring is performed to assess whether it is epileptic seizures or non-epileptic events.,	neurology, non-epileptic events, temporal spike, eeg monitoring, video eeg, epilepsy, frequency, eeg, epileptic,
EEG during wakefulness, drowsiness, and sleep with synchronous video monitoring	Neurology	Video EEG - 2	IMPRESSION: ,EEG during wakefulness, drowsiness, and sleep with synchronous video monitoring demonstrated no evidence of focal or epileptogenic activity.	neurology, ekg artifact, video monitoring, wakefulness, drowsiness,

DESCRIPTION	MEDICAL_ SPECIALTY	SAMPLE_NAME	TRANSCRIPTION	KEYWORDS
demonstrated no evidence of focal or epileptogenic activity.				
Chronic venous hypertension with painful varicosities, lower extremities, bilaterally. Greater saphenous vein stripping and stab phlebectomies requiring 10 to 20 incisions, bilaterally.	Neurology	Vein Stripping	PREOPERATIVE DIAGNOSIS: , Chronic venous hypertension with painful varicosities, lower extremities, bilaterally.,POSTOPERATIVE DIAGNOSIS: , Chronic venous hypertension with painful varicosities, lower extremities, bilaterally.,PROCEDURES,1. Greater saphenous vein stripping and stab phlebectomies requiring 10 to 20 incisions, right leg.,2.	neurology, chronic venous hypertension, varicosities, stab phlebectomies, greater saphenous vein stripping, lower extremities, vein stripping, saphenous vein, vein, incisions, hemostasis, stripping, branches, phlebectomies, thigh, calf, saphenous,
The patient is a 17-year-old female, who presents to the emergency room with foreign body and airway compromise and was taken to the operating room.	General Medicine	Airway Compromise & Foreign Body - ER Visit	HISTORY OF PRESENT ILLNESS:, The patient is a 17-year-old female, who presents to the emergency room with foreign body and airway compromise and was taken to the operating room. She was intubated and fishbone.,PAST MEDICAL HISTORY: , Significant for diabetes, hypertension, asthma, cholecystectomy, and total hysterectomy and cataract.,ALLERGIES: ,	general medicine, diabetes, hypertension, asthma, cholecystectomy, fishbone, foreign body, airway compromise, airway,
Sore throat - Upper respiratory infection.	General Medicine	URI - SOAP	SUBJECTIVE:, Mom brings patient in today because of sore throat starting last night. Eyes have been very puffy. He has taken some Benadryl when all of this congestion started but with a sudden onset just yesterday. He has had low-grade fever and just felt very run down, appearing very tired. He is still eating and drinking well, and his voice has been hoarse but no coughing. No shortness of breath, vomiting, diarrhea or abdominal pain.,PAST MEDICAL HISTORY:,	general medicine, soap, uri, upper respiratory infection, water's view, congestion, light reflex, sore throat, respiratory, strep, infection,
Patient with worsening shortness of breath and cough.	General Medicine	Trouble breathing	CHIEF COMPLAINT:, "Trouble breathing.",HISTORY OF PRESENT ILLNESS:, A 37-year-old German woman was brought to a Shock Room at the General Hospital with worsening shortness of breath and cough. Over the year preceding admission, the patient had begun to experience the insidious onset of shortness of breath. She had smoked one half pack of cigarettes per day for 20 years, but had quit smoking approximately 2 months prior to admission.	

DESCRIPTION	MEDICAL_ SPECIALTY	SAMPLE_NAME	TRANSCRIPTION	KEYWORDS
Vacuum-assisted vaginal delivery of a third-degree midline laceration and right vaginal side wall laceration and repair of the third-degree midline laceration lasting for 25 minutes.	Obstetrics / Gynecology	Vaginal Delivery - Vacuum-Assisted	PREOPERATIVE DIAGNOSES,1. A 40 weeks 6 days intrauterine pregnancy.,2. History of positive serology for HSV with no evidence of active lesions.,3. Non-reassuring fetal heart tones.,POST OPERATIVE DIAGNOSES,1. A 40 weeks 6 days intrauterine pregnancy.,2. History of positive serology for HSV with no evidence of active lesions.,3. Non-reassuring fetal heart tones.,PROCEDURES,1.	obstetrics / gynecology, intrauterine pregnancy, non-reassuring fetal heart tones, vacuum-assisted vaginal delivery, vaginal side wall laceration, fetal heart tones, vaginal delivery,
Well-woman check up for a middle-aged woman, status post hysterectomy, recent urinary tract infection.	Obstetrics / Gynecology	Well-woman checkup	CHIEF COMPLAINT:, The patient comes for her well-woman checkup.,HISTORY OF PRESENT ILLNESS:, She feels well. She has had no real problems. She has not had any vaginal bleeding. She had a hysterectomy. She has done fairly well from that time till now. She feels like she is doing pretty well. She remains sexually active occasionally. She has not had any urinary symptoms. No irregular vaginal bleeding. She has not had any problems with vasomotor symptoms and generally, she just feels like she has been doing pretty well. She sometimes gets a catch in her right hip and sometimes she gets heaviness in her calves. She says the only thing that works to relieve that is to sleep on her tummy with her legs pulled up and they relax and she goes off to sleep.	
A 21-year-old female was having severe cramping and was noted to have a blighted ovum with her first ultrasound in the office.	Obstetrics / Gynecology	Vacuum D&C	PREOPERATIVE DIAGNOSIS: , Blighted ovum, severe cramping.,POSTOPERATIVE DIAGNOSIS:, Blighted ovum, severe cramping.,OPERATION PERFORMED: , Vacuum D&C.,DRAINS: , None.,ANESTHESIA: , General.,HISTORY: , This 21-year-old white female gravida 1, para 0 who was having severe cramping and was noted to have a blighted ovum with her first ultrasound in the office. Due to the severe cramping, a decision to undergo vacuum D&C was made. At the time of the procedure, moderate amount of tissue was obtained.,	obstetrics / gynecology, pitocin, single tooth tenaculum, vaginal vault, vacuum d&c, blighted ovum, speculum, tenaculum, curetting, blighted, cramping,
Laparoscopic-assisted vaginal hysterectomy. Abnormal uterine bleeding. Uterine fibroids.	Obstetrics / Gynecology	Vaginal Hysterectomy - Laparoscopic-Assisted	PREOPERATIVE DIAGNOSES,1. Abnormal uterine bleeding.,2. Uterine fibroids.,POSTOPERATIVE DIAGNOSES,1. Abnormal uterine bleeding.,2. Uterine fibroids.,OPERATION PERFORMED: , Laparoscopic-	obstetrics / gynecology, abnormal uterine bleeding, laparoscopic-assisted vaginal hysterectomy, uterine

DESCRIPTION	MEDICAL_ SPECIALTY	SAMPLE_NAME	TRANSCRIPTION	KEYWORDS
			assisted vaginal hysterectomy.,ANESTHESIA: , General endotracheal anesthesia.,	fibroids, bipolar electrocautery, vaginal hysterectomy, vicryl sutures, tooth, uterine, uterosacral, laparoscope, electrocautery, hysterectomy, laparoscopic, coagulated, vaginal, ligament, transected
Surgical removal of completely bony impacted teeth #1, #16, #17, and #32. Completely bony impacted teeth #1, #16, #17, and #32.	Dentistry	Bony Impacted Teeth Removal	PREOPERATIVE DIAGNOSIS:, Completely bony impacted teeth #1, #16, #17, and #32.,POSTOPERATIVE DIAGNOSIS: , Completely bony impacted teeth #1, #16, #17, and #32.,PROCEDURE: , Surgical removal of completely bony impacted teeth #1, #16, #17, and #32.,ANESTHESIA: , General nasotracheal.,COMPLICATIONS: , None.,CONDITION: ,Stable to PACU.,DESCRIPTION OF PROCEDURE: , Patient was brought to the operating room, placed on the table in a supine position, and after demonstration of an adequate plane of general anesthesia via the nasotracheal route, patient was prepped and draped in the usual fashion for an intraoral procedure. A gauze throat pack was placed and local anesthetic was administered in all four quadrants, a total of 7.2 mL of lidocaine 2% with 1:100,000 epinephrine, and 3.6 mL of bupivacaine 0.5% with 1:200,000 epinephrine.	dentistry, intraoral, bony impacted teeth, throat pack, buccal aspect, saline solution, gut sutures, envelope flap, periosteal elevator,
Patient has had multiple problems with his teeth due to extensive dental disease and has had many of his teeth pulled, now complains of new tooth pain to both upper and lower teeth on the left side for approximately three days..	Dentistry	Toothache - ER Visit	CHIEF COMPLAINT:, Toothache.,HISTORY OF PRESENT ILLNESS: ,This is a 29-year-old male who has had multiple problems with his teeth due to extensive dental disease and has had many of his teeth pulled. Complains of new tooth pain. The patient states his current toothache is to both upper and lower teeth on the left side for approximately three days. The patient states that he would have gone to see his regular dentist but he has missed so many appointments that they now do not allow him to schedule reg are no new dental fractures. The oropharynx is normal without any sign of infection.	dentistry, odontalgi, multiple dental caries, dentist, dental disease, extensive dental disease, teeth pulled, lower teeth, cervical lymphadenopathy, dental caries, toothache, erythema, swelling, teeth, dental,

DESCRIPTION	MEDICAL_ SPECIALTY	SAMPLE_NAME	TRANSCRIPTION	KEYWORDS
Extraction of teeth. Incision and drainage (I&D) of left mandibular vestibular abscess adjacent to teeth #18 and #19.	Dentistry	Teeth Extraction & I&D	PREOPERATIVE DIAGNOSES,1. Carious teeth #2, #5, #12, #15, #18, #19, and #31.,2. Left mandibular vestibular abscess.,POSTOPERATIVE DIAGNOSES,1. Carious teeth #2, #5, #12, #15, #18, #19, and #31.,2. Left mandibular vestibular abscess.,PROCEDURE,1. Extraction of teeth #2. #5, #12, #15, #18, #19, #31.,2. Incision and drainage (I&D) of left mandibular vestibular abscess adjacent to teeth #18 and #19.,ANESTHESIA:, General nasotracheal.,COMPLICATIONS: , None.,DRAIN:, Quarter-inch Penrose drain place in left mandibular vestibule adjacent to teeth #18 and #19, secured with 3-0 silk suture.,CONDITION:, The patient was taken to the PACU in stable condition	dentistry, mandibular, vestibular, abscess, throat pack, purulent material, forceps extraction, nasogastric tube, carious teeth, incision, teeth, nasogastric, carious, extraction
Removal of cystic lesion, removal of teeth, modified Le Fort I osteotomy.	Dentistry	Teeth Extraction	PREOPERATIVE DIAGNOSES,1. Basal cell nevus syndrome.,2. Cystic lesion, left posterior mandible.,3. Corrected dentition.,4. Impacted teeth 1 and 16.,5. Maxillary transverse hyperplasia.,POSTOPERATIVE DIAGNOSES,1. Basal cell nevus syndrome.,2. Cystic lesion, left posterior mandible.,3. Corrected dentition.,4. Impacted teeth 1 and 16.,5. Maxillary transverse hyperplasia.,PROCEDURE,1. Removal of cystic lesion, left posterior mandible.,2. Removal of teeth numbers 4, 13, 20, and 29.,3. Removal of teeth numbers 1 and 16.,4. Modified Le Fort I osteotomy	dentistry, nevus syndrome, basal cell, mandible, teeth, hyperplasia, cystic lesion, osteotomy, le fort, le fort osteotomy, orotracheal route, bony crypt, watertight, removal of cystic lesion, le fort i osteotomy, aspect of the maxilla, modified le fort, molar tooth, posterior mandible, maxillary, molar, tooth,
2-D M-Mode. Doppler.	Cardiovascular / Pulmonary	2-D Echocardiogram - 1	2-D M-MODE: , ,1. Left atrial enlargement with left atrial diameter of 4.7 cm.,2. Normal size right and left ventricle.,3. Normal LV systolic function with left ventricular ejection fraction of 51%.,4. Normal LV diastolic function.,5. No pericardial effusion.,6. Normal morphology of aortic valve, mitral valve, tricuspid valve, and pulmonary valve.,7. PA systolic pressure is 36 mmHg.,DOPPLER: , ,1. Mild mitral and tricuspid regurgitation.,2. Trace aortic and pulmonary regurgitation.	cardiovascular / pulmonary, 2-d m-mode, doppler, aortic valve, atrial enlargement, diastolic function, ejection fraction, mitral, mitral valve, pericardial effusion, pulmonary valve, regurgitation, systolic function, tricuspid, tricuspid valve, normal lv

DESCRIPTION	MEDICAL_ SPECIALTY	SAMPLE_NAME	TRANSCRIPTION	KEYWORDS
2-D Echocardiogram	Cardiovascular / Pulmonary	2-D Echocardiogram - 2	1. The left ventricular cavity size and wall thickness appear normal. The wall motion and left ventricular systolic function appears hyperdynamic with estimated ejection fraction of 70% to 75%. There is near-cavity obliteration seen. There also appears to be increased left ventricular outflow tract gradient at the mid cavity level consistent with hyperdynamic left ventricular systolic function. There is abnormal left ventricular relaxation pattern seen as well as elevated left atrial pressures seen by Doppler examination.,2. The left atrium appears mildly dilated.,3. The right atrium and right ventricle appear normal.,4. The aortic root appears normal.,5. The aortic valve appears calcified with mild aortic valve stenosis, calculated aortic valve area is 1.3 cm square with a maximum instantaneous gradient of 34 and a mean gradient of 19 mm.,6. There is mitral annular calcification extending to leaflets and supportive structures with	cardiovascular / pulmonary, 2-d, doppler, echocardiogram, annular, aortic root, aortic valve, atrial, atrium, calcification, cavity, ejection fraction, mitral, obliteration, outflow, regurgitation, relaxation pattern, stenosis, systolic function, tricuspid, valve, ventricular, ventricular cavity, wall motion, pulmonary artery
2-D Echocardiogram	Cardiovascular / Pulmonary	2-D Echocardiogram - 3	2-D ECHOCARDIOGRAM,Multiple views of the heart and great vessels reveal normal intracardiac and great vessel relationships. Cardiac function is normal. There is no significant chamber enlargement or hypertrophy. There is no pericardial effusion or vegetations seen. Doppler interrogation, including color flow imaging, reveals systemic venous return to the right atrium with normal tricuspid inflow. Pulmonary outflow is normal at the valve. Pulmonary venous return is to the left atrium. The interatrial septum is intact. Mitral inflow and ascending aorta flow are normal. The aortic valve is trileaflet. The coronary arteries appear to be normal in their origins. The aortic arch is left-sided and patent with normal descending aorta pulsatility.	cardiovascular / pulmonary, 2-d echocardiogram, cardiac function, doppler, echocardiogram, multiple views, aortic valve, coronary arteries, descending aorta, great vessels, heart, hypertrophy, interatrial septum, intracardiac, pericardial effusion, tricuspid, vegetation, venous, pulmonary

DESCRIPTION	MEDICAL_ SPECIALTY	SAMPLE_NAME	TRANSCRIPTION	KEYWORDS
A white woman in her 47th year who is concerned about a probable spider bite on the left side of her neck shows herself to the hospital..."	General Medicine	Possible Spider Bite	This white female, age 47, arrives with a potential spider bite on the left side of her neck. It's unclear if she's been hurt, and she has no idea what kind. Her left rear shoulder has been bothering her for the last two days, and it's become sensitive and red.	basic medicine, spider bite, damage, soreness, redness, insect bite.
Hysterectomy via the vaginal canal. After doing so, a weighted speculum was inserted into the posterior vaginal vault...."	Obstetrics / Gynaecology	Vaginal Delivery - Vacuum-Assisted	Preoperative Diagnosis: 1. Intrauterine gestation lasting 40 weeks and 6 days. HSV-positive history without current signs of infection.	"obstetrics/gynecology, intrauterine pregnancy, non-reassuring foetal heart tones, vacuum-assisted vaginal birth."
"Fractures to both sides of the jaw, as well as to the left angle and the symphysis, were all open. A closed reduction of the mandibular fracture was performed with MMF..."	Dentistry	Closed Reduction - Mandible Fracture	Diagnosis: "Bilateral open mandible fracture, open left angle, and open symphysis." Diagnosis: open left angle and symphysis fractures, open both mandibles. Multiple-Mode Fluoroscopic-Assisted Anesthesia (MMF.ANESTHESIA): Induction of general anaesthesia by means of..."	*dentistry, closed reduction, mmf, endotracheal, pacu, bilateral open mandible fracture, symphysis fracture, mandible fracture*
"Right supraclavicular lymphadenopathy was discovered during a regular checkup. Right supraclavicular lymphadenopathy reappeared during her follow-up."	Cardiovascular / Pulmonary	Supraclavicular Lymphadenopathy	As we go over the patient's medical history, we find that they have hypertension and suffer from occasional heartburn. Her regular mammograms have all come back cancer-free.	*congenital heart disease, cyanotic, ductal-dependent, pulmonary blood flow, ventricular septal defect, blood flow*

2.7. Conclusion

In conclusion, this literature survey delved into the significance of automatic extractive summarization within the biomedical field, underscoring its importance in Natural Language Processing and Artificial Intelligence. The core objective of automatic text summarization is to distill essential information from source texts while maintaining their original meaning and context. The survey reviewed two primary approaches—extractive and abstractive—as well as the different types of summarization based on document type (mono-document vs. multi-document) and purpose (generic vs. query-based).

Key challenges in extractive summarization were identified, including issues of redundancy, coherence, loss of critical information, and reliance on sentence length and structure. The evolution of summarization techniques from traditional probabilistic models to sophisticated deep neural networks, such as sequence-to-sequence models with LSTM and GRU architectures, was discussed. Additionally, the survey emphasized the need for robust evaluation methods for automatically generated summaries, highlighting the benefits and limitations of both manual and automatic evaluation systems.

The findings indicate a need for further research to develop improved extractive summarization techniques, enhancing information retrieval and knowledge dissemination in the biomedical sector. Effective data collection is vital in medical and life sciences research, influencing the accuracy and reliability of research outcomes. Researchers can choose between primary and secondary data collection methods, based on factors like the research question, data type, target population, and available resources. Making informed decisions in data collection ensures the acquisition of reliable data, leading to better clinical practices and improved patient outcomes. This survey lays the groundwork for three proposed approaches to summarizing medical transcripts, detailed in chapters 3, 4, and 5.

CHAPTER 3

A Novel Method for Text Summarization using Masked Language Modelling & UML Metathesaurus

3.1. Introduction

In the realm of biomedical research, there's a growing demand for effective summarization techniques to handle the vast amount of generated data. In this chapter, a new corpus specific to the biomedical domain has been proposed that aims to identify significant features for summarizing biomedical transcripts. PubMed databases from BioMed Central and MTsamples data is used for the features identification then these features are passed Masked Language Modeling (MLM) to provide suitable summarization of the required text data. This chapter addresses the challenge of extractive summarization for biological materials by leveraging characteristics specific to the biomedical domain. The proposed solution comprises two key steps. Initially, utilizing the Metathesaurus from the Unified Medical Language System (UMLS), named entities' concepts are extracted, focusing on frequently occurring ideas. The collection of common concepts forms the basis for constructing an initial extraction summary. The shortest path in the graph, determined by the weights of connecting edges among common idea sets, contributes to preliminary summary. The output from the first phase is then transitioned to the second phase using a transfer learning-based approach based on BERT that provide the brief and accurate summary of the provided text. An overall ROUGE score of 74.80% is attained. The results indicate that the proposed strategy enhances the comprehension of key ideas and sentences in biological data and a concise extractive summary of the text.

The number of electronic documents in the biomedical area has greatly expanded as a result of the rapid growth of the Internet and other technologies. Nowadays, a variety of services (online clinical reports, biomedical literature databases, and electronic health record systems) provide access to medical information and biomedical literature in a variety of formats, including research papers for patient health records. However, getting the necessary and pertinent information from such kind of data set is very time-consuming and stressful for clinical researchers due to the volume of biological data, and the frequency of their updates. The key to solving this issue is text summarization. Humans can automatically and effectively identify and extract pertinent information from a vast volume of textual material with the use of text-summarizing technologies. Text summarizing strategies aim to condense

the information included in one or more papers by focusing only on the most crucial ideas and concepts. Extracting the key ideas and details from the reference material and then interpreting them to create an integrated summary are the two challenges of the summarizing assignment. The work of text summarization can be approached in one of two ways: extractive or abstractive. In contrast to the abstractive work, which requires creating new sentences from significant data retrieved from the corpus, the extractive based task is merging the key sentences retrieved from the collection of documents into a summary. In recent years, a number of summarizing techniques that choose the most pertinent and non-redundant phrases by combining the information supplied by an item set oriented model embedded with a statistical evaluator. The majority of these techniques are based on the frequently used item set in the mining. UMLS was utilized by Nasr Azadani et al. to develop a concept-oriented framework for processing input documents. A graph-based similarity indicator is then created using frequent item set mining approach. The extractive-oriented summarization process then uses the minimal spanning tree clustering approach to identify the document's subtopics. This method has been tested against general purpose graphs and graphs specifically geared for biomedicine [117].

Deep learning techniques, which are now widely used for text summaries, include the capacity to learn language-oriented models that enable the creation of digestible summaries. With the aid of a deep learning model, Verma et al. proposed an extractive-based text summarizing approach for factual-oriented reports and looked into several aspects to enhance the collection of sentences chosen for the information summary [7]. Similarly, the transfer learning-based ULMFiT framework was presented which may be used for any NLP job. Instead of training a model from scratch for a subsequent task, it offers the option of doing such tasks.

A deep-reinforced abstract summarization technique was also that analyses the abstract of various biomedical publications and generates a summary of those studies into a title or headline. To do this, they developed a novel reinforcement-based learning incentive using biomedical expert resources like the UMLS and demonstrated that their proposed model can generate domain-centric abstractive-based summaries. Moreover, they included a compensation scheme based on the TF-IDF and demonstrated that their model was capable of learning domain-specific data without the aid of experts or specialized equipment. An NLP-trained generative pre-trained transformer (GPT) was proposed. Delvin et al. developed a method based on this technique called the bidirectional encoder-based

representation for transformers called as BERT, which uses language models to learn bi-directional encoder representations [141]. In BERT, computation is more difficult, but its practical correctness is substantially stronger. The failure of this paradigm to recognize and comprehend the negation is one of its flaws. The GPT-2, extremely similar to GPT but has a few alterations, was introduced at the beginning of year 2019 by Radford et al. The GPT-2 language model has around 1.5 billion input parameters and that was trained using 40 GB of text to anticipate the next word. In order to create Wikipedia articles, Liu et al. introduced an abstract-based summarization scheme that included an extract-oriented pre-processing stage [142]. Ranking paragraphs based on how important a paper's reference links are is the goal of the pre-processing stage. To do this, they chose indices subset among the ranking graphs and fed it into the transforming decoder. In another paradigm proposed, every NLP job is viewed by way of a text-to-text issue [164]. While processing a series of text, self-attention replaces each input with a weighted sum of the remainder of the sequence. The Colossal Clean Crawled Corpus data set was used to train this model, which is referred to as T-5 (text-to-text transfer-based transformer). In NLP, it has improved self-supervised learning through rather comprehensive trials. With 11 billion parameters in 17 of 24 tests, T5 performs at the highest level. To make an abstract summary with many sentences in this respect, Zolotareva et al. employed the T-5 model [5]. Although massive research has been done in the past however, their exist some area of improvements in text summarization in biomedical data are;

- **Word-Level Analysis Limitation**: The majority of existing summarization technologies operate by analyzing text at the word level. This means that the input document is modelled based on individual words, without considering their semantic links and meanings [143]. This approach's performance is deemed unsatisfactory as it solely relies on vocabulary and neglects the specific characteristics of the document's domain.
- **Unsuitability for Specialized Fields**: In fields like biomedicine, where documents are highly specialized and interpretative, the interdependence of grammar and meaning becomes a significant challenge [144]. The prevalent word-level analysis approach struggles to capture the intricacies of such documents, leading to suboptimal summarization.
- **Biomedical Complexity**: The biomedical field introduces additional complexities, such as a wide range of synonyms, abbreviations, acronyms, and the necessity of incorporating domain-specific attributes. These intricacies pose a considerable hurdle for

summarization technologies that do not account for the nuanced and domain-specific nature of biomedical documents.

To address these issues, a new method for text summarization has been proposed. The proposed approach involves two primary steps. Firstly, it leverages the Metathesaurus sourced from the Unified Medical Language System (UMLS) to extract concepts related to named entities, with an emphasis on frequently recurring ideas. Then these commonly identified concepts serve as the foundation for creating an initial extractive summary deep learning method and using the BERT method. Then the concise summary is generated for the biomedical text data.

3.2. Proposed Approach

Two methods can be used: direct, and indirect. In the direct method, it employs abstract summarizing and NLP algorithms on the entire material to produce the summary. On the other hand, the indirect method involves using an extractive summarizing technique to generate an extractive summary first, followed by applying an abstractive summarizing approach to produce the final summary. The extractive and abstractive summarizing techniques play a crucial role in creating a useful final summary. If the extractive summarizing technique accurately chooses the most valuable terms of the material, the outcomes of applying NLP methodologies will be of higher efficiency. A contextual embedding model incorporates domain knowledge during pre-training through a unique knowledge augmentation approach. This involves enhancing with the Unified Medical Language System (UMLS) Metathesaurus in two ways: (i) establishing connections between words that share the same underlying 'concept' in UMLS, and (ii) utilizing semantic type knowledge from UMLS to generate input embeddings with clinical relevance. Through these strategies, UML and BERT effectively encode clinical domain knowledge into word Embeddings, demonstrating superior performance compared to existing domain-specific models in text summarization. This investigation employs the indirect method, where graph creation and frequent item set mining approaches are utilized for extract-oriented summarization, and a BERT learning-based methodology is employed for abstract-oriented summarization. The Metathesaurus is currently extensive, containing 15.5 million atoms organized into 4.28 million concepts sourced from 214 vocabularies. However, the sheer size poses challenges in terms of maintenance, proving to be a resource-intensive, time-consuming, and demanding task for human expert editors.

3.2.1. Bidirectional Encoder Representations from Transformers (BERT)

BERT, an innovative natural language processing (NLP) pre-training technique based on transformers, was introduced by Google in 2018. Jacob Devlin and his team are recognized for its development. Contextual word embedding models like ELMo [147] and BERT [148], [149] have demonstrated exceptional performance across various NLP tasks, outperforming existing methods. The advancements in BERT research have significantly impacted NLP tasks such as MNLI, sentiment analysis, and text summarization. The BERT process is illustrated in Figure 3.1.

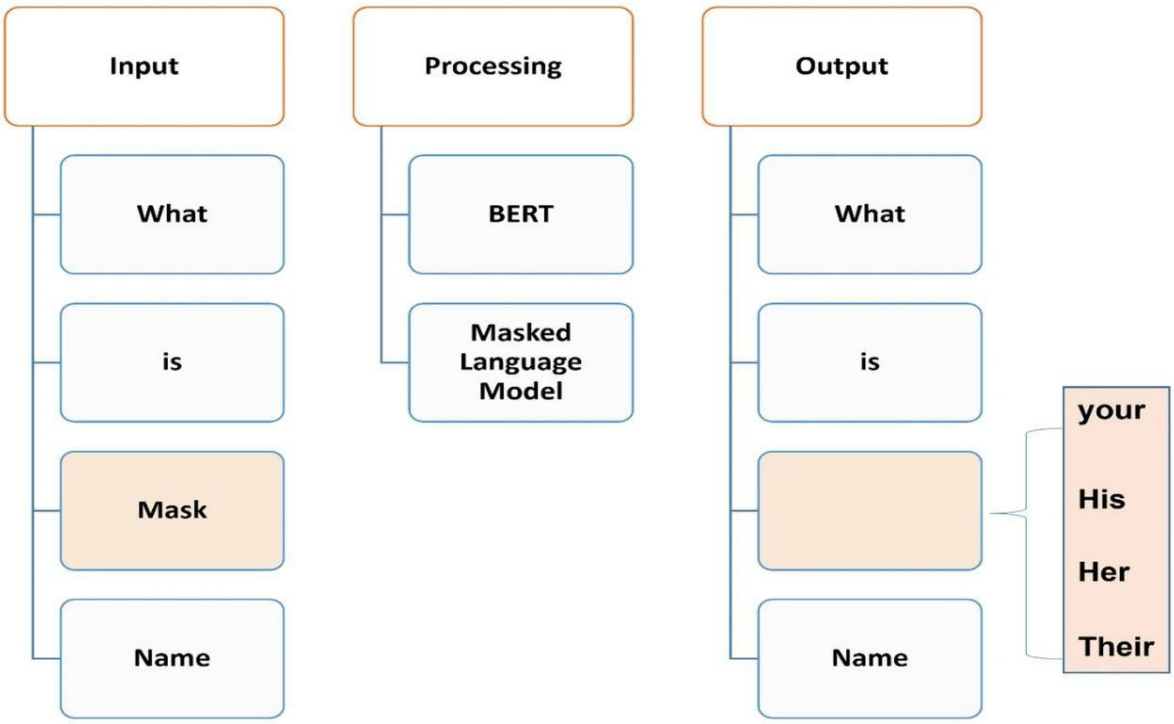


Fig. 3.1. Process of BERT [141]

BERT's architecture is based on the Transformer model and its key innovation lies in bidirectional training for language modelling. This differs from previous approaches which only considered either left-to-right or combined left-to-right and right-to-left training. MLM enables representations to incorporate contextual information from both the left and right sides, facilitating deep bidirectional Transformer pre-training. By leveraging bidirectional representations, the BERT model improves its capacity to understand the meaning of a word within its contextual context, particularly within a sentence. In the pre-training phase of BERT, two self-supervised tasks are utilized. The first task involves Masked Language Modeling (LM), followed by the second task, Next Sentence Prediction.

Before inputting word sequences into BERT, each sequence has 15% of its words replaced with [MASK] tokens. BERT then utilizes the context by non-masked words to predict the original values of the masked words. Within the Masked Language Modeling (LM) task, 15% of the tokens in each sentence are replaced with a [MASK] token. For the U^{jth} input token in the sentence, an input embedding is then generated.

$$u_{input}^{(j)} = p^{(j)} + SEGseg_{Id}^{(J)} + Ew_j \quad (3.1)$$

BERT employs a procedure that involves adding a classification layer on top of the encoder output, transforming output vectors using an embedding matrix to match the lexical dimension, and calculating the probability of each word within the vocabulary using softmax. Here, p(j) $\in \mathbb{R}^d$ represents the position embedding of the jth token in the sentence, where d is the hidden dimension of the transformer. SEG $\in \mathbb{R}^{(d\times 2)}$ is known as the segment embedding, and $SEG_{id} \in \mathbb{R}^2$, a one-hot vector, signifies the segment ID indicating the sentence to which the token belongs. In the context of Masked LM, the model operates with a single sentence, implying that the segment ID indicates that all tokens belong to the first sentence. This process results in a prediction for the original value of the masked word.

The input embedding vectors pass through multiple transformer layers, utilizing attention mechanisms. Each layer generates a contextualized embedding for each token. Subsequently, for every masked token w, the model generates a score vector $y_w \in \mathbb{R}$^D, aiming to minimize the cross-entropy loss between the softmax of y_w and the one-hot vector corresponding to the masked token (h_w).

$$loss = -\log\left(\frac{\exp(y_w[W])}{\exp(y_w[W'])}\right) \quad (3.2)$$

In this way tokens are masked by the training words, here we have trained the BERT using the UML Metathesaurus for the biomedical concepts.

3.2.2. Unified Medical Language System (UMLS) Metathesaurus

The UMLS Metathesaurus, developed by the National Library of Medicine, serves as a comprehensive system for integrating biomedical terminologies from more than 200 sources. At the core of the Metathesaurus is the "atom," which represents a term originating from a source vocabulary. This system facilitates the linkage of words representing identical or similar concepts. For example, terms like 'lungs' and 'pulmonary,' which share a similar

meaning, can be associated with the same concept unique identifier (CUI) such as C0024109.

Furthermore, UMLS enables the classification of concepts based on their semantic type. For instance, 'skeleton' and 'skin' are grouped under the 'Body System' semantic type. Each row in the matrix corresponds to a distinct semantic type in UMLS to which a word is linked.

As mentioned earlier, the UMLS Metathesaurus is built around the concepts of "atom" and "concept." An "atom" represents a term from a specific source vocabulary while a "concept" is a grouping of synonymous atoms. Table 3.2 provides examples of atoms and the various types of identifiers assigned to them.

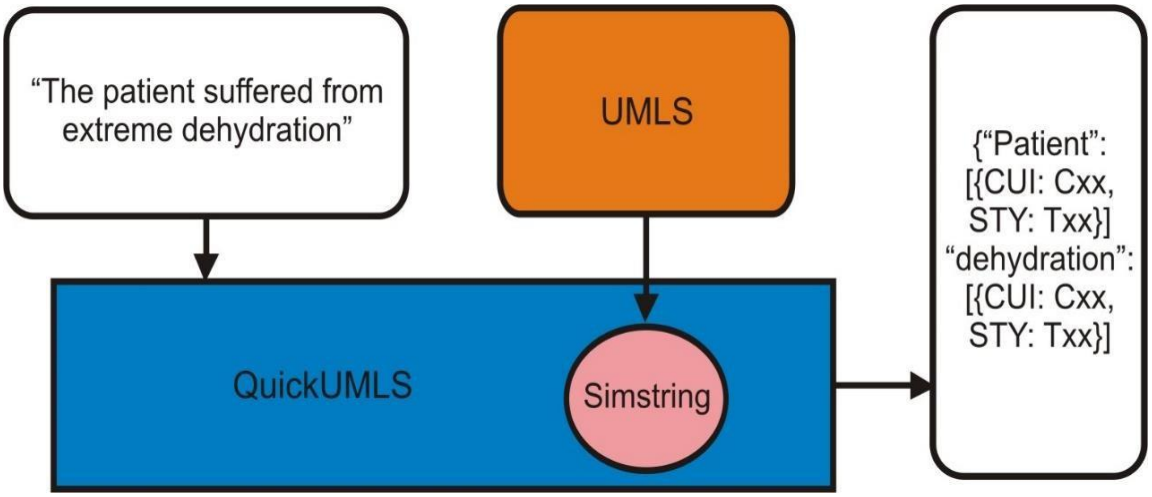


Fig. 3.2. UMLS structurer

As Metathesaurus editors also assign semantic types to each UMLS concept. Table 3.1 illustrates examples of atoms and the diverse types of identifiers assigned to them. Semantic types are associated with Concept Unique Identifiers (CUIs) rather than Atom Unique Identifiers (AUIs). However, understanding the semantics of an atom can be approximated by deducing it from the source vocabulary, especially for vocabularies that have consistent semantic content like anatomy ontologies. Another approach is to consider the highest-level categories of a vocabulary for those that encompass a wide range of topics.

Table 3.1. Examples of atoms and the diverse types of identifiers

Tuple	String	Source	SCUI	AUI	SUI	LUI	CUI	Semantic Group
t_1	Headache	MSH	M0009824	A0066000	S0046854	L0018681	C0018681	Disorders
t_2	Headaches	MSH	M0009824	A0066008	S0046855	L0018681	C0018681	Disorders
t_3	Cranial Pains	MSH	M0009824	A1641924	S1680379	L1406212	C0018681	Disorders
t_4	Cephalodynia	MSH	M0009824	A26628141	S0475647	L0380797	C0018681	Disorders
t_5	Cephalodynia	SNOMEDCT_US	25064002	A2957278	S0475647	L0380797	C0018681	Disorders
t_6	Headache (finding)	SNOMEDCT_US	25064002	A3487586	S3345735	L3063036	C0018681	Disorders

Let us consider three tuple pairs (t_1, t_3), (t_4, t_5), and (t_1, t_5) from Table 4.1 with

t_1 = (“Headache”, “MSH”, “M0009824”, “Disorders”)

t_3 = (“Cranial Pains”, “MSH”, “M0009824”, “Disorders”)

t_4 = (“Cephalodynia”, “MSH”, “M0009824”, “Disorders”)

t_5 = (“Cephalodynia”, “SNOMEDCT_US”, “25064002”, “Disorders”).

UMLS concepts in response to a given text, along with their similarity to the query string and other relevant information. To illustrate, for the text "The patient had a haemorrhage," UMLS produces candidate concepts using a default string similarity threshold of 0.7, shown in Fig 3.3.

```
{‘term’: ‘Inpatient’, ‘cui’: ‘C1548438’, ‘similarity’: 0.71,
‘semtypes’: {‘T078’}, ‘preferred’: 1},
{‘term’: ‘Inpatient’, ‘cui’: ‘C1549404’, ‘similarity’: 0.71,
‘semtypes’: {‘T078’}, ‘preferred’: 1},
{‘term’: ‘Inpatient’, ‘cui’: ‘C1555324’, ‘similarity’: 0.71,
‘semtypes’: {‘T058’}, ‘preferred’: 1},
{‘term’: ‘*^patient’, ‘cui’: ‘C0030705’, ‘similarity’: 0.71,
‘semtypes’: {‘T101’}, ‘preferred’: 1},
{‘term’: ‘patient’, ‘cui’: ‘C0030705’, ‘similarity’: 1.0, ‘semtypes’:
{‘T101’}, ‘preferred’: 0},
{‘term’: ‘inpatient’, ‘cui’: ‘C0021562’, ‘similarity’: 0.71,
‘semtypes’: {‘T101’}, ‘preferred’: 0}
```

Fig. 3.3. Words and masked words

To comprehend the semantic connections among words sharing the same Concept Unique Identifier (CUI) in a biomedical context, the UML and BERT model is employed. An illustrative scenario involves predicting the masked word 'lungs' both with and without the inclusion of clinical information, as depicted in Fig. 4.3. In this representation, model

endeavors to recognize words such as 'lung,' 'lungs,' and 'pulmonary' since all three are linked to the identical CUI (C0024109).

Introducing a unique framework to enhance contextual embeddings with clinical domain expertise, we have incorporated domain knowledge from a clinical Metathesaurus during the pre-training stage of a BERT-based model. This approach aims to construct 'semantically enriched' contextual representations that draw advantages from both the contextual learning offered by the BERT architecture and the domain-specific knowledge encapsulated in the UMLS Metathesaurus. Recent experiments, utilizing supervised learning approaches with word embeddings, have demonstrated promising results in the context of the Metathesaurus. These findings affirm that such approaches exhibit reasonably good performance for aligning selected subsets of source vocabularies within the Metathesaurus. Upon identifying all word ngram candidates, we executed a query across the entire UMLS database to locate concepts that partially correspond to these word ngrams. Given the inefficiency and difficulty of exact matching on such an extensive database, employed approximate string matching through simstring shown in Fig. 3.4.

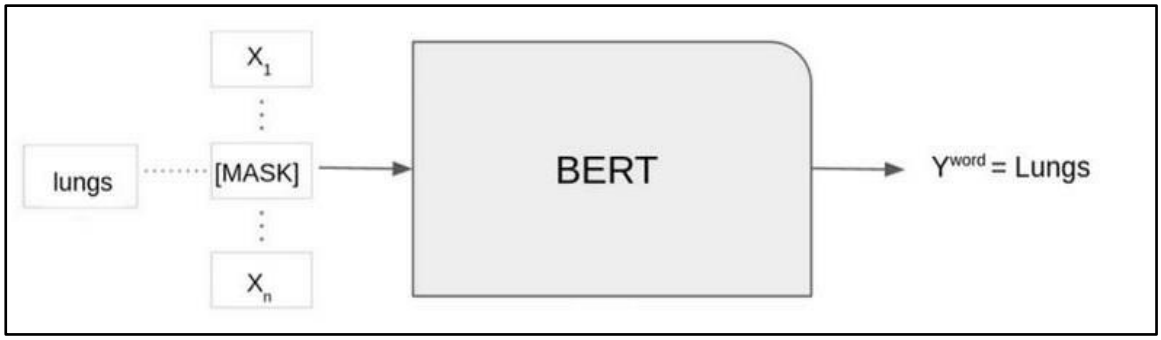


Fig. 3.4. UMLS MASK with BERT

This study primarily focuses on evaluating the viability of implementing deep learning (DL) techniques for large-scale terminology integration within the UMLS Metathesaurus. Unlike typical DL benchmarking studies, our investigation is not primarily technical. Instead, our aim to explore whether a straightforward DL approach can surpass the established editorial rules guiding the construction of the UMLS Metathesaurus shown in Fig 3.5.

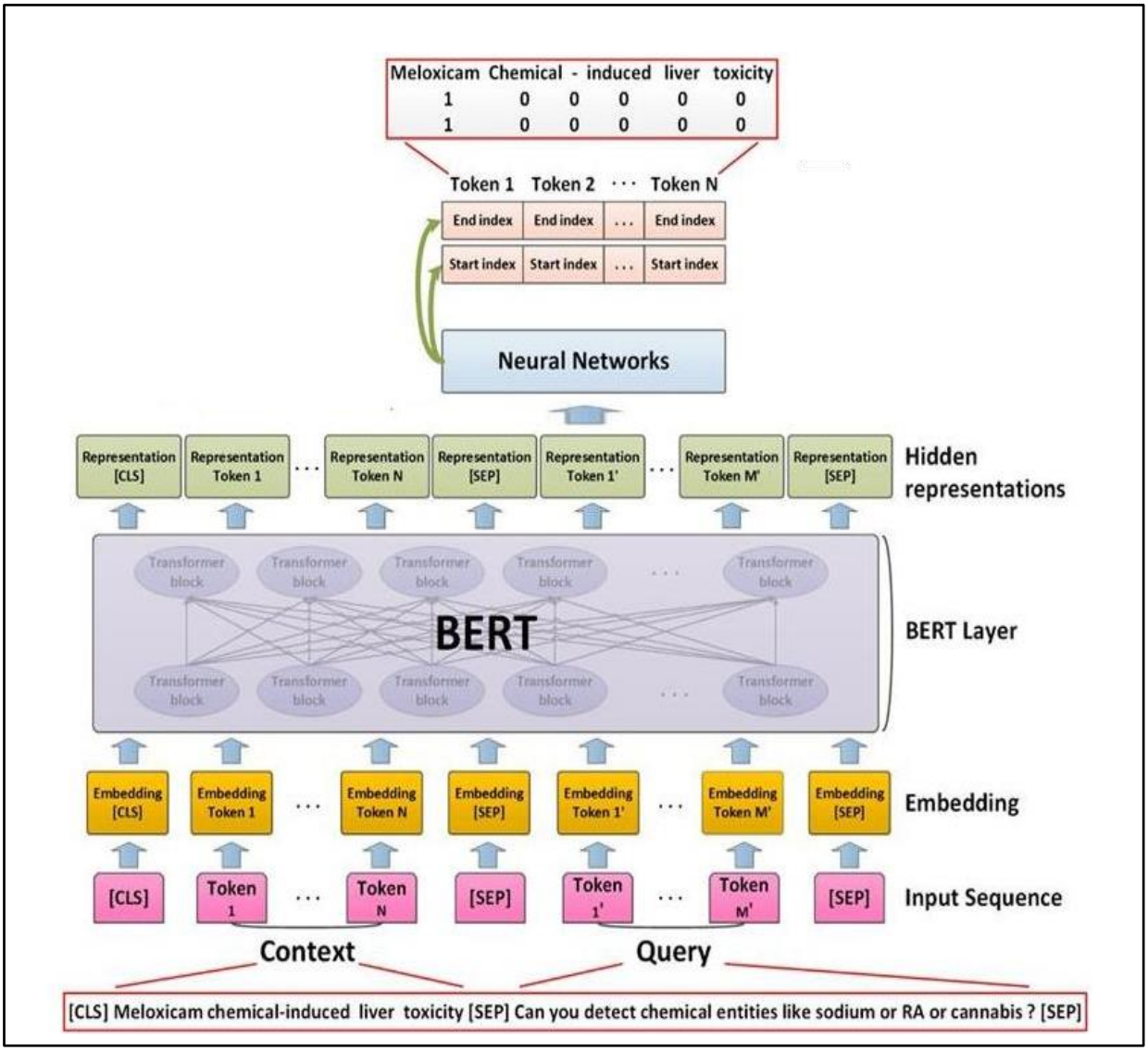


Fig 3.5. Proposed Framework

3.2.3. Algorithm for the proposed Approach

Input: Biomedical text T, UMLS Metathesaurus U, BERT model B

Output: *Summary of Transcripts S*

Concept Extraction:

$C = \{c1, c2, ..., cn\} \leftarrow$ ExtractConcepts(T, U) # where c_i is a concept from UMLS

Graph Construction:

$G(V, E) \leftarrow$ ConstructGraph(C) # where $V = C$ and $E = \{(c_i, c_j) \mid c_i, c_j \in C$ and are semantically related in U}

Sentence Ranking:

Let Sent = {s1, s2, ..., sm} be the set of sentences in T

For each $s_i \in$ Sent:

score(s_i) = Σ PageRank(cj, G) for all cj ∈ si ∩ C

ScoredSent = {(si, score(si)) | si ∈ Sent}

Initial Extractive Summary:

E ← Top-k(ScoredSent) #where k is a predefined number of sentences

BERT-based Summarization:

TokensE = Tokenize(E) using BERT tokenizer

MaskedE = ApplyMasking(TokensE) with probability p

EmbeddingsE = B(MaskedE)

for each masked token ti in MaskedE:

P(ti) = SoftMax(Linear(EmbeddingsE[i]))

B = GenerateSummary(P(ti) for all masked ti)

Summary Integration:

S = λE + (1-λ)B, where λ ∈ [0,1] is a weighting factor

Evaluation:

for reference summary R:

ROUGE-1(S, R) = F1-score of unigram overlap

ROUGE-2(S, R) = F1-score of bigram overlap

Return S

In the algorithm, A sequential model is appropriate for building a linear stack of layers, simplifying the construction of a text processing pipeline. The Embedding layer transforms words into dense vectors, capturing semantic relationships and preparing the text for further processing. The Convolutional layer detects local patterns and features within the text sequences, enhancing the model's ability to learn important characteristics. MaxPooling reduces the size of the feature maps, lowering computational complexity and helping prevent overfitting by focusing on the most significant features. The LSTM layer captures long-term dependencies and sequential patterns, which are crucial for understanding the context and

meaning in text. The Dense layer serves as the final layer to convert the extracted features into the desired output format, such as a summary. The softmax activation function in the final layer normalizes the output probabilities, ensuring they sum to one and facilitating multi-class classification. Compiling the model with the appropriate loss function, optimizer, and metrics prepares it for training, aiming to minimize the loss and improve accuracy. Training the model on labelled data allows it to learn the mapping from input transcripts to summaries, and validation ensures it generalizes well to new data. Padding ensures consistent input length, and generating summaries for test data demonstrates the model's practical application and performance. Evaluation metrics like ROUGE and F1 score provide quantitative measures of the model's performance, enabling comparison with other methods and understanding its effectiveness.

Embedding matrix into our model's input embedding, we identify all words with clinical meanings as defined in UMLS. For each of these identified words, we extract the corresponding Concept Unique Identifier (CUI) and semantic type. BERT utilizes the GELU (Gaussian Error Linear Unit) activation function. In BERT, the loss function focuses exclusively on predicting masked values while disregarding the prediction of non-masked words. Consequently, the model converges at a slower pace compared to unidirectional models. However, this drawback is mitigated by its heightened contextual awareness.

PageRank is a well-established algorithm used to assess the importance of vertices in a graph [151]. This is achieved by evaluating both the quantity and quality of links each vertex possesses. Vertices with higher scores are considered more significant due to their connections with other high-quality vertices. The PageRank score is recursively computed for each vertex Vi, with the damping factor regulating the likelihood of further graph traversal.

It's noteworthy that we selected the UMLS Metathesaurus and BERT model for two primary reasons:

- To develop a clinical contextual embedding model capable of seamlessly integrating domain-specific (medical) knowledge.
- The UMLS Metathesaurus serves as a comprehensive compilation of numerous renowned biomedical vocabularies (e.g., MeSH [152]).

Our goal is to underscore the positive influence of incorporating domain knowledge in our study. Instead of exploring complex layers, such as the Bi-LSTM layer as utilized in [145].

we have integrated domain knowledge. Our emphasis lies in demonstrating that the combination of UML and BERT surpasses other medical-based BERT models in performance across diverse medical NLP tasks.

3.3. Results

The Recall-Oriented Understudy for Gisting Evaluation metric is used to assess the proposed approach's effectiveness. As per the literature, to evaluate the quality of generated summary, ROUGE metric is most commonly used. ROUGE counts the number of overlapping units such as word-sequences, n-grams and word-pairs between automatically generated summary and human – generated summaries. It is computed as:

$$ROUGE - N = \frac{Number\ of\ overlapping\ units}{number\ of\ words\ in\ reference\ summary} \quad (3.3)$$

Table 3.2. displays ROUGE scores obtained by different biomedical summarizers, utilizing context-free language-based models and various graph based ranking algorithms. The results are shown for the top K values for each pairing of a language-based model.

Table 3.2. Comparison with the State of the Art Methods

Language-based model	**Ranking Algorithm**	**ROUGE-1**	**ROUGE-2**	***Best K***
CBOW	HITS	0.716	0.3042	**0.6**
CBOW	PPF	0.722	0.3094	**0.6**
Skip-gram	HITS	0.722	0.3118	**0.7**
Skip-gram	PPF	0.731	0.3155	**0.6**
CBOW	PageRank	0.730	0.3157	**0.7**
Skip-gram	PageRank	0.736	0.3204	**0.7**
Proposed Approach	**PageRank**	0.781	0.3341	**0.7**

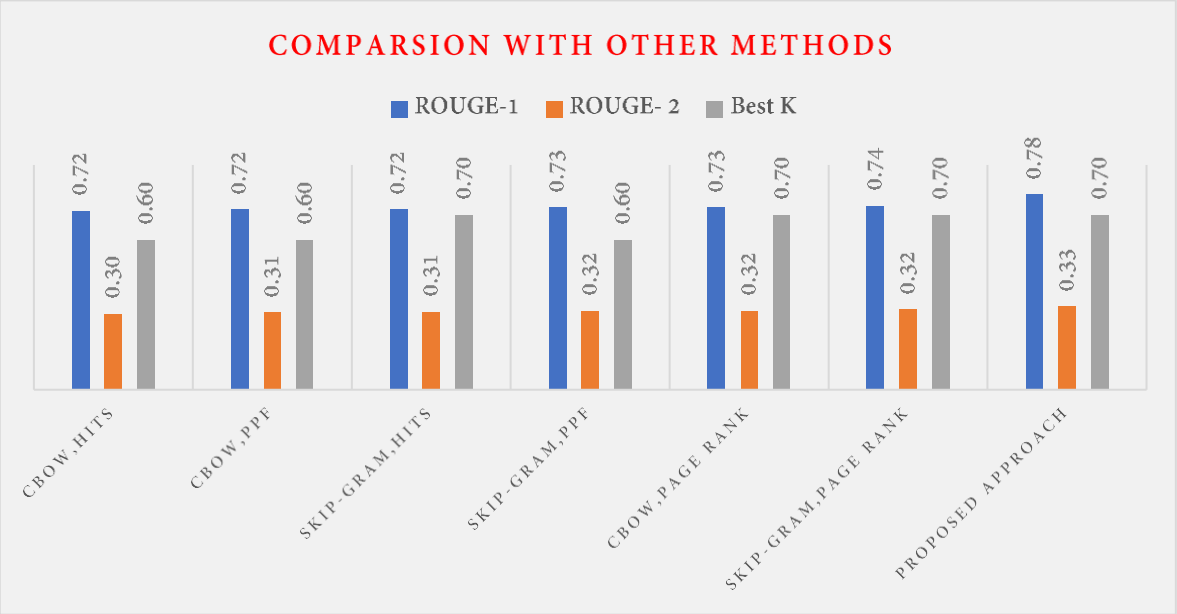


Fig 3.6. Comparison with the State of the Art Methods

In Table 3.2., the performance of the graph-based biomed summarizer is evaluated using three different language-oriented models produced by BioBERT, and via different graph ranking algorithms. Only the best K value for all possible combinations of the language model, and the ranking algorithm is reported. The results show that the performance of the summarizer varies depending on the language-based model and ranking based algorithm used. The same observation be made for Table 3.3, where a balance between the number of edges involved in the ranking process and their weights is important for achieving informative and accurate summaries. When too many or too few edges are incorporated, the algorithm may not select the most valuable and highly correlated sentences, leading to less informative summaries. In Table 3.3, the performance of the graph-based biomed summarizer is evaluated using three different language-based models produced by BioBERT, and via various graph ranking algorithms. Only the best K value for all possible combination of language, ranking based algorithm is reported. The results show that the performance of the summarizer varies depending on the language-based models, and ranking based algorithm used. In Table 3.3, where a balance between the number of edges involved in the ranking process and their weights is important for achieving informative and accurate summaries and Fig. 3.7. shows the comparative analysis.

Table 3.3. Comparison while selecting the Best K values

S. No.	Language model	*Best K*	Ranking algorithm	ROUGE
1	BioBERT(PubMed)	0.7	PageRank	**0.7418**
2	BioBERT(PubMed)	0.6	HITS	**0.7322**
3	BioBERT(PubMed)	0.5	PPF	**0.7402**
4	BioBERT(PMC)	0.6	PageRank	**0.7346**
5	BioBERT(PMC)	0.6	HITS	**0.7277**
6	BioBERT(PMC)	0.6	PPF	**0.7308**
7	**Masked Language Modeling (PubMed)**	**0.7**	**PageRank**	0.7480

When too many or too few edges are incorporated, the algorithm may not select the most important and highly related sentences, leading to less informative summaries.

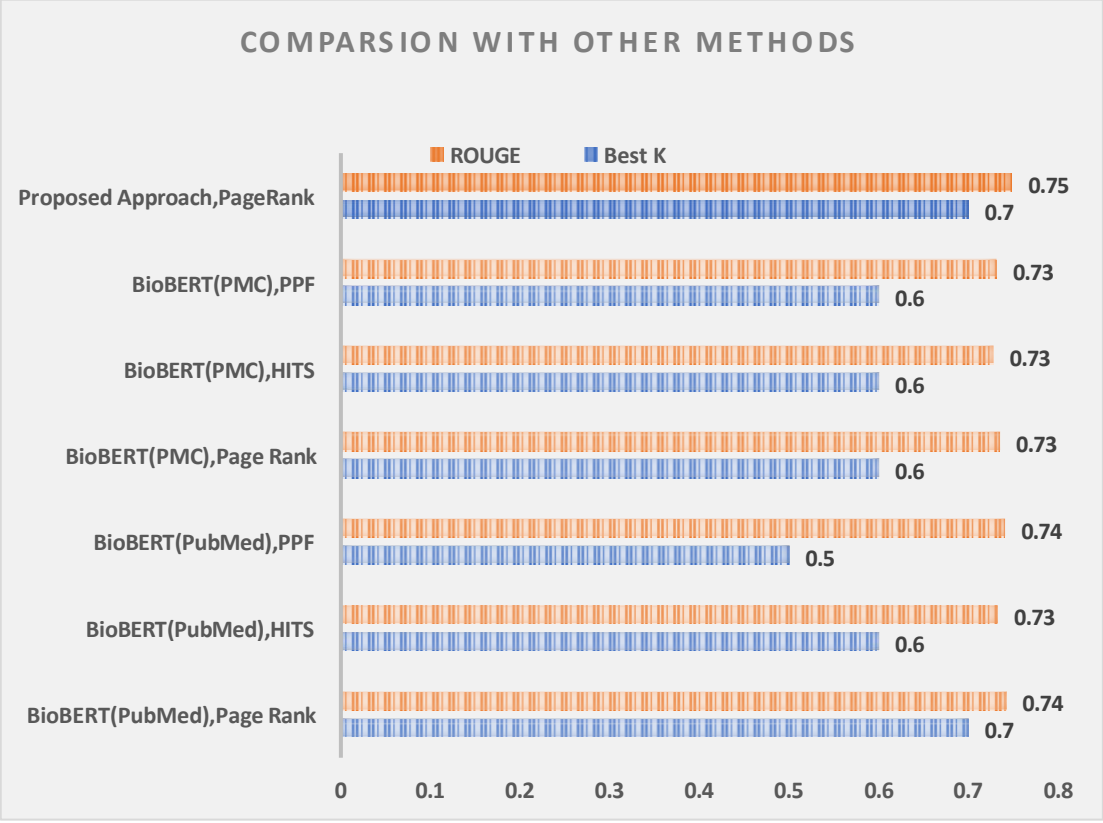


Fig 3.7. Comparison with the State of the Art Methods selecting the Best K Value

3.4. Conclusion

In this chapter, we have proposed a novel framework designed to augment contextual embeddings with specialized clinical domain expertise by integrating knowledge from the UMLS Metathesaurus during the pre-training phase of a BERT-based model. This approach aims to create 'semantically enriched' contextual representations, leveraging both the contextual learning capabilities of the BERT architecture and the domain-specific knowledge embedded in the UMLS Metathesaurus. Recent experiments, employing supervised learning techniques with word embeddings, have yielded promising outcomes when applied to the Metathesaurus context. These results validate the effectiveness of such approaches in achieving satisfactory performance in aligning specific subsets of source vocabularies within the Metathesaurus.

To identify potential word ngram candidates, conducted a comprehensive query across the entire UMLS database. Recognizing the challenges and inefficiencies associated with exact matching on such a vast database, we adopted an approximate string-matching approach using simstring. This strategy enhances the efficiency of the matching process, overcoming the difficulties associated with exact matching within the extensive UMLS database. The approach achieves a ROUGE score of 74.80% and demonstrates the potential for better interpretation of key ideas and sentences in biomedical papers. Future research in the field can focus on developing more advanced summarization models to improve the accuracy further.

CHAPTER 4

A Novel Method for Text Summarization using Extractive Summarization Using Concept-Space and Keyword Phrase

4.1. Introduction

In previous chapter, a novel framework was proposed to augment contextual embeddings with specialized clinical domain expertise by integrating knowledge from the UMLS Metathesaurus during the pre-training phase of a BERT-based model. This approach aimed to create 'semantically enriched' contextual representations, leveraging both the BERT architecture's contextual learning capabilities and the UMLS Metathesaurus' domain-specific knowledge. Recent experiments using supervised learning techniques with word embeddings yielded promising results, validating the effectiveness of this method in aligning specific subsets of source vocabularies within the Metathesaurus. In the biomedical domain, where summarization is based on word embeddings, several embedded models have been developed, leveraging recurrent neural networks, recursive networks, and convolution networks to learn the semantic representation of sentences. Despite these advancements, supervised extractive summarization in the biomedical domain faces challenges such as i) the unavailability of manually annotated medical health records for identifying concepts and their relationships. Additionally, ii) assessing the informativeness of sentences based on concepts and their relationships poses a hurdle. To overcome these limitations, unsupervised extractive summarization methods have been proposed. An unsupervised deep learning model that leverages word embeddings from BERT, named BioBERT have been proposed that effectively captures sentence context and quantifies relatedness and informativeness. Also, multi-document summarization using sentence embeddings and a centroid-based approach, considering content relevance, novelty, and sentence position have been proposed. While these methods are based on word embeddings emphasizing lexical similarity, previous researches primarily concentrated on lexical similarity between sentence concepts. Similarly, a domain-dependent graph-based approach utilizing UMLS and frequent-itemset mining for biomedical text summarization have also been proposed. However, limitations persisted, including a focus solely on linguistic similarities in word embeddings and the domain-dependency of graph-based approaches.

To address these gaps, in this work, an unsupervised approach that prioritizes semantic similarity and keyword-phrase extraction through a domain-independent approach has been

proposed. The proposed method is tailored for both single-document and multi-document (generic) summarization, emphasizing a novel and versatile solution to overcome the limitations of previous researches.

The chapter is organised as follows: various algorithms used in the proposed novel methodology are explained in section 4.2 followed by the proposed methodology in section 4.3. Section 4.4 discusses various steps involved in the implementation of the methodology followed by results that consist of a golden standard summary and automated generated summary along with various performance metrics in section 4.5. The conclusion of the proposed work is presented in section 4.6.

4.2. Algorithms Used for Biomedical Summarization

Distinct concepts that are used for single and multi-document summarization of the biomedical domain have been investigated and studied in this section. It comprises preprocessing of textual data, latent semantic analysis, concept map, and rapid automatic keyword extraction.

4.2.1. Data Pre-processing

To summarize the textual document, text pre-processing is an important part. It identifies several characters and words that serve as the fundamental units for further processing. It includes various evolution steps such as tokenization (breaking up the string into pieces of words), stop word removal (elimination of frequently used words), and stemming (conversion to base form). Tokenization breaks up the string into pieces of words and phrases called tokens. It removes punctuation and converts all uppercase characters to lowercase characters. Stop word removal are frequently used words such as 'adverbs', 'verbs', 'conjunctions' etc. are removed from the list of tokens. For example, words like 'is', 'are', 'this', 'and' etc. This reduces the noisy data and the performance of the system is improved. Stemming- it converts any words to its base form. Suffixes such as ed, ly, ing are removed from the words [146].

For illustration,

Sentence: It is a sunny day

Tokenization: ‘It’, ‘is’, ‘a’, ‘sunny’, ‘day’.

Stop word removal: ‘it’, ‘is’, ‘a’

4.2.2. Latent Semantic Analysis (LSA)

A numerical analysis approach infers profound relationships among words in vast text data. The textual data is expressed in matrix M, where each row denotes a unique word W and each column depicts a sentence S. Each cell represents the frequency of each word in a sentence denoted by: $freq(W)$. LSA uses the singular vector decomposition technique as a dimensionality reduction technique which forms semantic generalizations from textual data. Semantic similarity is a measure that computes the likeliness between two words that are similar in meaning. LSA () function is used to extract main concepts from the biomedical texts as LSA uses the semantic similarity among the words. A score between (0,1) is assigned between a pair of words. If words are similar in connotation, '1' score is accredited and if words are not similar in connotation, '0' score is accredited. For illustration, words such as 'epidural' and 'transforaminal' are interchangeably used with a semantic similarity measure of 0.99 [147], [148].

Latent Semantic Analysis is a natural language processing method that analyzes and identifies the relationships between documents and terms that are contained within them. It is a mathematical technique that uses singular value decomposition, to understand unstructured data and thus, to find hidden relationships between terms and concepts. It also closely approximates several aspects of human language learning and understanding. Therefore, Latent Semantic analysis is used as compared to other methods such as statistical similarity, vector space model, and word alignment-based model. The concept is illustrated in Fig. 4.1.

In this work, to compute semantically similar neighborhood words, $compose()$ function is used. This function is a part of LSAFun package in R language. $compose()$ returns a vector in semantic space with the same dimensionality. It specifies the 'k' words included in the vector using the $predication()$ function. The output of $compose()$ function is fed as an input to $neighbors()$ function, which further computes the semantically similar neighbor words between two given words [149]. It is expressed as in equation (5.1).

$$compose(u, v, method) = predication, m, k, tvectors = tvectors \quad (4.1)$$

Where,

u – single word 1

v – single word 2

m – number of neighborhood words to be predicated

k – score of k-neighbourhood

$vectors$ – numeric matrix of word vector

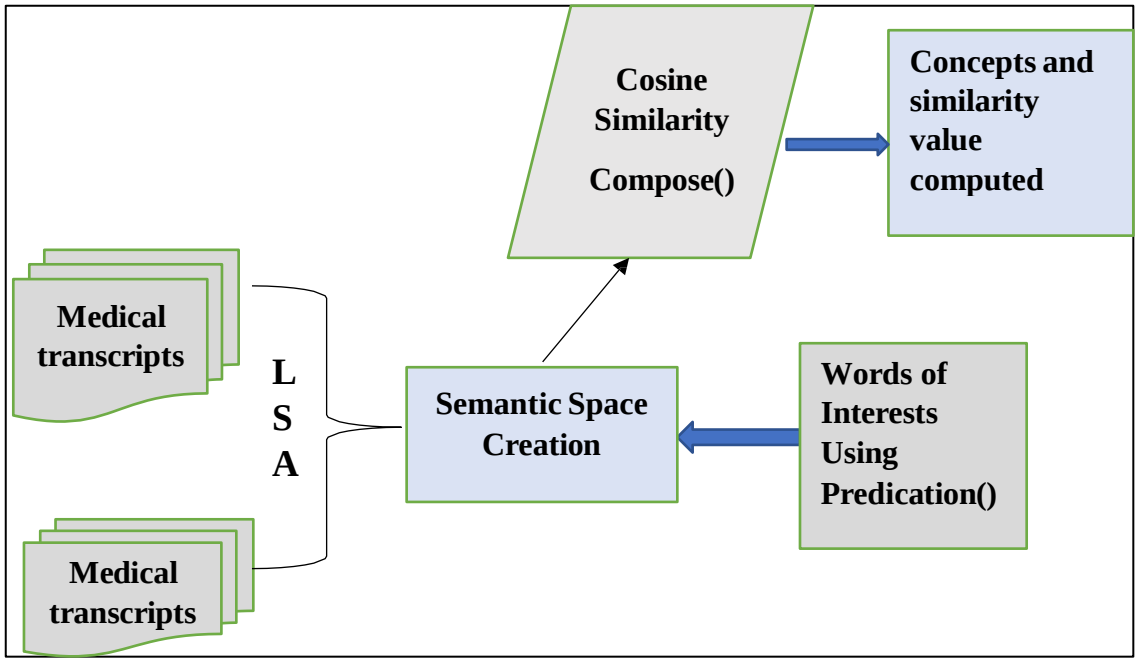


Fig. 4.1. Flow of computing Semantic similarity

4.2.3. Concept Map

A Concept Map is a visual representation of a meaningful relationship among the concepts of domain. It enables learners to focus only on the key concepts of a particular domain and organizes concepts into a structured form. Novak and Govin introduced it in year 1984. It is extensively used tool in education domain. It mainly constitutes of two things: concepts and their relationship. In a graph, G, $G \epsilon (u, v)$ where u are nodes that denotes concepts and v are edges that denotes the relationship between concepts [150] - [152]. Here in this research, concepts are the biomedical domain's neighboring words and edges represent the semantic similarity between various concepts. To map the concepts into concept map, identification of main concept is mandatory. After identification of main concept, subordinate and related concepts are identified and based on similarity values, these concepts are linked and are mapped to the concept map. For illustration, concept map is explained with an example of "Water" in Fig. 4.2. Water is made of molecules; it is used by living organisms and occur in various states. So, living things, molecules and state are semantically similar with water which can be depicted through a concept-map.

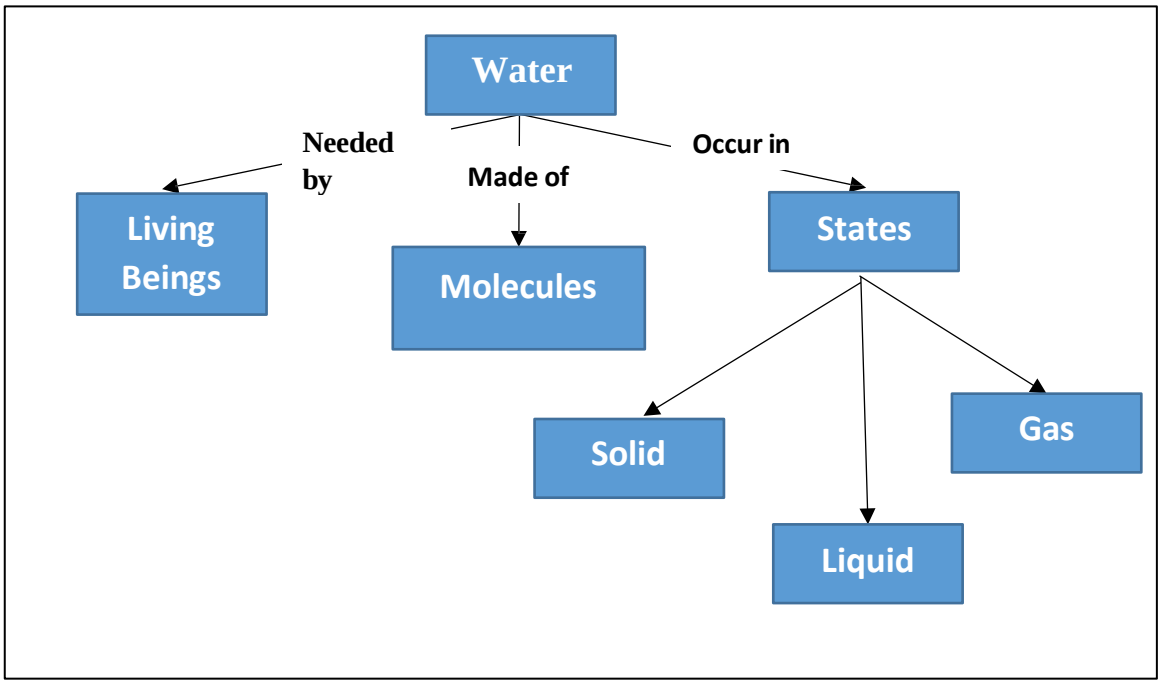


Fig. 4.2. Illustration of Concept-map with example of Water

4.2.4. Rapid Automatic Keyword Extraction (RAKE)

RAKE is the principle that extracts significant keyword phrases. It employs an archive of the concerning stop words and the phrase delineator to extract the utmost appropriate keywords extracted from the source data. It tokenizes text data along with removing stop words and phrase delimiters from the list of tokens. The remaining words in a list are called Content Words (CW). Then, a list of candidate words is created from textual data by splitting the text data at each phrase delimiter or stop word.

A co-occurrence matrix is established after creating a list of content words/terms and candidate words/terms. The matrix represents the frequency/regularity of co-occurrence of a word with another content word in sentences [153], [154]. A score for each content word is completed as follows:

i) Frequency/regularity of each content word/term is evaluated, represented by $freq(CW)$.

ii) Degree of word is computed as total number of words reflecting in postulant keyword comprising the content word/term, depicted as $\deg(CW)$.

iii) Ratio is computed as,

$$R=\frac{\deg(CW)}{freq(CW)} \tag{4.2}$$

The Complete procedure of content word, candidate word and keyword phrase extraction has been explained with an example in Table 4.1- 4.3. The example comprises of a medical transcript of neurology domain.

Table 4.1. Example of Transcript

1. Preoperative Diagnosis: Squamous cell carcinoma of right temporal bone/middle ear space.
2. Right temporal bone resection, rectus abdominis myocutaneous free flap for reconstruction of skull base defect right selective neck dissection zones 2 & 3.

Table 4.2. Score computation of Content Words

S. No.	Content Word	Degree (CW)	Frequency (CW)	Ratio (CW)
1	Squamous	3	1	3
2	Cell	3	1	3
3	Carcinoma	3	1	3
4	Right	8	3	2.4
5	Temporal	6	2	3
6	Bone	6	2	3
7	Middle	3	1	3
8	Ear	3	1	3
9	Space	3	1	3
10	Resection	1	1	1
11	Rectus	2	1	2
12	Abdominis	2	1	2
13	Myocutaneous	3	1	3
14	Reconstruction	1	1	1
15	Flap	3	1	3
16	Skull	2	1	2
17	Defect	2	1	2
18	Neck	2	1	2
19	Dissection	2	1	2
20	Zones	1	1	1
21	Selective	1	1	1

Table 4.3. List of candidate words

Candidate Words:
Squamous cell carcinoma
Right temporal bone
Middle ear space
Rectus abdominis
Skull base defect
Selective neck dissection

The computation of score of candidate word is based on scores of content word in Table 4.2. and is illustrated as follows:

$$score(squamous\ cell\ carcinoma) = score(squamous) + score(cell) + score(carcinoma) = 3 + 3 + 3 = 9 \quad (4.3)$$

4.3. Proposed Methodology

This section explains the disparate concepts invoked in this work. The framework of the proposed approach and various modules are explained in the following sub-sections which is superseded by Pseudocode of the proposed research paradigm. The approach proposed in this work focuses on three main concepts:

- Maximum content coverage achieved through information richness
- Covering diversified information from medical transcripts with maximum similarity in content
- A suitable compression ratio is achieved with respect to original transcripts.

Fig. 4.3. and Fig. 4.4. depict the framework. The methodology is explained in following subsections.

4.3.1. Corpus creation and Pre-processing

A corpus of transcribed medical reports is established for five biomedical domains: neurology, general medicine, dentistry, gynecology, and cardiovascular. A corpus of total 1040 transcribed reports is constructed[3]. The corpus comprises of short description, keywords, long transcriptions, medical-specialty and sample-name. From all these attributes, long-transcriptions are selected. These transcripts are then pre-processed using standard preprocessing steps: tokenization, stop-word removal and stemming. After pre-

[3] www.mtsamples.com

processing, Document Term Matrix (DTM) is constructed. The process is implemented in R language, which incorporates 'tm' and 'NLP' packages. After DTM is generated, sparsity is reduced using SparseM() function and further sparse matrix is constructed.

4.3.2. Feature Extraction

Features are extracted after preprocessing of text data. In feature extraction process, textual features are extracted which are categorized as word level features (keywords) and sentence level features (keyword phrases). Several approaches have been experimented with to achieve the best results in terms of information content and relevancy. The various features extracted in our proposed approach are explained in the following sub-sections.

4.3.2.1. Word level features (Keywords) – To generate multi-document (Generic) summary of transcripts, keywords are extracted and identified from sentences. In this work, keywords are the concepts which are identified through Latent Semantic Analysis (LSA). Several functions are used in R language. For example, isa() function extracts main concepts from the biomedical texts. To determine the correlation among a pair of concepts, semantic similarity is computed using compose() function that occurs in semantic space. It is completely unsupervised technique and no domain knowledge is required to train the system. It can be applied to any domain. Further, if two concepts are highly correlated, their neighborhood words are computed using neighbor() function. After identifying concepts and computing correlation among them, a concept map is constructed. Sentences comprising a number of concepts above a set threshold value, are extracted for the generic (multi-document) summarization. LSA and LsaFun packages are used in R language.

4.3.2.2. Sentence Level Feature Extraction – Sentence level features comprised of keyword phrases from the text document. To automatically keyword phrases from textual documents, Rapid Automatic Keyword Extraction (RAKE) approach is employed, which computes a score of every content word/term score. Score of each content word is computed as a ratio of degree of word ($\deg(CW)$) and ($\text{freq}(CW)$). Package rapidraker is installed and rapidrake() function is used in R language. Keyword phrases and concepts are fed to a rule engine and based on experiments; a threshold value is set. The process of content word identification and computation of score of keyword phrases is depicted as in section 4.2.4.

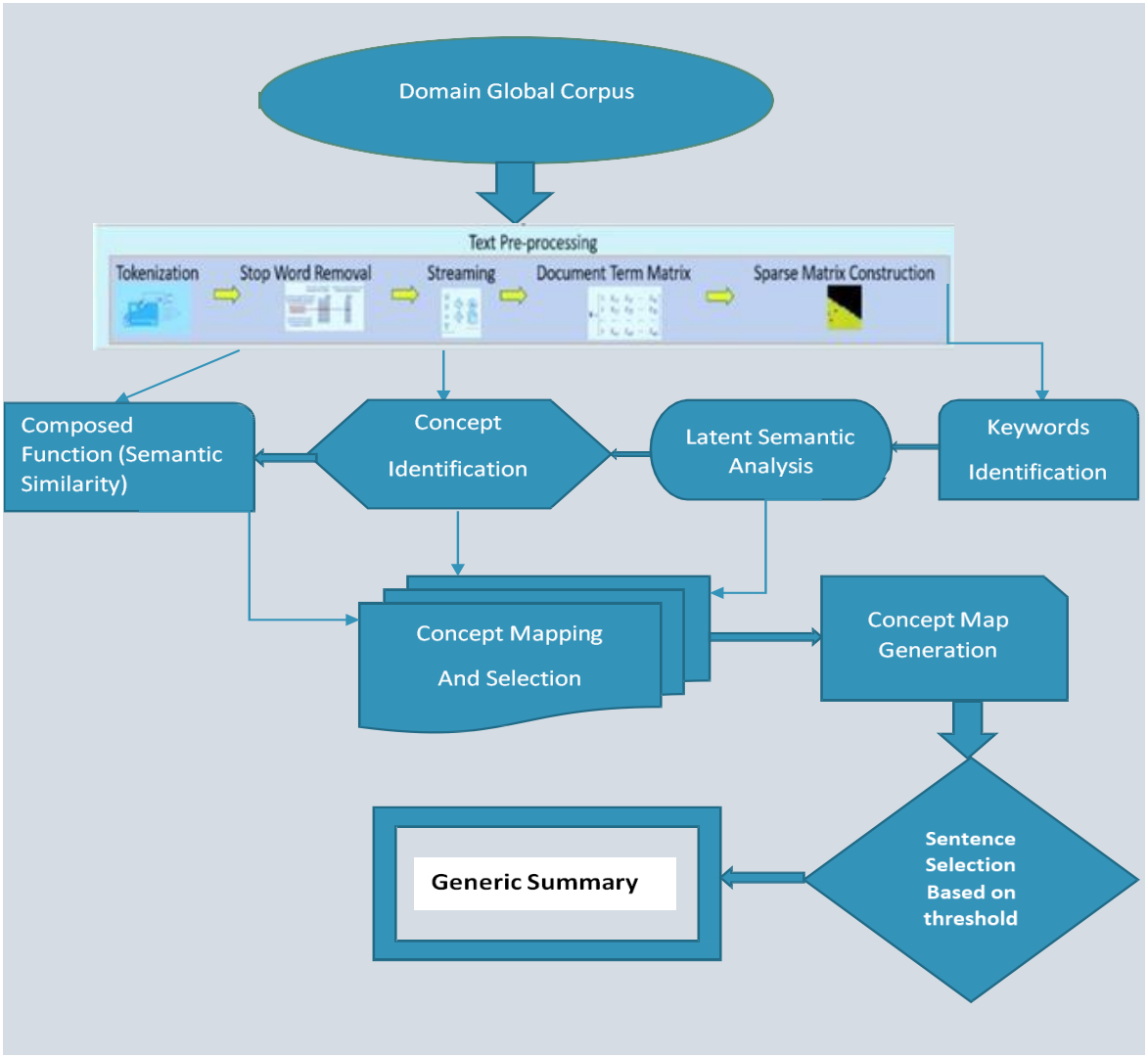


Fig. 4.3. Framework of Generation of Generic Summary

4.3.2.3. Rule Engine- After feature selection, the input is fed to Rule Engine to select sentences for Generic Summary and Single document summary. For generic summary, several concepts above a set threshold value are selected, and sentences comprising these concepts are selected for generating the Generic summary, as depicted in Fig. 4.4. A compression ratio of 10% is set as a selection criterion of selecting sentences of the whole corpus for Generic summary. For Single document summarization and to achieve a 10% compression ratio. The sentences are selected from both types of features i.e. word level features and sentence level features. For this, a threshold value of α and β, are set through experimentation procedure. To select sentences based on sentence level features (α is set to 0.6) and based on word level features (β is set to 0.4) as presented in Table 4.4. This means 60% of sentences are selected based on keyword phrase extraction, and 40% of sentences are selected based on concepts.

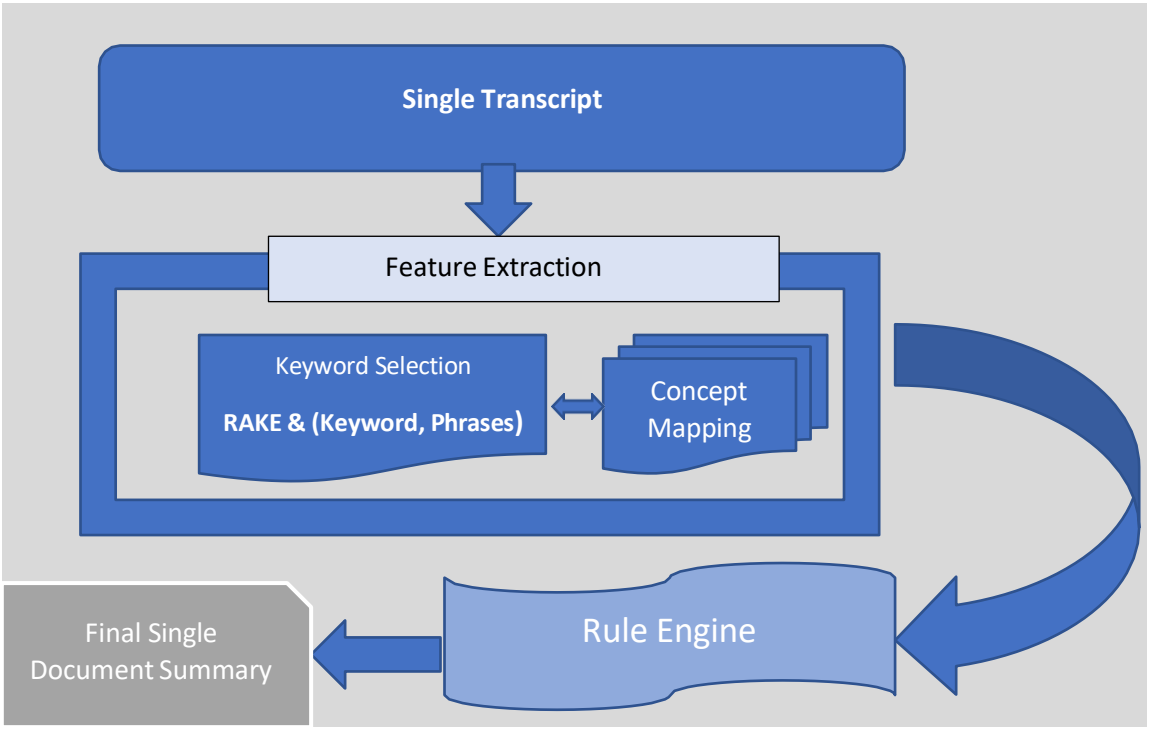


Fig. 4.4. Framework of generation of single-document Summary

Table 4.4. Parameters of sentence selection of rule engine

Feature extraction	Parameters	Threshold value
Sentence level	A	0.6
Word level	B	0.4

4.3.2.4. Pseudocode- The pseudocode for the proposed approach is shown below.

Input: G_C, Domain Global Corpus,
N-dimensional array of sentences
SENT, array of N sentences of transcripts,
Such that $G_C \supset$ SENT

Output: Gs: Generic Summary
Ts: Transcript summary

Notations:G_CP: processed global corpus

C: an array of Concepts identified
S_{ij}: Semantic similarity score among c_i and c_j
$\boldsymbol{\tau}$: Threshold value;
e_{ij}: $edgebetween\ c_i$ and c_j
$G(N, E)$: Graph of N nodes and E edges
$S(K)$: Score of keyword phrase
α, β : weighting parameter

Begin

1. G_CP← Pre-process the global corpus, G_c
2. DTM← Document Term Matrix (G_CP)
3. C ← LSA (DTM)
4. Semantic similarity between concepts is computed
5. $for\ each\ pair\ (C_i, C_j)$
6. {
7. $Compute\ S_{ij} \leftarrow compose(C_i, C_j)$
8. $for\ i\ in\ range\ i\ to\ N$
9. If
10. $Semantic\ Similarity\ (S_{ij}) \geq \tau$
11. $Then$
12. $E_{ij} <-- C_i, C_j$
13. $G(N, E) \leftarrow [N \in (C_i, C_j) \&\&\ (E_i, E_j)]$
14. }

$If \quad SENT_i \in (C_i \geq 4)$

15. $SENT_i\ is\ selected$
16. G_S←$SENT_i$
17. $Return\ G_S$

// **Automatic keyword extraction**

18. CW ← set of candidate words
19. SW←set of stop words
20. $for\ SENT_i \subseteq SENT$
21. $SENT_i$←$\sum CW_i$
22. for each $CW_i \in CW$
23. {
24. CW_i←$SENT_i \cap SW_i$
25. $Score(CW_i) = \frac{Degree(CW_i)}{Frequecy(CW_i)}$
26. $S(K) = \sum_{i=1}^{n} CW_i$
27. }
28. T_S: $\alpha * G(N, E) + \beta * S(K)$
29. $return\ T_s$
30. End

Explanation of Pseudocode

The pseudocode presented outlines a method for generating summaries from a domain-specific global corpus and a set of transcripts. The inputs include the domain global corpus G_C, an N-dimensional array of sentences, and an array of N sentences from the transcripts (SENT), where G_C encompasses SENT. The outputs are a generic summary Gs of the global corpus and a transcript summary T_S.Initially, the global corpus G_C undergoes pre-processing to form G_CP. From G_CP, a Document Term Matrix (DTM) is created. Concepts (C) are identified using Latent Semantic Analysis (LSA) on the DTM. The semantic similarity between concepts is then computed for each pair (Ci,Cj), resulting in a similarity score S_{ij}. If the similarity score S_{ij} exceeds a predefined threshold value **τ**, an edge e_{ij} is established between the concepts Ci and Cj. This process forms a graph G(N,E) with nodes representing concepts and edges representing semantic relationships between them.

The pseudocode then checks if any sentences from SENT are highly related to identified concepts. If a sentence $SENT_i$ corresponds to a concept with a similarity score above a certain level, it is selected for inclusion in the generic summary G_S.

For automatic keyword extraction, candidate words (CW) and stop words (SW) are identified within each sentence from SENT. Each candidate word's score is calculated based on its degree (number of connections) and frequency within the sentences. The keyword score S(K) is then determined by summing the scores of individual candidate words. Finally, the transcript summary T_S is generated by combining the graph-based score G(N,E) and the keyword score S(K), weighted by parameters α and β, respectively. The transcript summary T_S then returned as the final output.

4.4. Implementation

4.4.1. Data Collection

Medical data is always crucial as it contains the information regarding the human diseases and their symptoms. In earlier days, it does not get disclosed as none of human beings wants to discuss about it. Still, as time evolves, several transcripts are generated where the medical history, symptoms, and corrective measures are written to further be used by the humans, doctors, clinical experts, and researchers. MTSamples data has been for text summarization[4].

[4] www.mtsamples.com

MT sample data has 4996 real summaries of transcripts in 40 domains such as Allergy, Autopsy, Bariatrics, Cardio, Cosmetic, Neurology, Diet and Nutritious, Discharge summary, General medicine etc. To validate the research, five major transcripts have been selected. The five parameters in each sample are description, medical specialty, sample_ name, transcription, and keywords. The sample of transcripts of 5 major domains having larger samples such as Neurology samples, General medicine samples, Gynaecology, Dental, and Cardiovascular domains having 224, 260, 154, 28, and 372 transcripts respectively is being presented in Table 2.7 of chapter 2 of this thesis. A corpus of a total of 1,040 transcripts is constructed named as MT Corpus. In the earlier state-of-the-art techniques, research had been done on PubMed and BioMed articles. As per the knowledge of the authors, none of the work has been performed on medical transcripts. Therefore, to explore the research in the direction of biomedical transcripts and to reduce the time to read, comprehend, and provide diagnosis to the patients, a new corpus MTCorpus has been constructed.

MTSamples data is a dataset made by authors for evaluating the proposed approach. It consists of medical transcripts and is an open-source database of biomedical domain maintained under Kaggle Repository. This database is been used by several researchers and academicians for the clinical analysis, research and data available is authenticated and real reports of patients are posted by hiding their identity. An assessment of the newly constructed MTSample Corpus is performed to examine and analyze the efficacy of the proposed paradigm. Also, the approach is evaluated on the existing corpus of Biomed articles [155].

4.4.2. Research Questions

Some research queries have been composed to examine the efficacy of the contemplated approach on the biomedical domain.

RQ1: Does the proposed approach attain promising results on newly constructed MTCorpus?

RQ2: Does the proposed approach achieve improved results on existing Biomed articles compared to state-of-the-art approaches?

4.4.3. Evaluation Metrics

Recall-Oriented Understudy for Gisting Evaluation metric is used to assess the proposed approach's effectiveness. As per literature, to evaluate the quality of generated summary, ROUGE metric is most commonly used. ROUGE counts the number of overlapping units such as word-sequences, n-grams and word-pairs between automatically generated summary and human – generated summaries. It is computed as:

$$ROUGE - N = \frac{Number\ of\ overlapping\ units}{number\ of\ words\ in\ reference\ summary} \quad (4.4)$$

4.4.4. Process Illustration

Step1. Text pre-processing and keyword Identification.

For generation of multi-document summary, corpus is constructed for every domain. For the construction of text corpus, both the samples' descriptions and transcripts are used. Various R language packages have been used to create the corpus and for text mining process. NLP, quanteda, tm, snowballs are the common packages used for cleaning and creating the document term matrix (DTM). For data cleaning, the first text is converted in the plain-text, then all the sentences are changed into lower case followed by stemming and stop-word removal process. Some stop words of every domain are defined. DTM is constructed using Term Frequency -Inverse document frequency (Tf-Idf). After pre-processing steps, some of the keywords for the neurology domain are shown in Table 4.5.

Table 4.5. Keywords in Neurology

Epidural, transforaminal, decompression, steroid, frontal, adhesions, Brain, neuroplastic, intractable, Residual, preoperative, tumour, nerve, midline, extremity, discharged, resected.

Step 2. Latent Semantic Analysis and binding the concepts

After the dense DTM, the LSA space matrix is generated with the help of the LSA package in R. An informative and accurate latent space matrix is generated, having 877 concepts in the neurology domain. The LSA space matrix shows the most semantically similar words and their correlation with the other words. LSA space matrix has 877 concepts represented in 466-dimension places. Some of the identified concepts after the LSA space matrix are listed in Table 4.6.

Table 4.6. Concepts in neurology

Angiogram, cerebral, disease, abnormal, activity, ear, head, independent, light, positive, seen sharp, sleep, gentleman, pleasant, treated, concerning, ethology, monitoring, seizure epilepsy, past, patient, demonstrated, evidence, focal, bilaterally, chronic

To compute the association between different concepts, the *compose()* function is used from LSA Funpackage() in R language. Here, two concepts are selected, and using the prediction method, 30 most semantically similar concepts are computed from the created LSA space matrix as illustrated in Table 4.7.

Table 4.7. Association between concepts through compose()

```
comp1<-compose ("brain","lica", method="Predication",m=20,k=2,
tvectors=test_matrix_1)

neighbors(comp1, n=20,tvectors=test_matrix_1)
```

Brain and *Lisa* are two concepts in created space matrix, and m=20 is set to get the most semantically related concepts with these two words using the predication method in compose() function. Some concepts and their similarity are depicted in Table 4.8.

Table 4.8. Concepts and its similarity

Concept	Similarity	Concept	Similarity
Assessment	0.445	**tumour**	0.34
Deep	0.44	**complications**	0.29
Scan	0.38	**removed**	0.28
Subarachnoid	0.34	**Flow**	0.28
MRI	0.32	**vasculitis**	0.24
Therapy	0.29	**Lobes**	0.27

Similarly, different concepts and their association with other concepts are computed and combined to make a Concept Map.

Step 3. Concept Map Creation

A concept map for the Neurology domain is constructed using the proposed approach. A small part of ConceptNet using the proposed parameters is depicted in Fig. 4.5.

Step 4. Generating the Generic Summary for Multi-Documents

To create a generic summary for multiple documents, semantically similar concepts and associations are identified among the concepts. After constructing the concept map, most generic sentences from the corpus are selected. Sentences with 10 % threshold are selected that can vary based on domain and requirements. In Neuroscience domain, the corpus of 13000 sentences is constructed, therefore, in generic summary 130 sentences have been selected. Here, illustrated few of the sentences selected using the proposed approach.

Generated Generic Summary for multi-document

From a corpus consisting of 13,000 sentences in the neuroscience domain, a summary comprising 130 sentences has been generated. Below are some excerpts from this summary in Table 4.9

Table 4.9 Some part of summary

"Sleep study - patient with symptoms of obstructive sleep apnea with snoring. He suffered an intraventricular haemorrhage requiring shunt placement, and as a result, has developmental delay and left hemiparesis. Physical examination and radiographic findings are compatible with left shoulder pain and left upper extremity pain, due to a combination of left-sided rotator cuff tear and moderate cervical spinal stenosis. Chronic venous hypertension with painful varicosities, lower extremities, bilaterally. Massive intraventricular haemorrhage with hydrocephalus and increased intracranial pressure. Headaches, question of temporal arteritis. Bilateral temporal artery biopsies. Severe back pain and sleepiness. The patient, because of near syncopal episode and polypharmacy. Endoscopic exposure of sphenoid sinus with removal of tissue from within the sinus. The old female was referred to physical therapy following complications related to brain tumour removal. The patient with pseudotumor cerebri without papilledema, comes in because of new-onset of headaches."

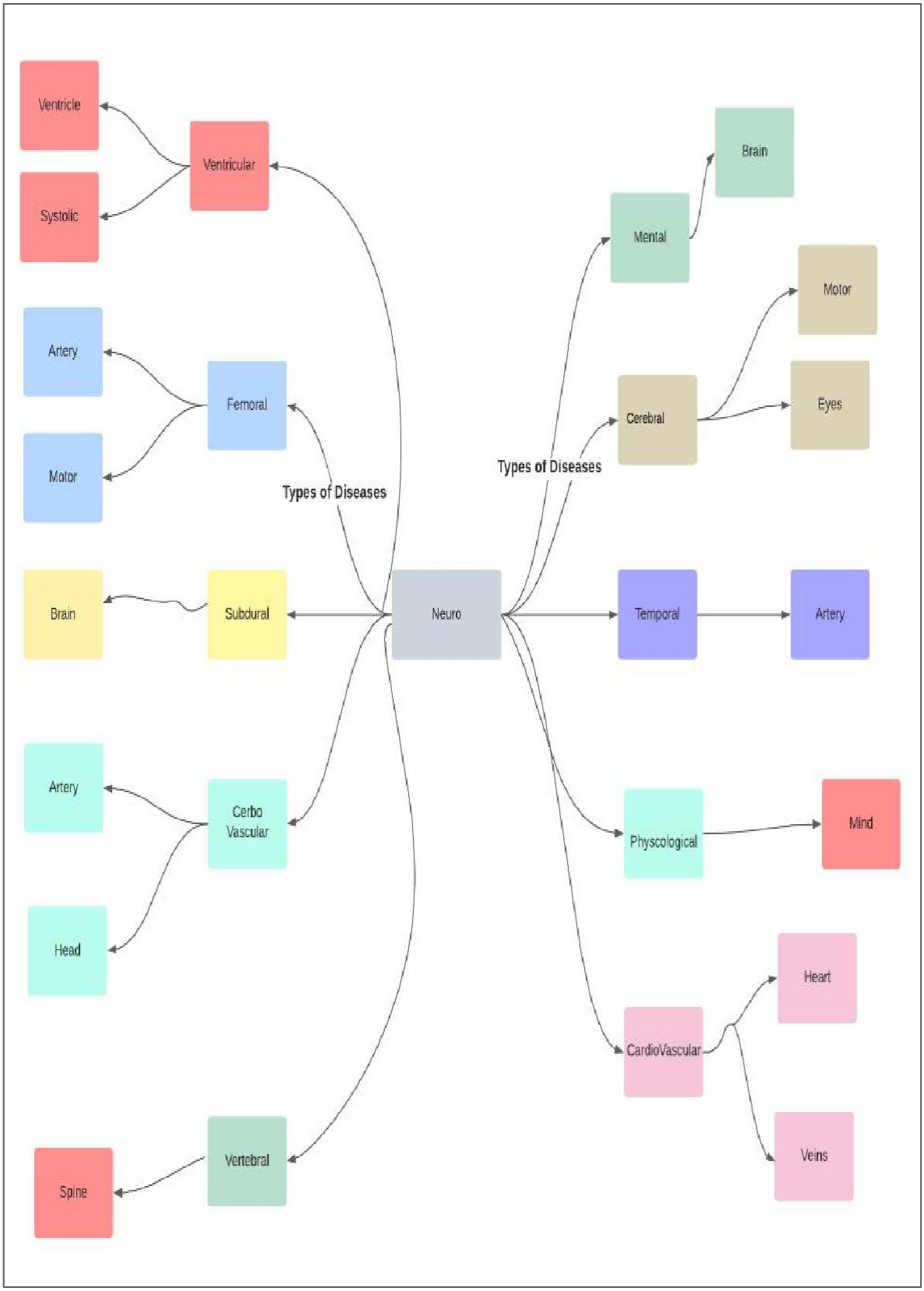


Fig. 4.5. Concept Map of Neurology Domain

Step 5. Single document summarization

A transcript is selected for single document summarization, and the proposed method is implemented as shown in Fig. 4.3. RAKE method and Concept Map are used for summary generation. Next, a transcript in neurology is depicted in Table 4.10.

Table 4.10. Sample transcript of neurology domain for single document

PREOPERATIVE DIAGNOSIS: , Chronic venous hypertension with painful varicosities, lower extremities, bilaterally., POSTOPERATIVE DIAGNOSIS: , Chronic venous hypertension with painful varicosities, lower extremities, bilaterally.,PROCEDURES,1. Greater saphenous vein stripping and stab phlebectomies requiring 10 to 20 incisions, right leg.,2. Greater saphenous vein stripping and stab phlebectomies requiring 10 to 20 incisions, left leg., PROCEDURE DETAIL: , After obtaining the informed consent, the patient was taken to the operating room where she underwent a general endotracheal anaesthesia. A time-out process was followed and antibiotics were given., Then, both legs were prepped and draped in the usual fashion with the patient was in the supine position. An incision was made in the right groin and the greater saphenous vein at its junction with the femoral vein was dissected out and all branches were ligated and divided. Then, an incision was made just below the knee where the greater saphenous vein was also found and connection to varices from the calf were seen. A third incision was made in the distal third of the right thigh in the area where there was a communication with large branch varicosities. Then, a vein stripper was passed from the right calf up to the groin and the greater saphenous vein, which was divided, was stripped without any difficultly. Several minutes of compression was used for haemostasis. Then, the exposed branch varicosities both in the lower third of the thigh and in the calf were dissected out and then many stabs were performed to do stab phlebectomies at the level of the thigh and the level of the calf as much as the position would allow us to do., Then in the left thigh, a groin incision was made and the greater saphenous vein was dissected out in the same way as was on the other side. Also, an incision was made in the level of the knee and the saphenous vein was isolated there. The saphenous vein was stripped and a several minutes of local compression was performed for haemostasis. Then, a number of stabs to perform phlebectomy were performed at the level of the calf to excise branch varicosities to the extent that the patient's position would allow us. Then, all incisions were closed in layers with Vicryl and staples., Then, the patient was placed in the prone position and the stab phlebectomies of the right thigh and calf and left thigh and calf were performed using 10 to 20 stabs in each leg. The stab phlebectomies were performed with a hook and they were very satisfactory. Haemostasis achieved with compression and then staples were applied to the skin.

For feature extraction, first RAKE method is applied. RapidRake package in R is used for finding the most important keyword phrases from the text. Setting the threshold value top sentences are selected for the summary. Table 4.11. depicts some of the most significant extracted keyword phrases of neurology domain.

Next, Concept Map is used for generating more sentences for the summary. Two main keywords from the transcripts are selected, thus, selecting the sentences which are highly similar to these keywords from the Domain Global summary. Most Semantically related words using the Concept Map are represented as:

Venous, sinuses, hypertension, bilaterally, collection, chronic, fluid, clear, extra-axial, midline, periventricular, cortical, vessel, Bony, abnormalities, process, flow, cells.

Table 4.11. Significant keyword phrases of Neurology Domain

Keywords Phrases	Score Value
Greater saphenous vein stripping	11.583333
Chronic venous hypertension	9.000000
Vein stripping	5.250000
Saphenous vein	4.583333
Lower extremities	4.000000
Stab phlebectomy	3.500000

The rule engine is applied to both concepts, and sentences are selected from the transcripts. Sentences selected from the rake and Concept Map are shown in Table 4.12. The final single document summary is generated by combining all the sentences.

Table 4.12. Generated Single document summary

Rake Sentences	Concept Map Sentences
Greater saphenous vein stripping and stab phlebectomies requiring 10 to 20 incisions, right leg chronic venous hypertension with painful varicosities, lower extremities.	Chronic venous hypertension with painful varicosities, lower extremities, bilaterally.
Then, an incision was made just below the knee where the greater saphenous vein was also found and connection to varices from the calf were seen.	An incision was made in the right groin and the greater saphenous vein at its junction with the femoral vein dissected out and all branches were ligated and divided.
Then in the left thigh, a groin incision was made and the greater saphenous vein was dissected out in the same way as was on the other side.	Several minutes of compression was used for haemostasis.

4.5. Results and Discussion

RQ1: Does the proposed approach attain promising results on newly constructed MTCorpus?

In the presented research, two innovative approaches are introduced for text summarization, targeting both generic summaries and single-document summaries. The evaluation of these approaches is conducted using biomedical text data; however, their applicability extends beyond this domain to any other. To assess our study, initially, the generic summary is compared against a golden summary. Notably, golden generic summaries are not available in these domains. Therefore, experts in the respective domains were enlisted to evaluate and approve these summaries. Three doctors, serving as experts, reviewed the generic summary in each domain, providing scores based on their knowledge. A sample of the generated generic summary is presented in Table 4.11. The scores range from 0 to 1, with 1 representing the maximum summary score for each summary. Table 4.13. and Fig. 4.6 illustrate the scores assigned by these experts to the generated generic summary in biomedical text data.

Table 4.13. Scores given by Annotators

Domain	Score_1	Score_2	Score_3
Neurology	0.76	0.81	0.78
General Medicine	0.7	0.67	0.64
Obstetrics / Gynaecology	0.78	0.73	0.71
Dentistry	0.72	0.74	0.69
Cardiovascular / Pulmonary	0.76	0.69	0.73

In the field of General Medicine, achieving a high score is challenging due to the broad scope encompassing various sub-domains within the field. Consequently, the average ROUGE score for the generic summary using the proposed method is 0.72.

To assess the single document summary, the MTSample dataset was employed for data collection and validation. The dataset includes five parameters: descriptions, medical specialty, Sample_Name, transcripts, and keywords. Evaluation of the summarization models was conducted using the Rouge method, an acronym for Recall Oriented Understudy

for Gisting Evaluation. This method compares the results of the automatic generic summary with the golden standard summary.

The transcripts represent the original medical reports containing comprehensive information about the patients' history, diagnosis, and treatment. In contrast, the golden summary is a concise overview of the pertinent information in a patient's report, created by medical professionals. The generated summary refers to the output produced by the proposed algorithm.

Rouge-1 Computes the overlap words in the golden summary and generated summary, in Rouge-1 Unigarm are considered for overlapping words. Rouge-2 compares the overlap words in golden and standard summary using the bi-gram words. Rouge-L compares the longest common subsequence between the referred and generated summary.

$$Rouge_1 = \frac{Number\ of\ Concepts\ in\ Generated\ summary}{number\ of\ Concepts\ in\ Golden\ Summary} \quad (4.5)$$

The Rouge was calculated on the MTSamples dataset initially, and subsequently, the proposed approach was assessed against the baseline approaches commonly employed in biomedical data for text summarization. For each transcript in the MTSample dataset, a golden summary was generated. The golden summary for a transcript was created by annotators by combining the Description and keywords. Table 4.15 presents the generated Golden summary along with its corresponding transcript and the summary generated in the neurology domain for single document. Due to space constrains, snapshot of multi document summary is presented in Table 4.14.

Table 4.14. Golden Summary and its Transcript of Neurology Domain for single document

Transcript	Golden Summary	Generated Summary
"Chronic venous hypertension with painful varicosities, lower extremities, bilaterally., POSTOPERATIVE DIAGNOSIS: Chronic venous hypertension with painful varicosities, lower extremities, bilaterally. PROCEDURES,1.Greater saphenous vein stripping and stab phlebectomy requiring 10 to 20 incisions, right leg.,2. Greater saphenous vein stripping and stab phlebectomy requiring 10 to 20 incisions, left leg. PROCEDURE DETAIL: After obtaining the informed consent, the patient was taken to the operating room where she	*"Chronic venous hypertension with painful varicosities, lower extremities, bilaterally. Greater saphenous vein stripping and stab phlebectomy requiring 10 to 20 incisions, bilaterally. A time-out process was followed and antibiotics were given. Then, both legs were prepped and draped in the usual fashion with the patient was in the supine*	Greater saphenous vein stripping and stab phlebectomies requiring 10 to 20 incisions, right leg chronic venous hypertension with painful varicosities, lower extremities. Chronic venous hypertension with painful varicosities,

underwent a general endotracheal anaesthesia. A time-out process was followed and antibiotics were given.,Then, both legs were prepped and draped in the usual fashion with the patient was in the supine position. An incision was made in the right groin and the greater saphenous vein at its junction with the femoral vein was dissected out and all branches were ligated and divided. Then, an incision was made just below the knee where the greater saphenous vein was also found and connection to varices from the calf were seen. A third incision was made in the distal third of the right thigh in the area where there was a communication with large branch varicosities. Then, a vein stripper was passed from the right calf up to the groin and the greater saphenous vein, which was divided, was stripped without any difficultly. Several minutes of compression was used for hemostasis. Then, the exposed branch varicosities both in the lower third of the thigh and in the calf were dissected out and then many stabs were performed to do stab phlebectomies at the level of the thigh and the level of the calf as much as the position would allow us to do. Then in the left thigh, a groin incision was made and the greater saphenous vein was dissected out in the same way as was on the other side. Also, an incision was made in the level of the knee and the saphenous vein was isolated there. The saphenous vein was stripped and a several minutes of local compression was performed for hemostasis. Then, a number of stabs to perform phlebectomy were performed at the level of the calf to excise branch varicosities to the extent that the patient's position would allow us. Then, all incisions were closed in layers with Vicryl and staples. Then, the patient was placed in the prone position and the stab phlebectomies of the right thigh and calf and left thigh and calf were performed using 10 to 20 stabs in each leg. The stab phlebectomies were performed with a hook and they were very satisfactory. Hemostasis achieved with compression and then**	*position. An incision was made in the right groin and the greater saphenous vein at its junction with the femoral vein was dissected out and all branches were ligated and divided.* *. Then, a vein stripper was passed from the right calf up to the groin and the greater saphenous vein, which was divided, was stripped without any difficultly. Several minutes of compression was used for hemostasis. Also, an incision was made in the level of the knee and the saphenous vein was isolated there. The saphenous vein was stripped and a several minutes of local compression was performed for hemostasis. Then, a number of stabs to perform phlebectomy were performed at the level of the calf to excise branch varicosities to the extent that the patient's position would allow us. Then, all incisions were closed in layers with Vicryl and staples., Hemostasis achieved with compression and then staples were applied to the skin. Then, the patient was rolled onto a stretcher where both legs were wrapped with the Kerlix, fluffs, and Ace bandages. Estimated blood loss probably was about 150 mL The patient tolerated the procedure well and was sent to recovery room in satisfactory condition*	lower extremities, bilaterally. An incision was made in the right groin and the greater saphenous vein at its junction with the femoral vein dissected out and all branches were ligated and divided. Then, an incision was made just below the knee where the greater saphenous vein was also found and connection to varices from the calf were seen. Then in the left thigh, a groin incision was made and the greater saphenous vein was dissected out in the same way as was on the other side. Several minutes of compression was used for haemostasis.

staples were applied to the skin.,Then, the patient was rolled onto a stretcher where both legs were wrapped with the Kerlix, fluffs, and Ace bandages.,Estimated blood loss probably was about 150 mL. The patient tolerated the procedure well and was sent to recovery room in satisfactory condition. The patient is to be observed, so a decision will be made whether she needs to stay overnight or be able to go home."		

ROUGE was employed as the standard method for evaluating the summarization models, despite the availability of other performance metrics such as precision and recall that could be used in the evaluation process. The primary reason for opting for ROUGE is its applicability to an unsupervised approach. Since the authors concentrated on an unsupervised approach to summarizing biomedical transcripts, the absence of a training dataset led to the utilization of ROUGE. ROUGE evaluates the model by calculating the overlap of words and does not necessitate any training data.

In contrast, using precision and recall requires the values of true positives, true negatives, false positives, and false negatives, which can only be computed with both training and test datasets. Consequently, these metrics were considered for evaluating our proposed approach, with a specific focus on the ROUGE method. Golden summaries were generated for each domain in those samples where the description was neither too short nor too long, maintaining a compression ratio of 10% for each summary. Table 4.15. provides an overview of the selected number of samples in each domain.

Table 4.15. Samples selected from each domain

Domain	Number of Samples
Neurology	34
General Medicine	45
Obstetrics / Gynaecology	46
Dentistry	45
Cardiovascular / Pulmonary	42

A total of 212 transcripts and 4563 concepts were utilized in this research within the biomedical domain. The ROUGE_1 scores for all samples across the five domains are illustrated in Fig. 4.6. The average ROUGE_1 and ROUGE_2 scores within these domains are presented in Table 4.16.

Table 4.16. Average Rouge_1 and Rouge_2 Scores

Domain	Rouge_1	Rouge_2
Neurology	0.78	0.63
General Medicine	0.776	0.62
Obstetrics / Gynaecology	0.752	0.532
Dentistry	0.76	0.591
Cardiovascular	0.741	0.578

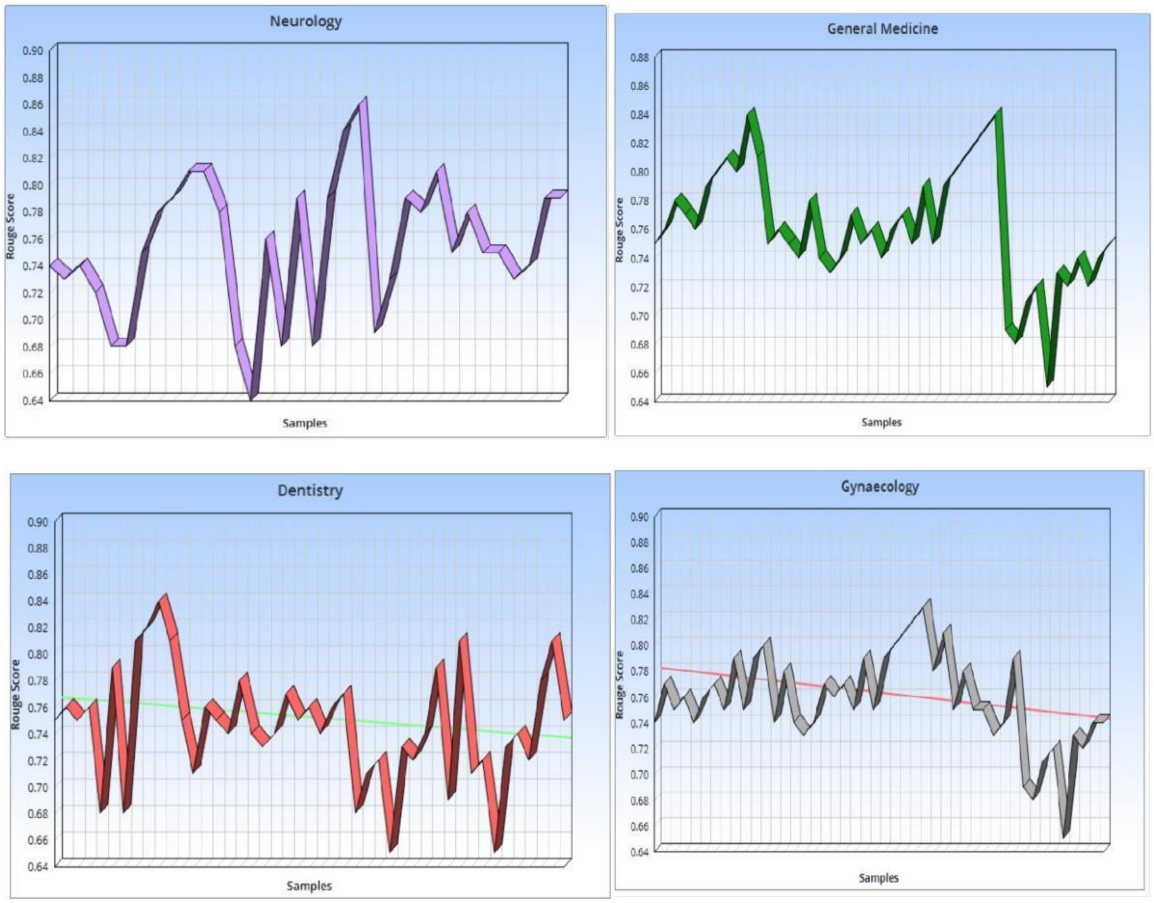


Fig. 4.6. Rouge_1 scores for different domains

RQ2: Does the proposed approach achieve improved results on existing Biomed articles compared to state-of-the-art approaches?

The results were compared with baseline approaches in biomedical text summarization to validate our proposed approach. In a recent research paper, a methodology for text summarization was proposed using a graph-based approach with the FP-Growth method. The study validated its approach using 400 biomedical research papers. Similarly, we identified research papers in our domain and compared the results using our methodology. We collected 167 research papers in our domain and applied the single transcript summarization approach. In this case, the introduction part of the research paper was considered as transcripts, and the abstract was considered as the golden summary. Table 4.17. presents a small excerpt of the biomedical research paper summary and the proposed research work summary to evaluate various baseline approaches and proposed approach. Table 4.18. and Fig. 4.7 show the comparison between the baseline approaches and the proposed approach in terms of ROUGE metrics.

Table 4.17. Baseline and generated summary (proposed method) for BioMed article for single document

Summary generated by baselines method	**"Poor adherence is a major issue and is associated with increased morbidity, mortality, and immense costs for the healthcare system. Due to demographic changes, the burden of neurological diseases increases with a crucial worsening of nonadherence. However, comprehensive data on geriatric patients with neurological disorders do not exist to date. This cross-sectional observational study aims to identify disease-specific adherence-modulating factors in neuropediatric patients. In addition, disease-specific data will be derived from medical records…".**
Introduction/ Transcript	"The treatment of chronic disorders commonly includes the long-term use of pharmacotherapy and non-pharmacological therapy. However, their full benefits are often not realized because approximately up to 50% of patients either do not take medications as prescribed or do not follow recommendations. In the geriatric population, nonadherence contributes to adverse drug events, increased length of stay and readmissions to hospitals, and a lower quality of life. However, physicians often do not routinely inquire about and are unaware of the extent of patients' nonadherence to medication. Factors contributing to nonadherence are numerous. Nonadherence is a dynamic process and maybe intentional (when the patient purposefully decides not to follow the recommended treatment) or unintentional ..... "
Proposed Method Summary	In the geriatric population, nonadherence contributes to adverse drug events, increased length of stay and readmissions to hospitals, and a lower quality of life. The treatment of chronic disorders commonly includes the long-term use of pharmacotherapy and non-pharmacological therapy. Physicians often do not routinely inquire about and are unaware of the extent of patients' nonadherence to medication. Factors contributing to

	nonadherence are numerous. This is probably due to the lack of care and routine available during the patient's stay in hospital, poor communication between different players in medical care and feedback from practitioners to the hospital has to date not been sufficiently studied in neuropediatric patients……

Table 4.18. Comparison of Proposed approach with Baseline approaches for single document

S. No.	Systems	ROUGE-1	ROUGE-2	ROUGE-W-1-2
1	Proposed approach	0.767	0.56	0.21
2	Graph and Item Set [108]	0.7648	0.3524	0.0913
3	LexRank [39]	0.7528	0.3482	0.0891
4	GraphSum [75]	0.7442	0.3361	0.0884
5	TextRank [75]	0.7394	0.3312	0.0804
6	ItemSum [75]	0.7291	0.3198	0.078
7	BioChain [127]	0.7184	0.2967	0.0764
8	SweSum [170]	0.7132	0.3118	0.075
9	TexLexAn [171]	0.6998	0.2884	0.0705
10	Lead baseline [172]	0.6922	0.2879	0.0723
11	AutoSummarize [173]	0.6891	0.2458	0.0697
12	Random baseline [174]	0.6302	0.2119	0.0653

From Table 4.18. it can be observed that our approach was comparable with baseline approaches, showing a slight improvement in the Rouge-1 score but a significant improvement in the Rouge-2 and Rouge-W-1-2 methods. Despite the biomedical data abstract not being an extractive summary of the introduction, our proposed algorithm demonstrated better performance than the baseline approaches. A ROUGE score of 0.75 was achieved for the multi-document summary. Since no prior research has focused on multi-

document summarization in this context, comparisons with previous results were not possible due to the lack of baseline approaches.

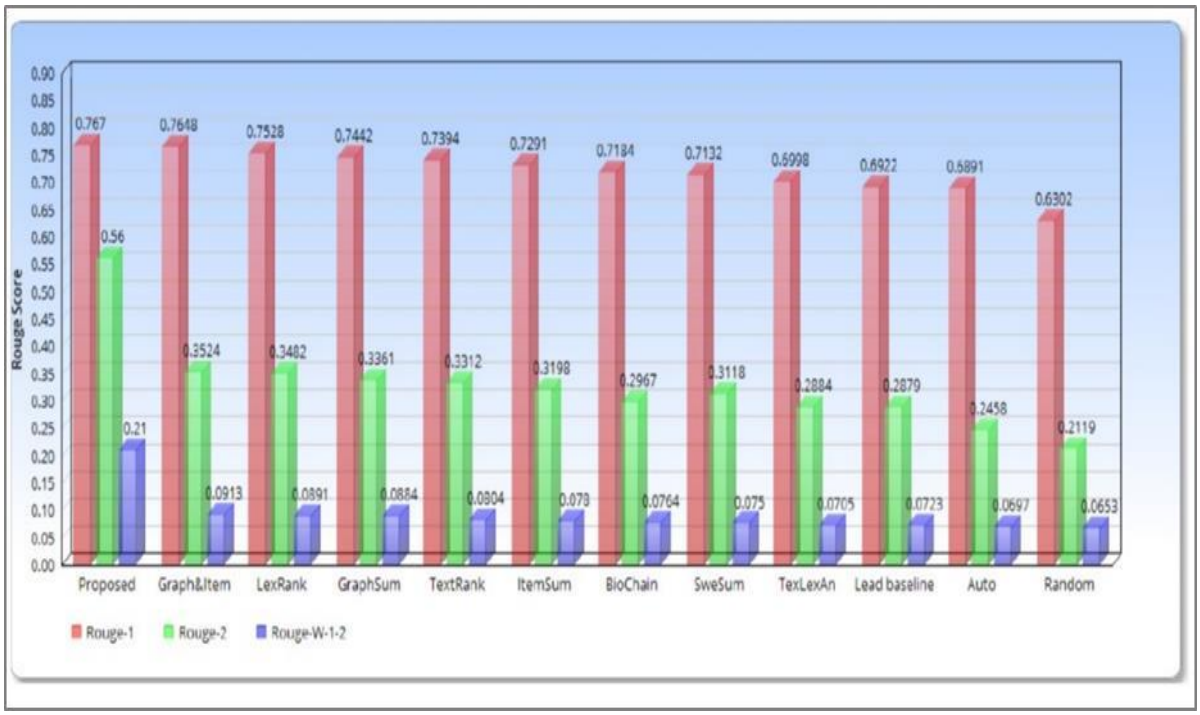


Fig. 4.7. Comparison of proposed approach with Baselines approaches

4.6. Conclusion

An unsupervised approach for single and multi-document summarization based on semantic similarity and keyword-phrase extraction was proposed. The evaluation was conducted on medical data containing information about human diseases and their symptoms. The merger of Concept Map and the RAKE method was utilized to generate a generic summary with the application of threshold values. The unsupervised approach was tested on various biomedical transcripts from neuro-science, general medicine, gastroenterology, orthopaedics, and radiology domains, encompassing 1,040 different transcripts from the MT Sample Dataset. The single-document summarization achieved an average ROUGE score of 0.77, while the generic summary achieved an average ROUGE of 0.72. The method was further validated on a previous corpus of BioMed articles, exhibiting superior results compared to state-of-the-art techniques. The multi-document summary achieved a rouge score of 0.75. The proposed unsupervised approach is poised to benefit the research community and health experts by saving considerable time and resources in computing patient summaries during diagnosis. The time and effort saved by the proposed unsupervised approach provide valuable benefits to researchers, facilitating the extraction of concise

information. This methodology can be replicated across various domains, including education, software, biomedical articles, and journal summarization. To enhance it further, applied other natural language and deep learning techniques to these medical transcripts.

CHAPTER 5

A Novel Method for Text Summarization using Deep Dense LSTM-CNN framework

5.1. Introduction

This chapter introduces a second approach to the extractive summarization of biomedical transcripts. The proposed approach is innovative and results in a more enhanced summary compared to our initial approach outlined in Chapter 4 of this thesis.

Our proposed approach is motivated by the pressing need to effectively summarize the vast volumes of fragmented data prevalent in the biomedical field, particularly in medical transcript summarization. This challenge is paramount as the information contained within health records is crucial for comprehending various diseases and their manifestations. By leveraging NLP-based deep learning algorithms and customizing them for biomedical-specific text summarization, our approach aims to deliver a concise and contextually relevant summary of biomedical literature. Incorporating techniques such as topic modelling, phrase selection, and punctuation restoration further enhances the accuracy and relevance of the produced summaries.

The integration of Dense CNN and LSTM architecture for clinical document summarization holds significant novelty for several reasons. This architecture amalgamates three distinct types of neural network layers—Convolutional Neural Networks (CNN), Dense layers, and Long Short-Term Memory Networks (LSTM)—to extract features from input data and generate the summary [156], [157]. This innovative approach remains relatively unexplored in the context of clinical document summarization. By utilizing CNNs to extract features from the input text, the model can discern important phrases and concepts within the document, subsequently utilized by the LSTM layer to produce a summary.

The inclusion of Dense layers within this architecture offers an additional degree of adaptability and flexibility, enabling the model to learn intricate relationships between the input data and the target summary. This aspect is particularly crucial for clinical document summarization, given the highly variable language present in medical records, which may necessitate more sophisticated modelling techniques for accurate summarization. Overall, the incorporation of CNN, Dense, and LSTM architecture for clinical document summarization presents a novel and innovative solution to this challenge, with the potential

to significantly enhance the accuracy and efficiency of summarization within the medical domain.

In this study, a unique approach to extractive summarization for medical transcript summarization is proposed. The main contributions and advantages include:

- A Biomed-Summarizer is introduced which is, a distinctive framework enabling intelligent and contextually aware summarization of biomedical literature.
- Biomed-Summarizer integrates a predictive quality assessment algorithm with a clinical context-aware model to identify relevant text segments within biomedical publications for inclusion in the final summary.
- A deep neural network binary classifier is developed for quality detection, aiming to distinguish scientifically valid papers from others.
- For the clinical context-aware classifier, a bidirectional long-short term memory recurrent neural network is constructed which is trained on semantically enriched features generated by a word-embedding tokenizer, enabling the identification of meaningful sentences representing textual sequences.

5.2. Research Questions

Research Question 1 How does the algorithm fare in comparison to current state-of-the-art methods for summarizing biomedical text?

Research Question 2 What contributions do the end-to-end summarization approach employing Deep Dense Long Short Term Memory Network (LSTM) and Convolutional Neural Network (CNN) models make towards enhancing the accuracy and usefulness of the summarization procedure?

To answer these questions, we have proposed a new method of text summarization and in the next part explain the various techniques used in proposed approach.

5.3. Various Techniques Used

The underlying technology used for biomedical summarization in this work involves a combination of deep learning techniques, specifically the Deep Dense LSTM-CNN architecture and ELMo (Embeddings from Language Models) Sentence Representation.

5.3.1. LSTM

Traditional Recurrent Neural Networks (RNNs) face challenges in retaining long-term dependencies due to the vanishing gradient problem, where gradients diminish exponentially as they propagate through the network during training. To address this issue, Long Short-Term Memory networks (LSTMs) were introduced.

LSTMs overcome the vanishing gradient problem by incorporating memory cells and gating mechanisms that allow them to selectively retain or forget information over time. These memory cells are equipped with three gates: forget gate, input gate, and output gate, in addition to the memory cell itself. Each gate is responsible for regulating the flow of information into and out of the memory cell, enabling LSTMs to effectively capture long-range dependencies in sequential data. The forget gate determines which information from the previous time step should be discarded, while the input gate controls which new information should be stored in the memory cell. The memory cell stores the current state of the network, and the output gate determines which information from the memory cell should be passed on to the next time step. By incorporating these gating mechanisms, LSTMs are able to learn and retain information over long sequences, making them well-suited for tasks such as natural language processing, time series prediction, and speech recognition [158]. The four gates are represented mathematically as:

If we have an old memory C_{t-1}, we can calculate the new cell memory C_t, as:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5.1}$$

Forget Gate: specifies which data will be purged from working memory

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{5.2}$$

Memory Gate: creates a fresh pool of possible memories.

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{5.3}$$

Input Gate: This gate controls the amount of new data that will be stored in the updated memory from the candidate memory.

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{5.4}$$

Output Gate: limits how much information may be retrieved from the cell's memory

$$o_t = \sigma(W_o x_t + U_o h_{t-1}) \tag{5.5}$$

5.3.2. ELMo

ELMo is a state-of-the-art deep contextualized word representation technique that captures the meaning of words in context. In this approach, ELMo word embeddings are used to represent each word in a sentence. These embeddings are pre-trained on a large corpus of text using a bi-directional LSTM (Long Short-Term Memory) model with a language modelling objective. By aggregating ELMo word embeddings for each phrase, sentence-level word embeddings are computed. ELMo consists of two LSTM networks arranged in a stacked configuration. These LSTM networks operate bidirectionally, meaning they analyze input data both forward and backward in sequence. The upper layer of this bidirectional architecture generates ELMo word vectors, also known as biLM (bidirectional Language Model), based on a two-layer bidirectional word embedding. Each layer in the biLM template is composed of two passes: a forward pass and a backward pass. During the forward pass, information about the word and its preceding terms with similar meanings is provided, while the backward pass includes information about the word and the context that follows it. The final ELMo description is obtained by combining the basic word predictions with the likely accompanying word indexes [159].

Additionally, the ELMo architecture includes several key components:

- Dropout Layer: This layer introduces randomness to the network during training by randomly disconnecting a certain percentage of connections between neurons in each layer. This helps prevent overfitting and improves the model's ability to generalize to unseen data.
- LSTM Layer: A single LSTM layer, operating bidirectionally, is essential for creating the ELMo representations.
- Bidirectional Layer: This layer allows the LSTM layers to form bidirectional models without the need for separate forward and backward layers. It combines the outputs from both directions in a single layer.
- Dense Layer: A fully connected vanilla artificial neural layer that follows the LSTM layer.
- Embedding Layer: Responsible for converting positive integers (such as word indices) into floating-point vectors.

- Conv1D Layer: Implementation of a one-dimensional convolutional neural network layer.
- MaxPooling1D Layer: Performs maximum pooling in a single dimension.

These components work together within the ELMo architecture to generate contextualized word representations that capture the meaning of words in context, making it a powerful tool for various natural language processing tasks.

In mathematical terms, an instance of convolutional neural network (CNN) operation can be represented as follows:

$$Yi = f(Xi * K + b) \tag{5.6}$$

Where, Xi is the output of the previous layer, Yi is the output of the current layer, K is the kernel for the current layer, b is the bias for the current layer, and f represents a selection of input maps. Convolving a text with multiple filters in various combinations can aid in tasks such as recognition and identification [160], [161].

The subsequent layer, known as the pooling layer, serves to reduce the number of parameters if the data are too large to be processed solely by the preceding layer. Spatial pooling, also referred to as sub-sampling or down-sampling, diminishes the number of dimensions in each map while retaining essential details. Pooling is a sampling-based technique in discretization aimed at reducing the number of dimensions in an input sequence (e.g., an image or the output matrix of a hidden layer). Features contained in sub-regions are binned, and common types of pooling include maximum pooling and minimum pooling. As its primary function is down sampling, this layer is often referred to as the subsampling layer.

For parameter estimation, we employ a supervised learning environment. In this setup, pre-labeled category targets at the segment level of the datasets serve as the supervisory signal. Possibilities based on the information gained retrospectively constitute the input data for training. N represents the total number of images used in the training process.

To achieve this, Total Squared Error (TSE) is used as a loss function. The training objective function is derived using L2 regularization:

$$J(\theta) = TSE + \lambda(\theta_{12} + \theta_{22}) \tag{5.7}$$

Where λ represents two experimental hyper-parameters, Lagrange multipliers, which are tuned using both training and validation data. Making minor adjustments to the loss function improves its effectiveness.

5.3.2.1. Deep Dense LSTM-CNN Architecture

The Deep Dense LSTM-CNN architecture combines two powerful deep learning models: LSTM and CNN. LSTM is capable of capturing long-range dependencies in sequential data, making it suitable for processing text data. CNN, on the other hand, is effective at capturing local patterns in data, making it suitable for tasks such as feature extraction.

5.3.2.2. Batch Normalization (BN)

Batch Normalization is a technique used to improve the training stability and speed of deep neural networks. It normalizes the activations of each layer in the network by adjusting and scaling the outputs, which helps in mitigating the vanishing gradient problem and enables faster convergence during training.

In summary, the process of biomedical summarization begins by computing sentence-level word embeddings using ELMo word embeddings. These embeddings are then fed into the Deep Dense LSTM-CNN architecture, along with the Batch Normalization technique, to learn text representations for summarization. This approach leverages the strengths of deep learning models and contextualized word embeddings to generate accurate and contextually relevant summaries of biomedical literature.

5.4. Proposed Methodology

High-quality representations in capturing complex nuances of word usage and their variations across linguistic contexts is of utmost importance. In this work, a novel form of deep contextualized word representation is introduced that effectively addresses these challenges by assigning each token a representation derived from the entire input phrase. This method utilizes ELMo (Embeddings from Language Models) abstractions, constructed from a bi-directional LSTM trained with a language modeling objective on a large text corpus. This approach can seamlessly integrate into existing models and has demonstrated enhancements to the current state-of-the-art in various language comprehension tasks.

Unlike traditional word embeddings that lack context awareness, ELMo embeddings capture the polysemy of words and offer a more nuanced understanding of language [162]. Fig. 5.1.

illustrates a memory network for long-term and short-term storage, providing insight into the components of LSTM. The gates, consisting of artificial neural networks with specific activation functions, convey related information.

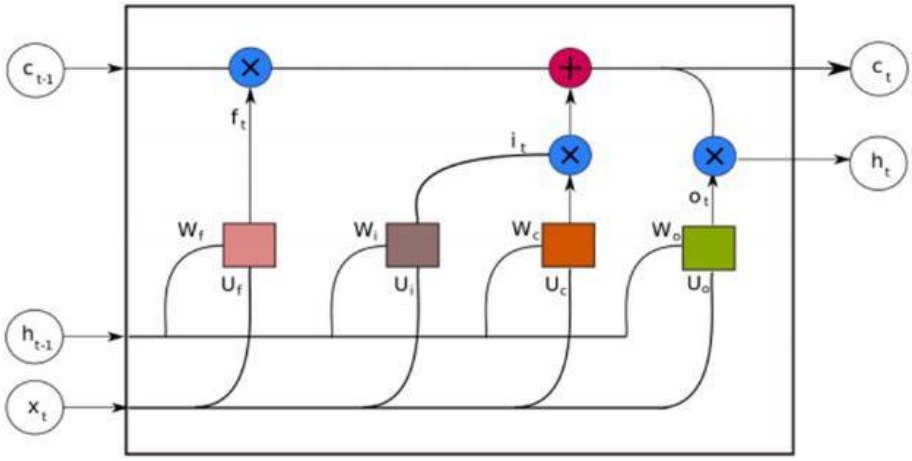


Fig. 5.1. Memory Networks for Long-Term and Short-Term Storage

In this work, the Deep Dense LSTM-CNN and ELMo Sentence Representation are introduced. The process begins with computing sentence-level word embeddings by aggregating ELMo word embeddings for each phrase and representing the text as a sequence of such embeddings. Text representations for summarization are then learned using Deep Dense LSTM-CNN and Batch Normalization (BN) techniques. The framework of the proposed approach is depicted in Fig. 5.2.

The proposed approach is explained further.

Preprocessing Biomedical Text: In this, biomedical text data is pre-processed which include tasks such as tokenization, removing stop words, and stemming or lemmatization to standardize the text.

Word Embeddings with ELMo - ELMo word embeddings are used to represent each word in the biomedical text. ELMo captures the contextual meaning of words in sentences, providing rich embeddings that account for the surrounding context.

Computing Sentence-Level Word Embeddings - For each sentence, the ELMo word embeddings of individual words are aggregated to compute a single vector representation for the entire sentence. This results in sentence-level word embeddings that capture the semantic meaning of each phrase.

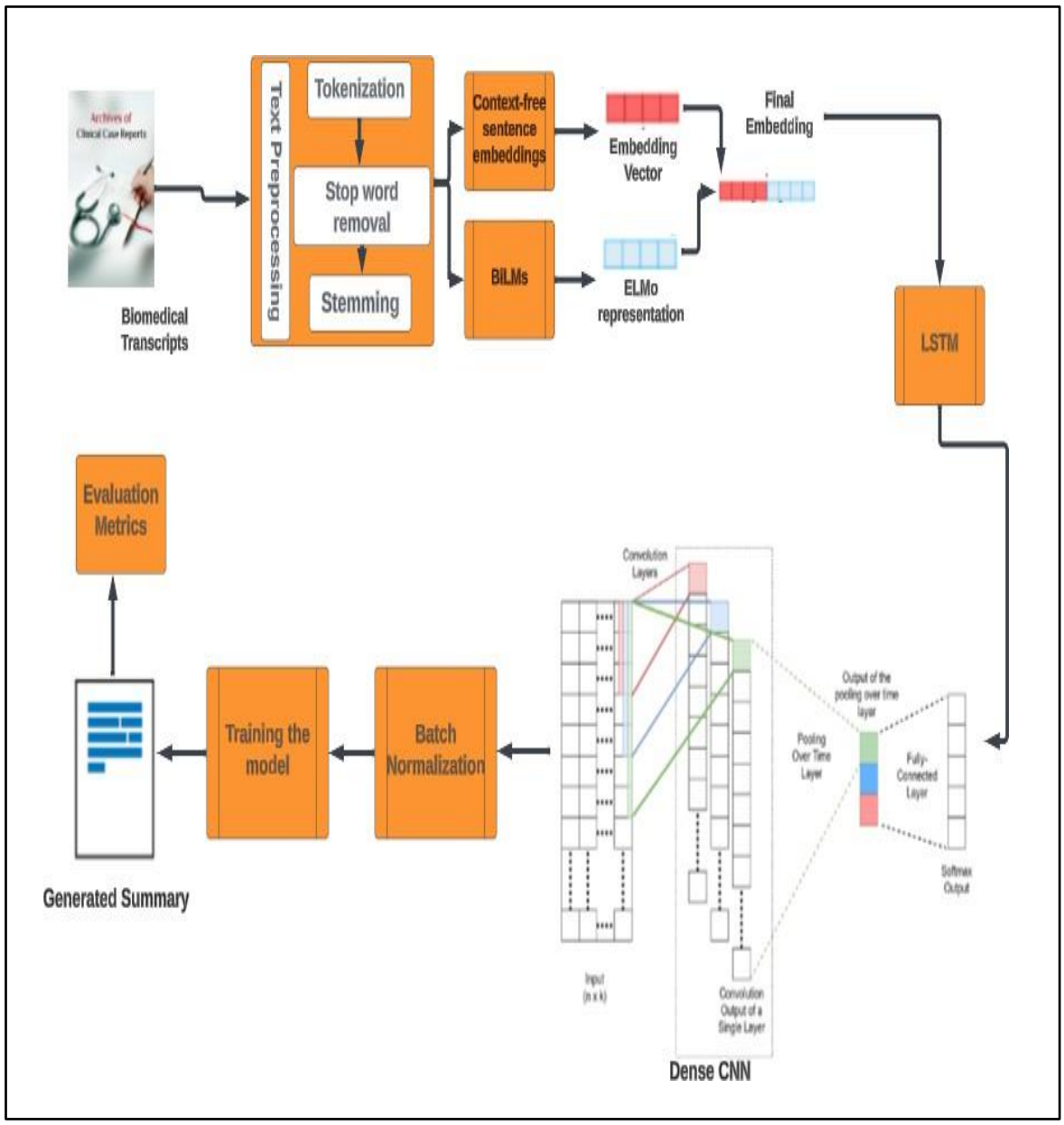


Fig. 5.2. Framework of the proposed approach

The algorithm for developing this design is presented as DDCNN algorithm, utilizing advanced domain-specific deep learning frameworks such as Keras. The algorithm is explained as follows:

- **Initializing Model:** Establishes a structured framework (Sequential) for building layers sequentially, essential for organizing neural network architecture.

5.4.1. Algorithm of the proposed approach

Proposed DDCNN Algorithm

Input: *Biomedical Transcripts*

Output: *Summary of transcripts*

Initialize a Sequential model (Model)

For feature extraction and sequence processing
Add an Embedding layer to convert words into numerical vectors

$$\boldsymbol{Embedding}(\boldsymbol{input_{dim}}, \boldsymbol{output_{dim}}, \boldsymbol{input_{length}})$$

Add 1D Convolutional layer to capture local patterns

Conv1D(filters, kernel_size, activation='relu')

Apply MaxPooling to reduce dimensionality

$$\boldsymbol{MaxPooling1D}(\boldsymbol{pool\_size})$$

Integrate an LSTM layer for sequential learning

$$\boldsymbol{LSTM}(\boldsymbol{hidden_{size}}, \boldsymbol{dropout}, \boldsymbol{recurrent_{dropout}})$$

Use a Dense layer for final output

$\boldsymbol{Dense}(\boldsymbol{output\_size})$ with softmax activation
Compile the model using categorical cross − entropy loss

$$\boldsymbol{compile}(\boldsymbol{loss} = '\boldsymbol{categorical\_crossentropy}', \boldsymbol{optimizer}, \boldsymbol{metrics} = ['\boldsymbol{accuracy}'])$$

Train the model using training data $(\boldsymbol{x\_train}, \boldsymbol{y\_train})$

$$\boldsymbol{fit}(\boldsymbol{x\_train}, \boldsymbol{y\_train}, \boldsymbol{epochs}, \boldsymbol{batch\_size}, \boldsymbol{validation\_data})$$

Initialize an empty list $\boldsymbol{summaries}$ to store generated summaries

For each transcript in test data:

$$\boldsymbol{pad\_sequences}(\boldsymbol{transcript}, \boldsymbol{maxlen})$$

$$\boldsymbol{Model.predict}(\boldsymbol{padded\_transcript})$$

Append the predicted summary to $\boldsymbol{summaries}$

Evaluate the generated summaries against reference summaries using

metrics like ROUGE or F1 score

- **Generating Summaries:** Applies the trained model to predict summaries for unseen test transcripts, translating learned patterns into actionable outputs.

- **Adding Layers:** Each layer (Embedding, Conv1D, MaxPooling, LSTM, Dense) serves a specific role in processing text data, ensuring the model captures relevant features and sequences effectively.

- **Compiling Model:** Specifies training configurations (loss, optimizer, metrics) crucial for optimizing model performance during training.

- **Training Model:** Fits the model to training data (x_train, y_train) to learn relationships between input transcripts and their summaries, crucial for model adaptation and learning.

- **Evaluating Model:** Assesses the quality of generated summaries using established metrics, providing quantitative insights into model effectiveness and summarization accuracy.

This algorithmic outline provides a structured approach to implementing a text summarization model for biomedical transcripts, ensuring clarity and coherence throughout the summarization process.

5.5. Implementation and Results

Both extractive and abstractive techniques for summarization focus on semantic qualities and connections between information components. The neural network model is well-suited for text processing due to its capability to handle sequences of varying lengths, making it widely utilized in the industry. RNNs, particularly the Bi-LSTM model, are commonly employed for multiclass text categorization. Despite being a widely used summarization model, it considers long-term text dependencies, distinguishing it from others in the field.

5.5.1. Datasets Used

The authors utilized MTSamples data to summarize texts, which encompasses forty diverse medical disciplines, including but not limited to allergies, autopsy, cardiology, and diet and nutrition. The study is based on five essential transcripts to substantiate their claims. Medium and small samples were obtained to cover a wide range of sample sizes, with 372 samples for cardiology and 28 for dentistry. However, the methodology employed for obtaining the five different samples proved applicable across all disciplines. Each sample was categorized based on five distinct criteria: description, medical specialty, sample name,

translation, and phrases. A subset of transcripts from five major domains with larger sample sizes—neurological, general medical, gynaecological, dental, and cardiovascular—consisting of 224, 260, 154, 28, and 372 transcripts, respectively, is presented in Table 2.7 of chapter 2 of this thesis. The MT Corpus comprises a total of 1,040 transcripts. Previously, innovative methodologies were employed to analyse published articles in PubMed and Biomed, leading to the creation of a new corpus known as MT Corpus, aimed at expediting the process of scanning, interpreting, and diagnosing patients.

To evaluate the effectiveness of their proposed approach, the researchers developed a dataset called MTSamples data, managed by the medical transcriptions collection of the Kaggle Repository, an open-source biomedical database. This database has been widely recognized for providing accurate patient records while maintaining anonymity, making it a valuable resource for clinical research and studies.

5.5.2. Steps for Summary Generation

A sample transcript was considered for summary generation as depicted in the Fig.5.3. The transcript was assigned to three different annotators to generate a golden standard summary as depicted in the Fig.5.4. Finally, the proposed model DDCNN was applied to generate the final summary as depicted in the Fig.5.5. The performance of the proposed approach and its comparison with state-of-the-art approaches are presented in section 5.5.3.

After obtaining the informed consent, the patient was taken to the operating room where she underwent a general endotracheal anaesthesia. A time-out process was followed and antibiotics were given. Then, both legs were prepped and draped in the usual fashion with the patient was in the supine position. An incision was made in the right groin and the greater saphenous vein at its junction with the femoral vein was dissected out and all branches were ligated and divided. Then, an incision was made just below the knee where the greater saphenous vein was also found and connection to varices from the calf were seen. A third incision was made in the distal third of the right thigh in the area where there was a where there was a communication with large branch varicosities. Then, a vein stripper was passed from the right calf up to the groin and the greater saphenous vein, which was divided, was stripped without any difficultly. Several minutes of compression was used for hemostasis. Then, the exposed branch varicosities both in the lower third of the thigh and in the calf were dissected out and then many stabs were performed to do stab phlebectomies at the level of the thigh and the level of the calf as much as the position would allow us to do. Then in the left thigh, a groin incision was made and the greater saphenous vein was dissected out in the same way as was on the other side. Also, an incision was made in the level of the knee and the saphenous vein was isolated there. The saphenous vein was stripped and a several minutes of local compression was performed for hemostasis. Then, a number of stabs to perform phlebectomy were performed at the level of the calf to excise branch varicosities to the extent that the patient's position would allow us. Then, all incisions were closed in layers with Vicryl and staples. Then, the patient was placed in the prone position and the stab phlebectomies of the right thigh and calf and left thigh and calf were performed using 10 to 20 stabs in each leg. The stab phlebectomies were performed with a hook and they were very satisfactory. Hemostasis achieved with compression and then staples were applied to the skin. Then, the patient was rolled onto a stretcher where both legs were wrapped with the Kerlix, fluffs, and Ace bandages. Estimated blood loss probably was about 150 mL. The patient tolerated the procedure well and was sent to recovery room in satisfactory condition. The patient is to be observed, so a decision will be made whether she needs to stay overnight or be able to go home.

Fig. 5.3. Sample transcript for Summarization

Chronic venous hypertension with painful varicosities, lower extremities, bilaterally. Greater saphenous vein stripping and stab phlebectomy requiring 10 to 20 incisions, bilaterally. A time-out process was followed and antibiotics were given. Then, both legs were prepped and draped in the usual fashion with the patient was in the supine position. An incision was made in the right groin and the greater saphenous vein at its junction with the femoral vein was dissected out and all branches were ligated and divided.

Then, a vein stripper was passed from the right calf up to the groin and the greater saphenous vein, which was divided, was stripped without any difficultly. Several minutes of compression was used for hemostasis. Also, an incision was made in the level of the knee and the saphenous vein was isolated there. The saphenous vein was stripped and a several minutes of local compression was performed for hemostasis. Then, a number of stabs to perform phlebectomy were performed at the level of the calf to excise branch varicosities to the extent that the patient's position would allow us. Then, all incisions were closed in layers with Vicryl and staples., Hemostasis achieved with compression and then staples were applied to the skin. Then, the patient was rolled onto a stretcher where both legs were wrapped with the Kerlix, fluffs, and Ace bandages. Estimated blood loss probably was about 150 mL The patient tolerated the procedure well and was sent to recovery room in satisfactory condition.

Fig. 5.4. Golden Summary

Bilateral chronic venous hypertension and painful varicosities were addressed with greater saphenous vein stripping and stab phlebectomy involving 10 to 20 incisions. The procedure began with a comprehensive time-out process and administration of antibiotics. Both legs were prepped and draped in the supine position before making an incision in the right groin. The greater saphenous vein at its junction with the femoral vein was dissected, ligated, and divided. Closure involved layering incisions with Vicryl and staples, ensuring hemostasis and successful completion. Post-procedure, the patient, with an estimated blood loss of around 150 mL, was sent to the recovery room in satisfactory condition. Hemostasis was reaffirmed through compression, and staples were applied to the skin, contributing to the overall success of the procedure.

Fig. 5.5. Summary generated by DDCNN

5.5.3. Results

Fig. 5.6 illustrates the performance of the proposed DDCNN during training over epochs. Initially, the model's performance shows improvement over time until it eventually stabilizes. It is evident that the accuracy remains consistently above 99% and remains stable for the majority of the training duration.

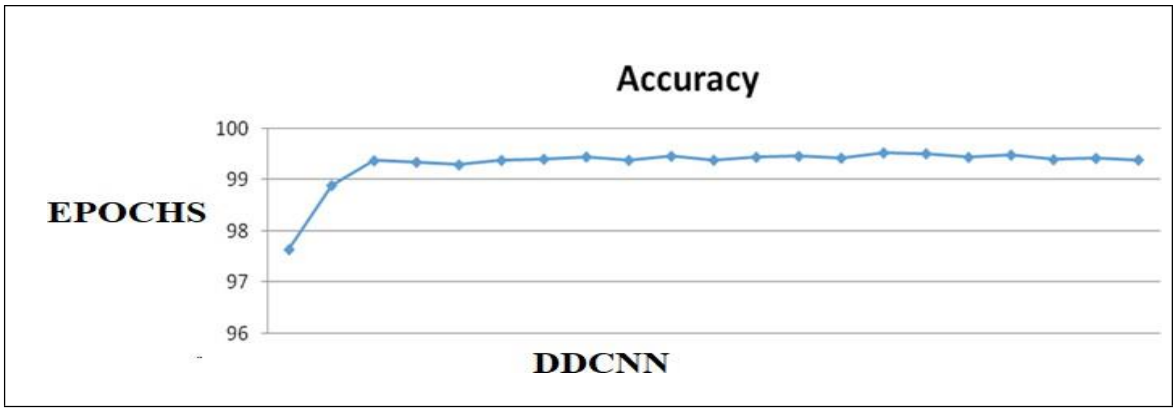


Fig. 5.6. Training Accuracy of DDCNN vs epochs

Further, Fig. 5.7 illustrates the training errors of the DDCNN across epochs. The training error is minimal throughout the entire training process, decreasing from 0.25% to 0.05% and stabilizing at 0.05% after the third epoch. This error remains constant for the majority of the subsequent epochs after the third epoch.

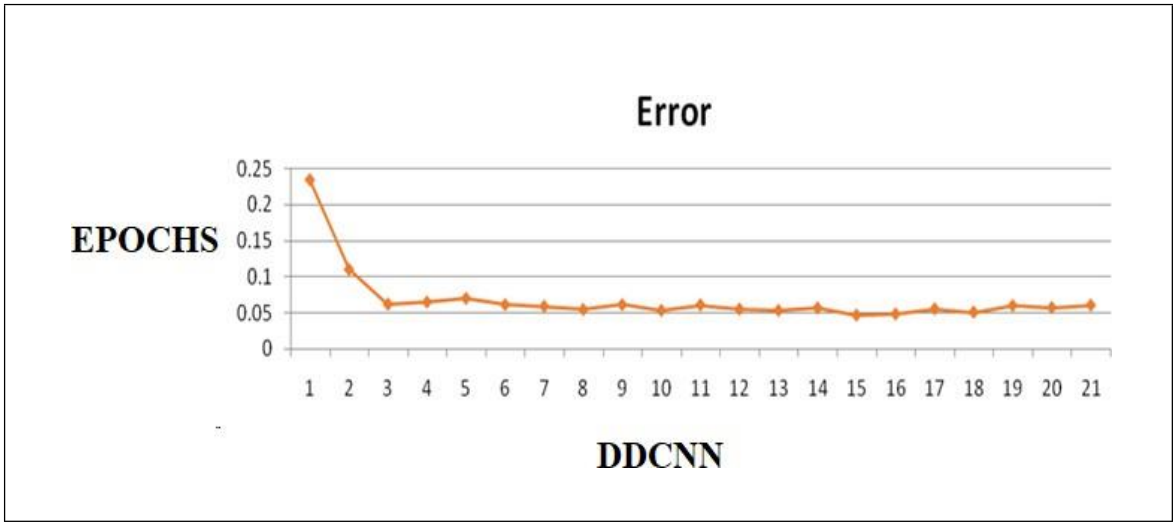


Fig. 5.7. Training Error of DDCNN vs epochs

In Fig. 5.8, the Rouge score of the proposed framework is compared to other state-of-the-art approaches. The proposed model achieves a score of 93.5%. In comparison, LSTM [27] achieves 89%, RNN [8] achieves 86.5%, and BioBERTSum [10] achieves 88.5%.

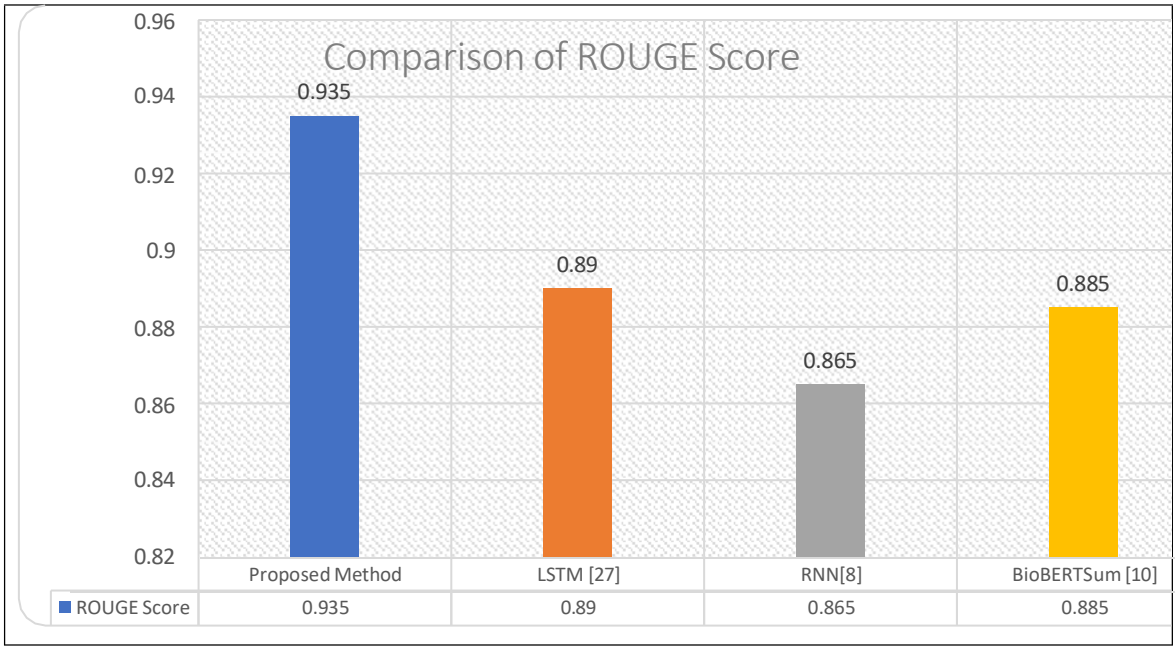


Fig. 5.8. Comparison of Rouge score of proposed models with state-of-the-art approaches

5.6. Conclusion

In this chapter, two distinct approaches to the challenge of summarizing clinical records were explored. Document summarization in this context could be challenging due to various factors: the linguistic preferences of physicians, the presence of succinct yet information-dense phrases alongside longer ones, abbreviations, misspellings, and more. An end-to-end summarization strategy, comprising Deep Dense Long Short Term Memory Network (LSTM) along with Convolutional Neural Network (CNN), was suggested for autonomously generating medical reports using biomedical transcripts. When trained on linguistically enriched features, modern deep neural network models can achieve remarkable accuracy for an Automatic Text Summarization (ATS) task compared to previous methods. Extensive testing, examination, and comparisons have indicated the effectiveness of this summarizer for medical transcript summarization. The proposed approach attained an average ROUGE score of 93.5% for single-document summarization. Additionally, comparing new techniques to previous ones demonstrates the utility and accuracy of novel strategies. The results indicate that models trained on general language can yield comparable results on a biomedical test set, with one model even outperforming the general language test set. The assessment findings highlight that the suggested Biomed-Summarizer framework significantly outperforms previous techniques.

There is potential for exploring the use of transfer learning techniques to enhance the performance of the summarization model. This could involve pre-training the model on extensive clinical data or on related tasks such as entity recognition or question answering. Furthermore, evaluating the performance of the summarization model on more diverse datasets, including records from various medical specialties or from different countries with diverse linguistic and cultural backgrounds, would offer insights into the model's generalizability and help identify areas for further improvement.

CHAPTER 6

Evaluation and Validation

6.1. Introduction

The abundance of information and sustained research focus on diverse health conditions has led to a steady rise in the volume of medical articles over the years. To stay abreast of the swift advancements in the medical field, practitioners and researchers must swiftly extract pertinent information from medical articles to advance their studies and enhance patient outcomes. Recent progress in artificial intelligence has made this task achievable through the development of Automatic Text Summarization (ATS). ATS, a main part in Natural Language Processing (NLP) research, aims to automatically generate concise summaries that highlight the most crucial information from lengthy source documents or document collections for the biomedical text data.

While automatic text summarization has advanced, there is a crucial need to develop mechanisms for the automatic assessment of the worth of generated summaries. This allows for comparisons and enhancements of different Automatic Text Summarization (ATS) systems. Human evaluation is widely regarded as the benchmark for assessing summaries, but it demands significant resources in terms of time, money, and effort. To address this challenge, the scientific community has developed various extrinsic and intrinsic methods for automatically evaluating summaries. Extrinsic evaluation involves assessing summaries in relation to another task, such as answer extraction, while intrinsic evaluation involves assessing summaries independently of any specific context, with or without human intervention. Both extrinsic and intrinsic methods aim to evaluate various characteristics in the summaries, including linguistic quality, content, coherence, and coverage.

In this chapter, focused on intrinsic methods for evaluating the quality of extractive summaries in the general domain, with some reliance on human intervention. When developing an automatic summarization or evaluation system, several considerations must be taken into account. Firstly, the source of evaluation texts can be digital documents obtained from the web, downloaded from public benchmarks, or transcribed automatically from audio sources. This raises ethical concerns regarding the use of these texts while ensuring the privacy of relevant parties is not violated. Secondly, the nature of evaluation texts varies across domains such as medicine, news, sports, literature, science, and dialogues.

As a result, the selection of an appropriate automatic system depends on factors like the text's nature, structure, and length. For example, the maximum input sequence length and the maximum length of generated summaries may differ from one system to another.

Our focus lies in summarizing extensive medical transcripts shown in chapter 3, and to address this, we employ different methodologies to generate extractive summaries from lengthy input text data. Numerous cutting-edge deep architectures, including BERT [163], T5 [164], and PEGASUS [165], have demonstrated adaptability for various NLP tasks, including text summarization. However, these models encounter challenges stemming from the intricacies of the summarization task:

Length of Input Text: Existing neural-network-based approaches face limitations in reading the entire source text due to memory explosion issues. The maximum input length documented in the literature is typically constrained to 2000 tokens, as seen in LSTM-based approaches [74], [165].

Redundant Information: An inherent drawback of existing summarization approaches is the prevalence of redundant information in generated summaries. Addressing this challenge necessitates the implementation of efficient techniques to mitigate repeated n-grams during the decoding process.

Choice of Output Summary: During the decoding stage, predicting the next word is influenced by what has already been generated. Multiple methods exist for predicting the next word, including greedy search (selecting the word with the highest probability each time) and more sophisticated algorithms like beam search (exploring a tree of possible summaries).

Computational Requirements: Unlike many NLP applications, text summarization is a demanding task requiring deep networks for effective learning. State-of-the-art results often rely on pre-trained models, such as the PEGASUS system from Google, pre-trained on a massive dataset of 1.5 billion articles (3.8 TB). Therefore, robust memory and computational resources are essential for effective summarization.

Numerical Data: A significant hurdle in medical article summarization lies in the abundance of numerical data, encompassing medication concentrations, patient ages, statistics, quantities, and dates. This poses a challenge due to the limited vocabulary used to

train the summarization model, which may struggle to retain comprehensive knowledge about all utilized numbers and accurately integrate them into generated summaries.

To tackle the challenges posed by Automatic Text Summarization, numerous systems have emerged in the past decade to address this issue. However, it's crucial to assess the quality of generated summaries to improve automatic summarization systems. Thus, the field of Automatic Summary Evaluation has developed alongside Text Summarization, aiming to ascertain whether automatically generated summaries are concise, meaningful, and coherent.

In the realm of automatic evaluation, determining an "ideal" or unequivocally "correct" summary is challenging, as summaries can be appraised based on diverse criteria like quality, in formativeness, and efficiency impact [166]. The effectiveness of evaluation metrics depends on specific criteria, and the quality of evaluation is influenced not only by the automatic system but also by human judgment, especially in cases where human competence is essential.

Concerning fairness, assessing extractive summaries becomes challenging when the evaluation approach relies on lexical content [167]. Dependency is another challenge, with many evaluation methods relying on human reference summaries, often termed gold standards [74], [167]. While some researchers have attempted automated methods without human intervention [168] the correlation with manual approaches tends to decrease in such instances.

The evaluation domain introduces variability, as the performance of each system is contingent on the domain to which candidate summaries belong. For instance, certain approaches excel in the biomedical domain [74] while others exhibit greater accuracy in the news domain [168].

Given the inherent connection between automatic text summarization and automatic summary evaluation, the challenges intensify, demanding comprehensive consideration of various aspects to deliver a summarization system that maximizes accuracy.

6.2. Evaluation Metrics

Precision and Recall, two widely recognized metrics for evaluating extractive summaries, involve comparing system-generated summaries to human-generated ones (gold standards) and calculating lexical overlap.

Precision is defined as the ratio of correctly chosen system sentences to those chosen by the system [169]:

$$Precision = \frac{\text{system selected}}{\text{sentences selected by system}} \quad (6.1)$$

Recall, on the other hand, represents the fraction of sentences selected by humans that were accurately identified by the system [169]:

$$Recall = \frac{\text{system selected}}{\text{sentences selected by human}} \quad (6.2)$$

According to [169], Precision and Recall have few drawbacks such as;

- **Human Variation:** The subjective nature of human sentence selection can lead to considerable variability, with different individuals choosing different sentences.
- **Granularity:** Sentences may vary in length, resulting in variations in information granularity.
- **Semantic Equivalence:** Two sentences with different wording may convey the same meaning.

In this thesis, we have contributed to the creation and implementation of an automatic extractive summarization system specifically designed for lengthy medical transcripts. Evaluating such a system requires an effective approach that offers a reasonable estimation of the quality of the generated summaries. ROUGE stood out as the predominant evaluation approach during the period under concern.

ROUGE, which stands for Recall-Oriented Understudy for Gisting Evaluation, was introduced by Lin in 2004 and has emerged as a highly influential method for assessing automatic summaries. It relies on word overlap between a candidate summary and reference summaries.

Various ROUGE variants exist, and elaborate on ROUGE-N[167], which is associated with recall between the candidate summary and reference summaries.

The equation above computes ROUGE-N using one reference summary. The following equation computes it using multiple references:

$$ROUGE - N_{Multi} = argmax_i(ROUGE - N\{R_i, S\}) \quad (6.3)$$

where:

• S is a candidate summary

• r_i is every reference summary in RS

ROUGE-L, which stands for Longest Common Subsequence, operates by examining two-word sequences, denoted as X and Y. Specifically, ROUGE-L searches for the longest common subsequence of X within Y, with the assumption that Y is the larger sequence compared to X.

6.3. Validation of Research

Medical data plays a crucial role as it encompasses information about human diseases and their symptoms. In the past, such data remained undisclosed as individuals were reluctant to discuss it. Over time, however, numerous transcripts have been generated containing medical history, symptoms, and recommended measures. These transcripts serve as valuable resources for individuals, doctors, clinical experts, and researchers. MTSamples data, utilized for text summarization, comprises 4996 real summaries from transcripts covering 40 domains, including Allergy, Autopsy, Bariatrics, Cardio, Cosmetic, Neurology, Diet and Nutrition, Discharge summary, General medicine, and more. To validate the research, five significant transcripts have been chosen.

Five major domains, comprising larger samples, include Neurology with 224 samples, General medicine with 260 samples, Gynaecology with 154 samples, Dental with 28 samples, and Cardiovascular with 372 samples. This results in the creation of a corpus named MT Corpus, consisting of a total of 1,040 transcripts. Notably, in previous state-of-the-art techniques, research predominantly focused on PubMed and BioMed articles. However, the authors exploration in the realm of real medical transcripts. To address this gap and streamline the process of reading, comprehending, and providing diagnoses to patients, a new corpus, MTCorpus, has been developed as discussed in chapter 3.

In this research thesis, we have utilized medical transcripts text data for summarization, facing challenges due to the length and complexity of the information. To overcome these challenges, we introduced three innovative methods for text summarization. The results of each method, along with their evaluations using the ROUGE method, are shown here.

Firstly, our initial approach involves leveraging the Metathesaurus from the Unified Medical Language System (UMLS) to extract concepts associated with named entities. We then apply the BERT method to generate a concise summary for biomedical text data.

Following that, we introduce a novel unsupervised approach that emphasizes semantic similarity and keyword-phrase extraction using a domain-independent methodology. This method is designed to cater to both single-document and multi-document summarization, providing a versatile solution that addresses limitations observed in prior research.

Lastly, we present a distinctive framework capable of intelligent and contextually aware summarization of biomedical literature. This involves the development of a deep neural network binary classifier, and the utilization of a bidirectional long-short term memory recurrent neural network to generate a concise summary of biomedical transcripts.

6.4. Results and Discussion

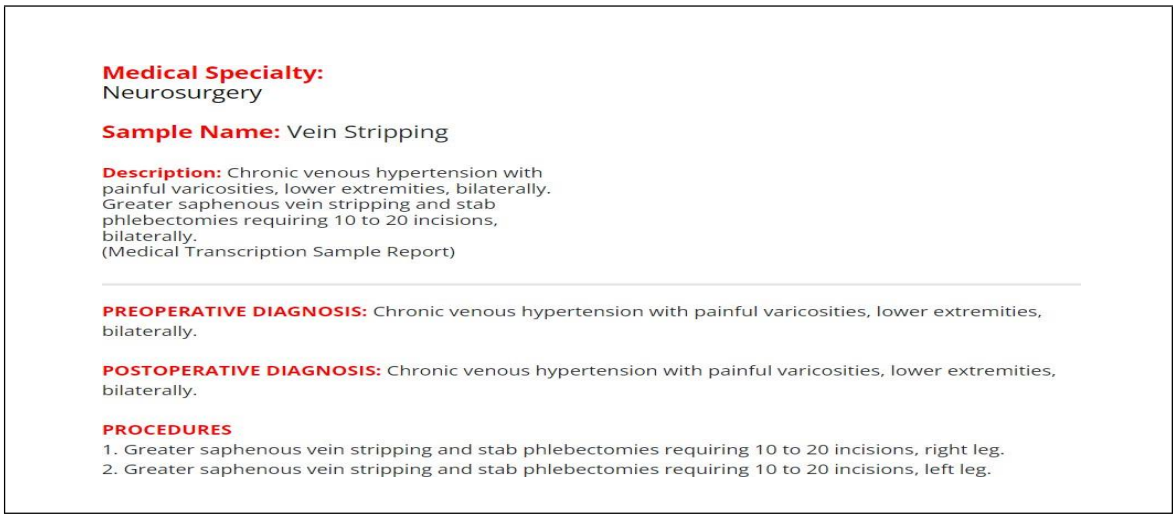
To validate the research two queries have been formulated to assess the effectiveness of the proposed approaches (3) in the biomedical domain.

Research Question 1 (RQ1): Does the envisioned approach yield promising outcomes when applied to the newly developed MTCorpus?

Research Question 2 (RQ2): Does the proposed approach demonstrate enhanced results when compared to state-of-the-art approaches on existing biomedical articles?

To answer these research questions, the process is illustrated in following steps:

Step 1: In each methodology, to generate a multi-document summary, a corpus is assembled for each domain, utilizing both sample descriptions and transcripts. The initial steps involve converting the text into plain text, converting all sentences to lowercase, and implementing stemming and stop-word removal processes. Specific stop words for each domain are defined. The Document Term Matrix (DTM) is constructed using Term Frequency-Inverse Document Frequency (Tf-Idf). After the pre-processing steps, some representative keywords for the neurology domain are presented. Next, the sample transcript for the text summarization is shown that has evaluated the research on all the proposed methods. Fig 6.1 shows the sample transcript and basic detail of the symptoms followed by Detailed Sample transcript in Fig. 6.2.



Medical Specialty:
Neurosurgery

Sample Name: Vein Stripping

Description: Chronic venous hypertension with painful varicosities, lower extremities, bilaterally. Greater saphenous vein stripping and stab phlebectomies requiring 10 to 20 incisions, bilaterally.
(Medical Transcription Sample Report)

PREOPERATIVE DIAGNOSIS: Chronic venous hypertension with painful varicosities, lower extremities, bilaterally.

POSTOPERATIVE DIAGNOSIS: Chronic venous hypertension with painful varicosities, lower extremities, bilaterally.

PROCEDURES
1. Greater saphenous vein stripping and stab phlebectomies requiring 10 to 20 incisions, right leg.
2. Greater saphenous vein stripping and stab phlebectomies requiring 10 to 20 incisions, left leg.

Fig 6.1. shows the sample transcript and basic detail of the symptoms.

After obtaining the informed consent, the patient was taken to the operating room where she underwent a general endotracheal anaesthesia. A time-out process was followed and antibiotics were given. Then, both legs were prepped and draped in the usual fashion with the patient was in the supine position. An incision was made in the right groin and the greater saphenous vein at its junction with the femoral vein was dissected out and all branches were ligated and divided. Then, an incision was made just below the knee where the greater saphenous vein was also found and connection to varices from the calf were seen. A third incision was made in the distal third of the right thigh in the area where there was a where there was a communication with large branch varicosities. Then, a vein stripper was passed from the right calf up to the groin and the greater saphenous vein, which was divided, was stripped without any difficultly. Several minutes of compression was used for hemostasis. Then, the exposed branch varicosities both in the lower third of the thigh and in the calf were dissected out and then many stabs were performed to do stab phlebectomies at the level of the thigh and the level of the calf as much as the position would allow us to do. Then in the left thigh, a groin incision was made and the greater saphenous vein was dissected out in the same way as was on the other side. Also, an incision was made in the level of the knee and the saphenous vein was isolated there. The saphenous vein was stripped and a several minutes of local compression was performed for hemostasis. Then, a number of stabs to perform phlebectomy were performed at the level of the calf to excise branch varicosities to the extent that the patient's position would allow us. Then, all incisions were closed in layers with Vicryl and staples. Then, the patient was placed in the prone position and the stab phlebectomies of the right thigh and calf and left thigh and calf were performed using 10 to 20 stabs in each leg. The stab phlebectomies were performed with a hook and they were very satisfactory. Hemostasis achieved with compression and then staples were applied to the skin. Then, the patient was rolled onto a stretcher where both legs were wrapped with the Kerlix, fluffs, and Ace bandages. Estimated blood loss probably was about 150 mL. The patient tolerated the procedure well and was sent to recovery room in satisfactory condition. The patient is to be observed, so a decision will be made whether she needs to stay overnight or be able to go home.

Fig. 6.2. Sample transcript for Summarization

Golden summary for this sample text is shown in Fig. 6.3 followed by keywords in Fig. 6.4.

Chronic venous hypertension with painful varicosities, lower extremities, bilaterally. Greater saphenous vein stripping and stab phlebectomy requiring 10 to 20 incisions, bilaterally. A time-out process was followed and antibiotics were given. Then, both legs were prepped and draped in the usual fashion with the patient was in the supine position. An incision was made in the right groin and the greater saphenous vein at its junction with the femoral vein was dissected out and all branches were ligated and divided.

Then, a vein stripper was passed from the right calf up to the groin and the greater saphenous vein, which was divided, was stripped without any difficultly. Several minutes of compression was used for hemostasis. Also, an incision was made in the level of the knee and the saphenous vein was isolated there. The saphenous vein was stripped and a several minutes of local compression was performed for hemostasis. Then, a number of stabs to perform phlebectomy were performed at the level of the calf to excise branch varicosities to the extent that the patient's position would allow us. Then, all incisions were closed in layers with Vicryl and staples., Hemostasis achieved with compression and then staples were applied to the skin. Then, the patient was rolled onto a stretcher where both legs were wrapped with the Kerlix, fluffs, and Ace bandages. Estimated blood loss probably was about 150 mL The patient tolerated the procedure well and was sent to recovery room in satisfactory condition.

Fig. 6.3. Golden Summary

Operating room, General endotracheal anesthesia, Antibiotics, Legs prepped and draped Supine position, Greater saphenous vein, Femoral vein, Varices, varicosities, Vein stripper, Right calf to the groin, Stripping of the greater saphenous vein, Compression, Exposed branch varicosities, Dissection, Stab phlebectomies, Hemostasis, Groin incision, saphenous vein, Phlebectomy, Prone position, Stab phlebectomies on both legs, 10 to 20 stabs in each leg,Hook for stab phlebectomies, Hemostasis achieved with compression, Legs wrapped with Kerlix, fluffs, and Ace bandages, Estimated blood loss, Procedure tolerance, Recovery room, Patient discharge

Fig.6.4. Keywords

Then using these keywords, using first methodology, the UMLS Metathesaurus is found and then using the BERT the text summary of the above text as depicted in Fig. 6.5.

After obtaining the informed consent, the patient was taken to the operating room where she underwent a general endotracheal anaesthesia. A time-out process was followed and antibiotics were given. Then, an incision was made just below the knee where the greater saphenous vein was also found and connection to varices from the calf were seen. Then, the exposed branch varicosities both in the lower third of the thigh and in the calf were dissected out and then many stabs were performed to do stab phlebectomies at the level of the thigh and the level of the calf as much as the position would allow us to do.

Fig. 6.5. Summary generated by BERT

For the second approach where semantic similarity and keyword-phrase extraction method is applied and final summary is generated as shown in Fig.6.6.

The surgical procedure involved making an incision in the right groin to dissect out the greater saphenous vein at its junction with the femoral vein. All branches were ligated and divided. Another incision, just below the knee, revealed the greater saphenous vein with connections to varices from the calf. Similar steps were taken in the left thigh, where a groin incision was made to dissect out the greater saphenous vein. To ensure haemostasis, several minutes of compression were applied during the procedure. Greater saphenous vein stripping, along with 10 to 20 stab phlebectomies involving incisions, was performed on the right leg due to chronic venous hypertension with painful varicosities in the lower extremities. The condition was also noted bilaterally.

Fig. 6.6. Summary generated based on semantic similarity and keyword phrase extraction

Similarly, in the third approach we have applied deep neural network binary classifier and bidirectional long-short term memory recurrent neural network to form the summary. The summary is shown in Fig. 6.7.

Bilateral chronic venous hypertension and painful varicosities were addressed with greater saphenous vein stripping and stab phlebectomy involving 10 to 20 incisions. The procedure began with a comprehensive time-out process and administration of antibiotics. Both legs were prepped and draped in the supine position before making an incision in the right groin. The greater saphenous vein at its junction with the femoral vein was dissected, ligated, and divided. Closure involved layering incisions with Vicryl and staples, ensuring hemostasis and successful completion. Post-procedure, the patient, with an estimated blood loss of around 150 mL, was sent to the recovery room in satisfactory condition. Hemostasis was reaffirmed through compression, and staples were applied to the skin, contributing to the overall success of the procedure.

Fig. 6.7. Summary generated by LSTM

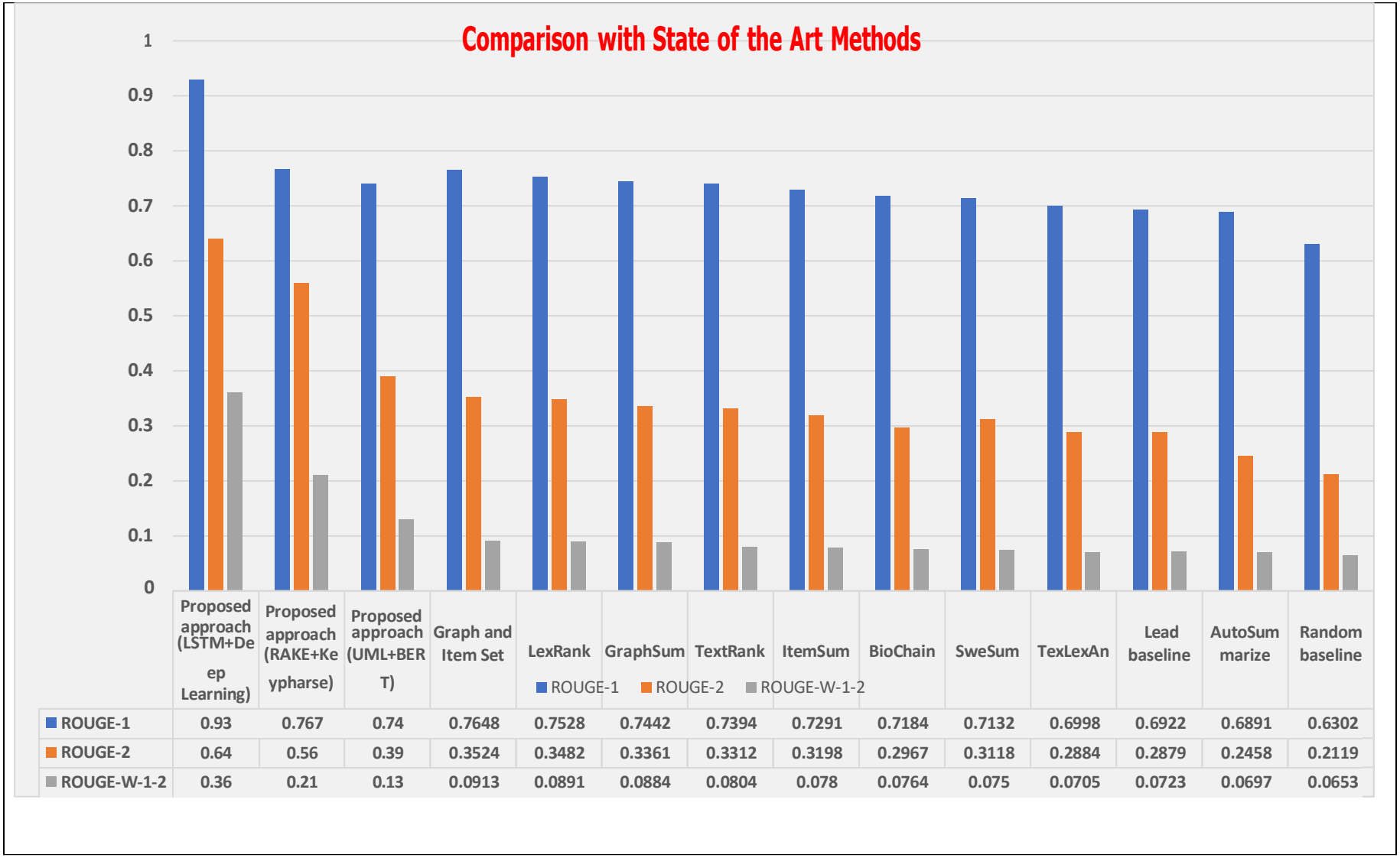
Upon comparison, it is evident that the final approach closely resembles the Golden Summary. However, for evaluation purposes, we applied the ROUGE score to each method.

Answer to the research question (1): In this research, three novel methods for text summarization are proposed, designed for single-document summaries. The evaluation of these methods involves biomedical text data, but their applicability extends to various domains. To assess our study, we initially compare the generic summary to a golden summary. Notably, golden generic summaries are unavailable in these domains. Therefore, domain experts were engaged to assess and approve these summaries. Three doctors, serving as experts, evaluated the generic summary in each domain and provided scores based on their expertise. A sample of the generated generic summary is presented in Fig 6.3. From the annotators comments authors can say that the constructed MTsample corpus attain promising results using the proposed approaches.

Answer to the research question (2): Our proposed approaches were subjected to comparison with baseline methods in biomedical text summarization for validation. In a recent research paper, a text summarization methodology utilizing a graph-based approach with the FP-Growth method was introduced [23]. This study validated its method using 400 biomedical research papers. Similarly, we have also identified research papers in our domain and compared the results employing our methodology. A collection of 167 research papers in our domain was utilized, applying the single transcript summarization approach. In this context, the introduction part of the research paper was treated as transcripts, and the abstract served as the golden summary. Table 6.1 provides a brief excerpt of the biomedical research paper summary and the summary generated by our proposed methods. Table 6.1. and Fig. 6.8. depicts the comparison between the baseline methods and our proposed approach in terms of ROUGE metrics.

Table 6.1. Comparison with the state-of -the-art methods

S. No.	Methods	ROUGE-1	ROUGE-2	ROUGE-W-1-2
1	**Proposed approach (LSTM + Deep Learning) [175]**	**0.93**	**0.64**	**0.36**
2	**Proposed approach (RAKE+ Keypharse) [176]**	0.767	0.56	0.2100
3	**Proposed approach (UML+BERT) [177]**	0.74	0.39	0.1300
4	Graph and Item Set [108]	0.7648	0.3524	0.0913
5	LexRank [39]	0.7528	0.3482	0.0891
6	GraphSum [75]	0.7442	0.3361	0.0884
7	TextRank [75]	0.7394	0.3312	0.0804
8	ItemSum [75]	0.7291	0.3198	0.0780
9	BioChain [127]	0.7184	0.2967	0.0764
10	SweSum [170]	0.7132	0.3118	0.0750
11	TexLexAn [171]	0.6998	0.2884	0.0705
12	Lead baseline [172]	0.6922	0.2879	0.0723
13	AutoSummarize [173]	0.6891	0.2458	0.0697
14	Random baseline [174]	0.6302	0.2119	0.0653



	Proposed approach (LSTM+Deep Learning)	Proposed approach (RAKE+Keypharse)	Proposed approach (UML+BERT)	Graph and Item Set	LexRank	GraphSum	TextRank	ItemSum	BioChain	SweSum	TexLexAn	Lead baseline	AutoSummarize	Random baseline
ROUGE-1	0.93	0.767	0.74	0.7648	0.7528	0.7442	0.7394	0.7291	0.7184	0.7132	0.6998	0.6922	0.6891	0.6302
ROUGE-2	0.64	0.56	0.39	0.3524	0.3482	0.3361	0.3312	0.3198	0.2967	0.3118	0.2884	0.2879	0.2458	0.2119
ROUGE-W-1-2	0.36	0.21	0.13	0.0913	0.0891	0.0884	0.0804	0.078	0.0764	0.075	0.0705	0.0723	0.0697	0.0653

Fig 6.8. Comparison with the state-of-the-art methods

6.5. Conclusion

In this research, the focus is on summarizing extensive medical transcripts, employing various methodologies, including cutting-edge deep architectures such as BERT, T5, and PEGASUS, known for their adaptability in natural language processing tasks. The challenges of Automatic Text Summarization are addressed through the development of an automatic extractive summarization system tailored for lengthy medical transcripts. The need for evaluating the quality of generated summaries is emphasized, leading to the emergence of the field of Automatic Summary Evaluation. Precision and Recall, common metrics for extractive summaries, are employed to assess the system-generated summaries against human-generated ones, utilizing lexical overlap.

The research contributes to the field by proposing different approaches. The initial approach combines the Unified Medical Language System (UMLS) Metathesaurus for named entity concept extraction and the BERT method for generating concise biomedical text summaries. Another approach introduces an unsupervised method emphasizing semantic similarity and keyword-phrase extraction, addressing limitations observed in prior research. The final approach involves a unique framework for intelligent and contextually aware summarization of biomedical literature, utilizing a deep neural network binary classifier and a bidirectional long-short term memory recurrent neural network. To validate the proposed approaches, comparisons are made with baseline methods in biomedical text summarization, including a recent graph-based approach with the FP-Growth method. The results indicate that the last proposed approach outperforms state-of-the-art methods, achieving the highest ROUGE score of 0.96, surpassing the scores of the first and second approach (0.74, 0.76).

The research concludes that the proposed methods demonstrate superior results in the medical domain compared to existing state-of-the-art techniques, highlighting the efficacy of the developed summarization approaches for biomedical literature.

CHAPTER 7

CONCLUSION and FUTURE WORK

In this work, the indispensable role of the healthcare sector and the biomedical domain in society was examined, with an emphasis on their critical contributions to public health, medical advancements, and overall well-being. Diverse health needs and challenges were addressed by these sectors, from routine check-ups to ground-breaking medical research, ultimately enhancing the quality of life for individuals worldwide.

Given the exponential growth of biomedical literature and research outputs, the need for efficient information retrieval and comprehension became paramount. Automatic text summarization emerged as a crucial solution to navigate and distill relevant insights from the vast amounts of available information. By utilizing domain-specific knowledge and advanced algorithms, automatic summarization systems were able to simplify complex biomedical texts into easily understandable summaries, facilitating knowledge dissemination and interdisciplinary collaboration.

The two main approaches of Automatic Text Summarization, Extractive and Abstractive, were explored. Our focus was on extractive summarization techniques in the biomedical domain, addressing issues such as redundancy, coherence, and the risk of overlooking crucial information. Various algorithms and approaches, including Frequency-based Methods, Graph-based Algorithms, and Machine Learning Approaches, were examined to identify and extract key sentences or phrases from biomedical documents. Additionally, hybrid approaches that combined multiple techniques were explored to improve accuracy and coverage while effectively summarizing complex biomedical texts.

To address identified research gaps, novel approaches for biomedical text summarization were proposed. These approaches included leveraging the Metathesaurus from UMLS to extract named entity concepts and applying the BERT method to generate concise summaries from Pubmed and Mtsamples. Furthermore, an unsupervised approach focusing on semantic similarity and keyword-phrase extraction for both single-document and multi-document summarization was proposed. Additionally, a distinctive framework utilizing deep neural networks for contextually aware summarization of biomedical literature was introduced, employing a binary classifier and bidirectional long-short term memory recurrent neural network.

In validation of the proposed approaches, comparisons were made with baseline methods in biomedical text summarization, including a recent graph-based approach with the FP-Growth method. The results showcased the superior performance of the last proposed approach, which achieved the highest ROUGE score of 0.96, surpassing the scores of the first and second approaches (0.74, 0.76).

In conclusion, the findings of this thesis demonstrated the effectiveness and superiority of the developed summarization approaches for biomedical literature. These advancements hold great promise in enhancing information retrieval, knowledge dissemination, and interdisciplinary collaboration within the medical domain, ultimately contributing to improved healthcare outcomes and advancements in medical research.

REFERENCES

[1] W. Hoorn van and Wesley, "Automatic Text Summarization as a Text Extraction Strategy for Effective Automated Highlighting," 2018, [Online]. Available: https://theses.ubn.ru.nl/handle/123456789/5560

[2] S. Sonawane, P. Kulkarni, C. Deshpande, and B. Athawale, "Extractive summarization using semigraph (ESSg)," Evolving Systems, vol. 0, no. 0, p. 0, 2018, doi: 10.1007/s12530-018-9246-8.

[3] A. Sinha, A. Yadav, and A. Gahlot, "Extractive Text Summarization using Neural Networks," 2018, [Online]. Available: http://arxiv.org/abs/1802.10137

[4] R. Ferreira et al., "Assessing sentence scoring techniques for extractive text summarization," Expert Systems with Applications, vol. 40, no. 14. pp. 5755–5764, 2013. doi: 10.1016/j.eswa.2013.04.023.

[5] E. Zolotareva, T. M. Tashu, and T. Horváth, "Abstractive Text Summarization using Transfer Learning." [Online]. Available: https://www.kaggle.com/pariza/bbc-news-summary

[6] M. T. Nayeem, T. A. Fuad, and Y. Chali, "Neural diverse abstractive sentence compression generation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2019, pp. 109–116. doi: 10.1007/978-3-030-15719-7_14.

[7] N. Kumar Nagwani and S. Verma, "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm," Int J Comput Appl, vol. 17, no. 2, pp. 36–40, 2011, doi: 10.5120/2190-2778.

[8] V. K. Gupta, S. Operation, and T. J. Siddiqui, "Multi-Document Summarization Using Sentence Clustering," 2012.

[9] S. Mangla, "Multi-document summarization using sentence embeddings," Project-Archive.Inf.Ed.Ac.Uk, 2012, [Online]. Available: http://project-archive.inf.ed.ac.uk/msc/20150155/msc_proj.pdf

[10] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01, pp. 19–25, 2001, doi: 10.1145/383952.383955.

[11] M. Afsharizadeh, H. Ebrahimpour-Komleh, and A. Bagheri, "Query-oriented text summarization using sentence extraction technique," 2018 4th International Conference on Web Research, ICWR 2018, pp. 128–132, 2018, doi: 10.1109/ICWR.2018.8387248.

[12] Y. Ouyang, W. Li, S. Li, and Q. Lu, "Applying regression models to query-focused multi-document summarization," Inf Process Manag, vol. 47, no. 2, pp. 227–237, 2011, doi: 10.1016/j.ipm.2010.03.005.

[13] X. Hu, "Biomedical Literature Mining for Disease Knowledge," Text, vol. 1159, no. May, 2014, doi: 10.1007/978-1-4939-0709-0.

[14] Y. Du, Q. Li, L. Wang, and Y. He, "Biomedical-domain pre-trained language model for extractive summarization," Knowl Based Syst, vol. 199, Jul. 2020, doi: 10.1016/j.knosys.2020.105964.

[15] R. Khan, Y. Qian, and S. Naeem, “Extractive based Text Summarization Using K-,” no. May, pp. 33–44, 2019, doi: 10.5815/ijieeb.2019.03.05.

[16] A. Kukkar and R. Mohana, for Unsupervised Automatic Extractive Bug Report Summarization. Springer Singapore. doi: 10.1007/978-981-13-2354-6.

[17] R. Ferreira et al., “Assessing sentence scoring techniques for extractive text summarization,” Expert Systems with Applications, vol. 40, no. 14. pp. 5755–5764, 2013. doi: 10.1016/j.eswa.2013.04.023.

[18] R. T. Anchiêta and R. S. Moura, “Exploring Unsupervised Learning Towards Extractive Summarization of User Reviews,” Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web, ’WebMedia’17, pp. 217–220, 2017doi: 10.1145/3126858.3131583.

[19] R. Khan, Y. Qian, and S. Naeem, “Extractive based Text Summarization Using KMeans and TF-IDF,” International Journal of Information Engineering and Electronic Business, vol. 11, no. 3, pp. 33–44, 2019, doi: 10.5815/ijieeb.2019.03.05.

[20] T. Uçkan and A. Karcı, “Extractive multi-document text summarization based on graph independent sets,” no. xxxx, 2019, doi: 10.1016/j.eij.2019.12.002.

[21] W. Li, M. Wu, Q. Lu, W. Xu, and C. Yuan, “Extractive summarization using inter- and intra-event relevance,” Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL ’06, no. July, pp. 369–376, 2006, doi: 10.3115/1220175.1220222.

[22] B. Mutlu, E. A. Sezer, and M. A. Akcayol, “Multi-document extractive text summarization: A comparative assessment on features,” Knowl Based Syst, vol. 183, 2019, doi: 10.1016/j.knosys.2019.07.019.

[23] F. Mohsen, J. Wang, and K. Al-Sabahi, “A hierarchical self-attentive neural extractive summarizer via reinforcement learning (HSASRL),” Applied Intelligence, vol. 50, no. 9, pp. 2633–2646, Sep. 2020, doi: 10.1007/s10489-020-01669-5.

[24] Luhn, H. P. The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165, 1958.

[25] K. Knight and D. Marcu, “Summarization beyond sentence extraction: A probabilistic approach to sentence compression ☆,” 2002. [Online]. Available: www.elsevier.com/locate/artint

[26] J. M. Conroy and D. P. O’leary, “Text Summarization via Hidden Markov Models,” 2001.

[27] H. Jing and K. R. Mckeown, “The Decomposition of Human-Written Summary Sentences,” 1999.

[28] C. Fang, D. Mu, Z. Deng, and Z. Wu, “Word-sentence co-ranking for automatic extractive text summarization,” Expert Syst Appl, vol. 72, pp. 189–195, 2017, doi: 10.1016/j.eswa.2016.12.021.

[29] D. Demner-Fushman and J. Lin, “Answer Extraction, Semantic Clustering, and Extractive Summarization for Clinical Question Answering,” Association for Computational Linguistics, 2006.

[30] Y.-W. Chen and C.-J. Lin, “Combining SVMs with Various Feature Selection Strategies,” Feature Extraction, no. 1, pp. 315–324, doi: 10.1007/978-3-540-35488-8_13.

[31] F. Ali, S. El-Sappagh, and D. Kwak, “Fuzzy ontology and LSTM-based text mining: A transportation network monitoring system for assisting travel,” Sensors (Switzerland), vol. 19, no. 2, 2019, doi: 10.3390/s19020234.

[32] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, “On the Properties of Neural Machine Translation: Encoder-Decoder Approaches,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1259

[33] S. Bakas et al., “Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge,” Nov. 2018, [Online]. Available: http://arxiv.org/abs/1811.02629

[34] V. V. Giri, Dr. M. M. Math, and Dr. U. P. Kulkarni, “A Survey of Automatic Text Summarization System for Different Regional Language in India,” Bonfring International Journal of Software Engineering and Soft Computing, vol. 6, no. Special Issue, pp. 52–57, Oct. 2016, doi: 10.9756/bijsesc.8242.

[35] S. Chopra, M. Auli, and A. M. Rush, “Abstractive Sentence Summarization with Attentive Recurrent Neural Networks.” [Online]. Available: http://torch.ch/

[36] J. Cheng and M. Lapata, “Neural Summarization by Extracting Sentences and Words,” Mar. 2016, [Online]. Available: http://arxiv.org/abs/1603.07252

[37] H. P. Edmundson, “New Methods in Automatic Extracting.”

[38] J. Kupiec, J. Pedersen, and F. Chen, “A Trainable Document Summarizer.”

[39] R. Nallapati, B. Zhou, C. N. dos santos, C. Gulcehre, and B. Xiang, “Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond,” Feb. 2016, [Online]. Available: http://arxiv.org/abs/1602.06023

[40] H. Lin, J. Bilmes, and S. Xie, “Graph-based Submodular Selection for Extractive Summarization.”

[41] P. Cimiano and H. S. Pinto, “Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6317 LNAI, no. October, 2010, doi: 10.1007/978-3-642-16438-5.

[42] K. Sarkar, “Automatic single document text summarization using key concepts in documents,” Journal of Information Processing Systems, vol. 9, no. 4, pp. 602–620, 2013, doi: 10.3745/JIPS.2013.9.4.602.

[43] A. Abuobieda, N. Salim, A. T. Albaham, A. H. Osman, and Y. J. Kumar, “Text summarization features selection method using pseudo genetic-based model,” in Proceedings - 2012 International Conference on Information Retrieval and Knowledge Management, CAMP’12, 2012, pp. 193–197. doi: 10.1109/InfRKM.2012.6204980.

[44] E. Cardinaels, S. Hollander, and B. J. White, “Automatic summarization of earnings releases: attributes and effects on investors’ judgments,” Review of Accounting Studies, vol. 24, no. 3, pp. 860–890, Sep. 2019, doi: 10.1007/s11142-019-9488-0.

[45] A. Tamura, K. Ishikawa, M. Saikou, and M. Tsuchida, “Extractive Summarization Method for Contact Center Dialogues based on Call Logs.”

[46] N. Vanetik and M. Litvak, “Query-based summarization using MDL principle,” 2017.

[47] G. Erkan and D. R. Radev, “LexRank: Graph-based lexical centrality as salience in text summarization,” Journal of Artificial Intelligence Research, vol. 22, pp. 457–479, 2004, doi: 10.1613/jair.1523.

[48] D. R. Radev, “LexRank : Graph-based Lexical Centrality as Salience in Text Summarization,” vol. 22, pp. 457–479, 2004, doi: 10.1613/jair.1523.

[49] M. M. Rahman and C. K. Roy, “TextRank Based Search Term Identification for Software Change Tasks,” 2018, doi: 10.1109/SANER.2015.7081873.

[50] P. Nakov, P. Nakov, A. Popova, A. Popova, P. Mateev, and P. Mateev, “Weight functions impact on LSA performance,” Proceedings of the EuroConference Recent Advances in Natural Language Processing, RANLP 2001, no. January 2001, pp. 187–193, 2001, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.69.9244

[51] Y. Liu, A. An, and X. Huang, “Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles,” Pacific-Asia Conference on Knowledge Discovery and Data Mining, vol. 3918, pp. 107–118, 2006, doi: 10.1038/ncb2329.

[52] W. Du, W. Du, Z. Zhan, and Z. Zhan, “Building decision tree classifier on private data,” Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14, pp. 1–8, 2002, [Online]. Available: http://portal.acm.org/citation.cfm?id=850784

[53] A. L. and M. Wiener, “Classification and Regression by randomForest. R News 2,” vol. 3, no. December 2002, pp. 18–22, 2003.

[54] Z. Feng, L. Mo, and M. Li, “A Random Forest-based ensemble method for activity recognition,” Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, vol. 2015-Novem, pp. 5074–5077, 2015, doi: 10.1109/EMBC.2015.7319532.

[55] T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse, “An Empirical Study of Learning from Imbalanced Data Using Random Forest,” in 19th IEEE International Conference on Tools with Artificial Intelligence, 2007. ICTAI 2007., 2007, pp. 310–317.

[56] X. Wang, T. Liu, X. Zheng, H. Peng, J. Xin, and B. Zhang, “Short-term prediction of groundwater level using improved random forest regression with a combination of random features,” Appl Water Sci, vol. 8, no. 5, pp. 1–12, 2018, doi: 10.1007/s13201-018-0742-6.

[57] D. Anand and R. Wagh, “Effective deep learning approaches for summarization of legal texts,” Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 5, pp. 2141–2150, May 2022, doi: 10.1016/j.jksuci.2019.11.015.

[58] A. Louis, S. K. Dash, E. T. Barr, and C. Sutton, “Deep Learning to Detect Redundant Method Comments,” 2018, [Online]. Available: http://arxiv.org/abs/1806.04616

[59] Q. Dong, S. Gong, and X. Zhu, “Imbalanced Deep Learning by Minority Class Incremental Rectification,” IEEE Trans Pattern Anal Mach Intell, vol. 41, no. 6, pp. 1367–1381, 2019, doi: 10.1109/TPAMI.2018.2832629.

[60] K. Sharma, A. Gaikwad, S. Patil, P. Kumar, and D. P. Salapurkar, “Automated Document Summarization and Classification Using Deep Learning,” International Research Journal of Engineering and Technology (IRJET), vol. 5, no. 6, pp. 1182–1185, 2018, [Online]. Available: https://sci-hub.tw/

[61] Y. P. Chen, Y. Y. Chen, J. J. Lin, C. H. Huang, and F. Lai, "Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): Development and performance evaluation," JMIR Med Inform, vol. 8, no. 4, Apr. 2020, doi: 10.2196/17787.

[62] R. Nallapati, F. Zhai, and B. Zhou, "SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents." [Online]. Available: https://github.com/deepmind/rc-data

[63] S. Number Cruncher Statistical Systems (NCSS) and A. R. Reserved, "Hierarchical Clustering/Dendrograms," NCSS Documentation, p. 15, 2019, [Online]. Available: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Hierarchical_Clustering-Dendrograms.pdf

[64] A. J. C. Trappey, C. V Trappey, F. Hsu, and D. W. Hsiao, "A Fuzzy Ontological Knowledge Document Clustering Methodology," vol. 39, no. 3, pp. 806–814, 2009.

[65] R. M. Alguliyev, R. M. Aliguliyev, N. R. Isazade, A. Abdi, and N. Idris, "COSUM: Text summarization based on clustering and optimization," Expert Syst, vol. 36, no. 1, pp. 1–17, 2019, doi: 10.1111/exsy.12340.

[66] S. Mani, R. Catherine, V. S. Sinha, and A. Dubey, "AUSUM: Approach for unsupervised bug report summarization," Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering, FSE 2012, no. May 2014, 2012, doi: 10.1145/2393596.2393607.

[67] Hahn, U., & Mani, I., "The Challenges of Automatic Summarization," 33(11), 29-36. 2000.

[68] I. Safder, J. Sarfraz, S. U. Hassan, M. Ali, and S. Tuarob, "Detecting target text related to algorithmic efficiency in scholarly big data using recurrent convolutional neural network model," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2017, pp. 30–40. doi: 10.1007/978-3-319-70232-2_3.

[69] E. Yulianti, R. C. Chen, F. Scholer, W. B. Croft, and M. Sanderson, "Document summarization for answering non-factoid queries," IEEE Trans Knowl Data Eng, vol. 30, no. 1, pp. 15–28, Jan. 2018, doi: 10.1109/TKDE.2017.2754373.

[70] T. Vodolazova, "The role of statistical and semantic features in single-document extractive summarization," Artif Intell Res, vol. 2, no. 3, Apr. 2013, doi: 10.5430/air.v2n3p35.

[71] J. Ouyang, B. Song, and K. Mckeown, "A Robust Abstractive System for Cross-Lingual Summarization," Association for Computational Linguistics.

[72] K. Moritz et al., "Teaching Machines to Read and Comprehend." [Online]. Available: http://www.github.com/deepmind/rc-data/

[73] C. Napoles, M. Gormley, and B. Van Durme, "Annotated Gigaword," NAACL-HLT. [Online]. Available: http://vtd-xml.sourceforge.net

[74] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and N. Goharian, "SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions," Jun. 2018, [Online]. Available: http://arxiv.org/abs/1806.05258

[75] X. Zhang, F. Wei, and M. Zhou, “HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization,” May 2019, [Online]. Available: http://arxiv.org/abs/1905.06566

[76] A. Kornilova and V. Eidelman, “BillSum: A Corpus for Automatic Summarization of US Legislation,” Oct. 2019, doi: 10.18653/v1/D19-5406.

[77] S. Narayan, S. B. Cohen, and M. Lapata, “Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization,” Aug. 2018, [Online]. Available: http://arxiv.org/abs/1808.08745

[78] M. Grusky, M. Naaman, and Y. Artzi, “Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies,” Apr. 2018, [Online]. Available: http://arxiv.org/abs/1804.11283

[79] E. J. Benjamin et al., “Heart disease and stroke statistics - 2018 update: A report from the American Heart Association,” Circulation, vol. 137, no. 12, pp. E67–E492, Mar. 2018, doi: 10.1161/CIR.0000000000000558.

[80] L. Zhou, E. Hovy, and M. Rey, “A Web-Trained Extraction Summarization System,” Proceedings of the HLT-NAACL conference, no. May, pp. 1–7, 2003.

[81] J. M. Sanchez-gomez, M. A. Vega-rodríguez, and C. J. Pérez, “Experimental analysis of multiple criteria for extractive multi-document text summarization,” Expert Syst Appl, vol. 140, p. 112904, 2020, doi: 10.1016/j.eswa.2019.112904.

[82] V. K. Gupta and T. J. Siddiqui, “Multi-document summarization using sentence clustering,” 4th International Conference on Intelligent Human Computer Interaction: Advancing Technology for Humanity, IHCI 2012, 2012, doi: 10.1109/IHCI.2012.6481826.

[83] T. J. Siddiqui, “Multi-Document Summarization (MLTA 2014),” no. Mlta, 2014.

[84] E. Baralis, L. Cagliero, N. Mahoto, and A. Fiori, “GraphSum: Discovering correlations among multiple terms for graph-based summarization,” Inf Sci (N Y), vol. 249, pp. 96–109, Nov. 2013, doi: 10.1016/j.ins.2013.06.046.

[85] C. Paul, A. R. B, A. M. B, C. A. K. B, and P. S. B, “Efficient Graph-Based Document Similarity,” vol. 1, pp. 334–349, 2016, doi: 10.1007/978-3-319-34129-3.

[86] A. M. Rush, S. Chopra, and J. Weston, “A neural attention model for sentence summarization,” Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, no. September, pp. 379–389, 2015, doi: 10.18653/v1/d15-1044.

[87] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, “Multiple documents summarization based on evolutionary optimization algorithm,” Expert Syst Appl, vol. 40, no. 5, pp. 1675–1689, Apr. 2013, doi: 10.1016/j.eswa.2012.09.014.

[88] R. Rautray and R. Chandra, “Cat swarm optimization based evolutionary framework for multi document summarization,” Physica A, vol. 477, pp. 174–186, 2017, doi: 10.1016/j.physa.2017.02.056.

[89] D. Shen, J. Sun, H. Li, Q. Yang, and Z. Chen, “Document Summarization using Conditional Random Fields,” Science (1979), vol. 7, pp. 2862–2867, 2004.

[90] D. Shen, J. Sun, H. Li, Q. Yang, and Z. Chen, “Document Summarization using Conditional Random Fields,” Science (1979), vol. 7, pp. 2862–2867, 2004.

[91] N. Mittal and N. Vijay, “Text Summarization with Semantics Information,” 2011.

[92] L. Al Qassem, D. Wang, H. Barada, A. Al-Rubaie, and N. Almoosa, “Automatic {A}rabic Text Summarization Based on Fuzzy Logic,” Proceedings of the 3rd International Conference on Natural Language and Speech Processing, pp. 42–48, 2019, [Online]. Available: https://www.aclweb.org/anthology/W19-7406

[93] G. L. De La Peña Sarracén and P. Rosso, “Automatic text summarization based on betweenness centrality,” ACM International Conference Proceeding Series, vol. Part F1377, 2018, doi: 10.1145/3230599.3230611.

[94] Y. Ouyang, W. Li, S. Li, and Q. Lu, “Applying regression models to query-focused multi-document summarization,” Inf Process Manag, vol. 47, no. 2, pp. 227–237, Mar. 2011, doi: 10.1016/j.ipm.2010.03.005.

[95] C. Yadav and J. N. U. Sc, “A New LSA and Entropy-Based Approach for Automatic Text Document Summarization,” vol. 14, no. 4, pp. 1–32, 2018, doi: 10.4018/IJSWIS.2018100101.

[96] J. Steinberger and K. Ježek, “Using Latent Semantic Analysis in Text Summarization,” In Proceedings of ISIM 2004, no. June, pp. 93--100, 2004, doi: 10.1177/0165551511408848.

[97] Y. Gong and X. Liu, “Generic text summarization using relevance measure and latent semantic analysis,” Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’01, pp. 19–25, 2001, doi: 10.1145/383952.383955.

[98] J. Steinberger and K. Ježek, “Using Latent Semantic Analysis in Text Summarization,” In Proceedings of ISIM 2004, no. June, pp. 93--100, 2004, doi: 10.1177/0165551511408848.

[99] J. Y. Yeh, H. R. Ke, W. P. Yang, and I. H. Meng, “Text summarization using a trainable summarizer and latent semantic analysis,” Inf Process Manag, vol. 41, no. 1, pp. 75–95, 2005, doi: 10.1016/j.ipm.2004.04.003.

[100] T. K. Landauer and P. W. Foltz, “An Introduction to Latent Semantic Analysis,” pp. 259–284, 1998.

[101] D. Patel, S. Shah, and H. Chhinkaniwala, “Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique,” Expert Syst Appl, vol. 134, pp. 167–177, 2019, doi: 10.1016/j.eswa.2019.05.045.

[102] F. B. Goularte, S. M. Nassar, R. Fileto, and H. Saggion, “A text summarization method based on fuzzy rules and applicable to automated assessment,” Expert Syst Appl, vol. 115, pp. 264–275, 2019, doi: 10.1016/j.eswa.2018.07.047.

[103] R. Abbasi-ghalehtaki, H. Khotanlou, and M. Esmaeilpour, “Fuzzy evolutionary cellular learning automata model for text summarization,” Swarm Evol Comput, vol. 30, pp. 11–26, 2016, doi: 10.1016/j.swevo.2016.03.004.

[104] E. Valladares-vald, A. Sim, J. A. Olivas, and F. P. Romero, “A Fuzzy Approach for Sentences Relevance Assessment in Multi-document Summarization,” vol. 1, pp. 57–67, 2020, doi: 10.1007/978-3-030-20055-8.

[105] R. Khan, Y. Qian, and S. Naeem, “Extractive based Text Summarization Using K-,” no. May, pp. 33–44, 2019, doi: 10.5815/ijieeb.2019.03.05.

[106] E. Valladares-vald, A. Sim, J. A. Olivas, and F. P. Romero, “A Fuzzy Approach for Sentences Relevance Assessment in Multi-document Summarization,” vol. 1, pp. 57–67, 2020, doi: 10.1007/978-3-030-20055-8.

[107] J. M. Sanchez-gomez, M. A. Vega-rodríguez, and C. J. Pérez, “Experimental analysis of multiple criteria for extractive multi-document text summarization,” Expert Syst Appl, vol. 140, p. 112904, 2020, doi: 10.1016/j.eswa.2019.112904.

[108] C. Hark and A. Karcı, “Information Processing and Management Karc ı summarization : A simple and effective approach for automatic text summarization using Karc ı entropy Karcı summarization : A simple and effective approach for automatic text summarization using Karcı entropy,” Inf Process Manag, vol. 57, no. 3, p. 102187, 2020, doi: 10.1016/j.ipm.2019.102187.

[109] T. Uçkan and A. Karcı, “Extractive multi-document text summarization based on graph independent sets,” no. xxxx, 2019, doi: 10.1016/j.eij.2019.12.002.

[110] R. Rautray and R. Chandra, “Cat swarm optimization based evolutionary framework for multi document summarization,” Physica A, vol. 477, pp. 174–186, 2017, doi: 10.1016/j.physa.2017.02.056.

[111] V. Gupta, L. C. Science, and G. S. Lehal, “A Survey of Text Mining Techniques and Applications,” vol. 1, no. 1, pp. 60–76, 2009.

[112] S. Gupta and G. S.K, “Summarization of Software Artifacts : A Review,” International Journal of Computer Science and Information Technology, vol. 9, no. 5, pp. 165–187, 2017, doi: 10.5121/ijcsit.2017.9512.

[113] K. Joshi, A. Verma, A. Kandpal, S. Garg, R. Chauhan, and R. H. Goudar, “Ontology based Fuzzy Classification of Web Documents for Semantic Information Retrieval,” pp. 1–5, 2013.

[114] M. Mohd, R. Jan, and M. Shah, “Text document summarization using word emb e dding,” Expert Syst Appl, vol. 143, p. 112958, 2020, doi: 10.1016/j.eswa.2019.112958.

[115] S. Mohammed, “Introducing the new JETWI associate editor-in-chief,” Journal of Emerging Technologies in Web Intelligence, vol. 5, no. 1, p. 1, 2013, doi: 10.4304/jetwi.5.1.1-1.

[116] M. Afzal, F. Alam, K. M. Malik, and G. M. Malik, “Clinical Context–Aware Biomedical Text Summarization Using Deep Neural Network: Model Development and Validation,” J Med Internet Res, vol. 22, no. 10, Oct. 2020, doi: 10.2196/19810.

[117] M. Nasr Azadani, N. Ghadiri, and E. Davoodijam, “Graph-based biomedical text summarization: An itemset mining and sentence clustering approach,” J Biomed Inform, vol. 84, pp. 42–58, Aug. 2018, doi: 10.1016/j.jbi.2018.06.005.

[118] R. Mishra et al., “Text summarization in the biomedical domain: A systematic review of recent research,” Journal of Biomedical Informatics, vol. 52. Academic Press Inc., pp. 457–467, Dec. 01, 2014. doi: 10.1016/j.jbi.2014.06.009.

[119] C. Mallick, A. K. Das, J. Nayak, D. Pelusi, and S. Vimal, “Evolutionary Algorithm based Ensemble Extractive Summarization for Developing Smart Medical System,” Interdiscip Sci, vol. 13, no. 2, pp. 229–259, Jun. 2021, doi: 10.1007/s12539-020-00412-5.

[120] C. Gulden et al., “Extractive summarization of clinical trial descriptions,” Int J Med Inform, vol. 129, pp. 114–121, Sep. 2019, doi: 10.1016/j.ijmedinf.2019.05.019.

[121] L. Bing, P. Li, Y. Liao, W. Lam, W. Guo, and R. J. Passonneau, "Abstractive multi-document summarization via phrase selection and merging," ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, vol. 1, pp. 1587–1597, 2015, doi: 10.3115/v1/p15-1153.

[122] M. Moradi, "CIBS: A biomedical text summarizer using topic-based sentence clustering," J Biomed Inform, vol. 88, pp. 53–61, Dec. 2018, doi: 10.1016/j.jbi.2018.11.006.

[123] M. Moradi, "Frequent itemsets as meaningful events in graphs for summarizing biomedical texts," in 2018 8th International Conference on Computer and Knowledge Engineering, ICCKE 2018, Institute of Electrical and Electronics Engineers Inc., Dec. 2018, pp. 135–140. doi: 10.1109/ICCKE.2018.8566651.

[124] M. Moradi and N. Ghadiri, "Quantifying the informativeness for biomedical literature summarization: An itemset mining method," 2017.

[125] E. K. Lee and K. Uppal, "CERC: an interactive content extraction, recognition, and construction tool for clinical and biomedical text," BMC Med Inform Decis Mak, vol. 20, Dec. 2020, doi: 10.1186/s12911-020-01330-8.

[126] M. Moradi and N. Ghadiri, Different approaches for identifying important concepts in probabilistic biomedical text summarization, vol. 84, no. March 2017. 2018. doi: 10.1016/j.artmed.2017.11.004.

[127] E. Davoodijam, N. Ghadiri, M. Lotfi Shahreza, and F. Rinaldi, "MultiGBS: A multi-layer graph approach to biomedical summarization," J Biomed Inform, vol. 116, Apr. 2021, doi: 10.1016/j.jbi.2021.103706.

[128] D. P. Purbawa, Malikhah, R. N. E. Anggraini, and R. Sarno, "Automatic Text Summarization using Maximum Marginal Relevance for Health Ethics Protocol Document in Bahasa," in Proceedings of 2021 13th International Conference on Information and Communication Technology and System, ICTS 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 324–329. doi: 10.1109/ICTS52701.2021.9607951.

[129] S. Rai, S. Chakraverty, and D. K. Tayal, "Supervised Metaphor Detection using Conditional Random Fields," Proceedings of The Fourth Workshop on Metaphor in NLP, no. June, pp. 18–27, 2016.

[130] P. Chen and R. Verma, "A Query-based Medical Information Summarization System Using Ontology Knowledge," 2006. [Online]. Available: www.blackwell-synergy.com/toc/dme/21/12.

[131] K. R. Mckeown, N. Elhadad, and V. Hatzivassiloglou, "Leveraging a Common Representation for Personalized Search and Summarization in a Medical Digital Library."

[132] M. Fiszman and T. C. Rindflesch, "Abstraction Summarization for Managing the Biomedical Research Literature."

[133] M. Fiszman, T. C. Rindflesch, and H. Kilicoglu, "Summarizing Drug Information in Medline Citations."

[134] A. Sarker, D. Mollá, and C. Paris, "Extractive summarisation of medical documents using domain knowledge and corpus statistics," Australasian Medical Journal, vol. 5, no. 9, pp. 478–481, 2012, doi: 10.4066/AMJ.2012.1361.

[135] L. Plaza, M. Stevenson, and A. Díaz, “Improving Summarization of Biomedical Documents using Word Sense Disambiguation,” Association for Computational Linguistics, 2010.

[136] L. Reeve, H. Han, and A. D. Brooks, “BioChain: Lexical Chaining Methods for Biomedical Text Summarization,” 2006.

[137] I. Spasic, S. Ananiadou, J. McNaught, and A. Kumar, “Text mining and ontologies in biomedicine: Making sense of raw text,” Brief Bioinform, vol. 6, no. 3, pp. 239–251, 2005, doi: 10.1093/bib/6.3.239.

[138] P. Chen and R. Verma, “A Query-based Medical Information Summarization System Using Ontology Knowledge,” 2006. [Online]. Available: www.blackwell-synergy.com/toc/dme/21/12.

[139] L. Plaza, M. Stevenson, and A. Díaz, “Improving Summarization of Biomedical Documents using Word Sense Disambiguation,” Association for Computational Linguistics, 2010.

[140] E. Alsentzer et al., “Publicly Available Clinical BERT Embeddings,” Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.03323

[141] H. Saleh, A. Alhothali, and K. Moria, “Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model,” Applied Artificial Intelligence, vol. 37, no. 1, 2023, doi: 10.1080/08839514.2023.2166719.

[142] Y. Liu, “Domain Ontology Concept Extraction Method Based on Text,” pp. 0–4, 2016.

[143] C. Fang, D. Mu, Z. Deng, and Z. Wu, “Word-sentence co-ranking for automatic extractive text summarization,” Expert Syst Appl, vol. 72, pp. 189–195, 2017, doi: 10.1016/j.eswa.2016.12.021.

[144] M. Ghazizadeh and A. Mahmoud, “Semantic similarity assessment of words using weighted WordNet,” 2012, doi: 10.1007/s13042-012-0135-3.

[145] A. Graves, A. R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Oct. 2013, pp. 6645–6649. doi: 10.1109/ICASSP.2013.6638947.

[146] A. Kaur and S. G. Jindal, “Text analytics based severity prediction of software bugs for apache projects,” International Journal of Systems Assurance Engineering and Management, vol. 10, no. 4, pp. 765–782, 2019, doi: 10.1007/s13198-019-00807-8.

[147] S. Dennis and T. Landauer, “Introduction to latent semantic analysis,” Slides from the tutorial, 2003, [Online]. Available: http://lsa3.colorado.edu/papers/LSATutorial.pdf

[148] T. K. Landauer and S. T. Dutnais, “A Solution to Plato ’ s Problem : The Latent Semantic Analysis Theory of Acquisition , Induction , and Representation of Knowledge,” vol. 1, no. 2, pp. 211–240, 1997.

[149] C. Dudschig and B. Kaup, “LSAfun - An R package for computations based on Latent,” pp. 930–944, 2015, doi: 10.3758/s13428-014-0529-0.

[150] E. G. M. Petrakis and G. Varelas, “X-Similarity : Computing Semantic Similarity between Concepts from Different Ontologies,” vol. 4, no. 4, pp. 233–237, 2006.

[151] J. E. Rome and R. M. Haralick, “Towards a Formal Concept Analysis Approach to Exploring Communities on the World Wide Web,” pp. 33–48, 2005.

[152] M. Priya and C. A. Kumar, “A Survey of State of the Art of Ontology Construction and Merging using Formal Concept Analysis,” vol. 8, no. September, 2015, doi: 10.17485/ijst/2015/v8i24/82808.

[153] S. Goyal, “Rapid automatic keyword extraction for extractive summarization of software bug reports of apache projects,” Academia Letters, Jun. 2022, doi: 10.20935/al2473.

[154] S. G. Jindal and A. Kaur, “Automatic Keyword and Sentence based Text Summarization for Software Bug Reports”.

[155] Y. Shang, Y. Li, H. Lin, and Z. Yang, “Enhancing biomedical text summarization using semantic relation extraction,” PLoS One, vol. 6, no. 8, Aug. 2011, doi: 10.1371/journal.pone.0023862.

[156] Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270.

[157] R. Jin, L. Lu, J. Lee, and A. Usman, “Multi-representational convolutional neural networks for text classification,” Comput Intell, vol. 35, no. 3, pp. 599–609, 2019, doi: 10.1111/coin.12225.

[158] A. S. Almasoud et al., “Automated Multi-Document Biomedical Text Summarization Using Deep Learning Model,” Computers, Materials and Continua, vol. 71, no. 2, pp. 5799–5815, 2022, doi: 10.32604/cmc.2022.024556.

[159] Y. Peng, S. Yan, and Z. Lu, “Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets,” Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.05474

[160] Y. Zhang, M. J. Er, R. Zhao, and M. Pratama, “Multiview Convolutional Neural Networks for Multidocument Extractive Summarization,” IEEE Trans Cybern, vol. 47, no. 10, pp. 3230–3242, Oct. 2017, doi: 10.1109/TCYB.2016.2628402.

[161] S. K. Sahu, A. Anand, K. Oruganty, and M. Gattu, “Relation extraction from clinical texts using domain invariant convolutional neural network,” Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.09370

[162] M. Gardner et al., “AllenNLP: A Deep Semantic Natural Language Processing Platform,” Mar. 2018, [Online]. Available: http://arxiv.org/abs/1803.07640

[163] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[164] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html.

[165] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” 2020.

[166] J. Leskovec, M. Grobelnik, and N. Milic-Frayling, “Learning Sub-structures of Document Semantic Graphs for Document Summarization.”

[167] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries.”

[168] L. A. Cabrera-Diego and J. M. Torres-Moreno, “SummTriver: A new trivergent model to evaluate summaries automatically without human references,” Data Knowl Eng, vol. 113, pp. 184–197, Jan. 2018, doi: 10.1016/j.datak.2017.09.001.

[169] A. Nenkova, L. Vanderwende, and K. Mckeown, “A Compositional Context Sensitive Multi-document Summarizer: Exploring the Factors That Influence Summarization,” 2006. [Online]. Available: http://duc.nist.gov

[170] Hassel, Martin. "Exploitation of named entities in automatic text summarization for swedish." NODALIDA’03–14th Nordic Conference on Computational Linguistics, Reykjavik, Iceland, May 30–31 2003.

[171] Baralis, Elena, et al. "Multi-document summarization exploiting frequent itemsets." Proceedings of the 27th annual ACM Symposium on Applied Computing. 2012.

[172] Koupaee, Mahnaz, and William Yang Wang. "Wikihow: A large scale text summarization dataset." arXiv preprint arXiv:1810.09305 (2018)

[173] Saggion, Horacio, and Thierry Poibeau. "Automatic text summarization: Past, present and future." Multi-source, multilingual information extraction and summarization (2013): 3-21.

[174] Antiqueira, Lucas, et al. "A complex network approach to text summarization." Information Sciences 179.5 (2009): 584-599.

[175] P.P.S. Bedi, M. Bala and K. Sharma, “Extractive text summarization for biomedical transcripts using deep dense LSTM-CNN framework”, Expert Systems, vol. 41, no.7, p.p.13490.

[176] P.P.S. Bedi, M. Bala and K. Sharma, “Extractive summarization using concept-space and keyword phrase”, Expert Systems, vol. 39, no. 10, p.p. 13110.

[177] P.P.S. Bedi, M. Bala and K. Sharma, “MLM: Masked Language Modeling Using Deep Learning for Efficient Summarization of Unstructured Data”, In International Conference on Data Analytics & Management, pp. 339-347.

[178] Yongkiatpanich, Chuleepohn, and Duangdao Wichadakul. "Extractive text summarization using ontology and graph-based method." *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*. IEEE, 2019.