IS71130A: Natural Language Processing (2023-24)

A Systematic Literature Review of

Deep Learning Medical Text Summarisation

Sandor Kanda

2024

Table of Contents

List of Figures

List of Tables

Abbreviations

- NLP: Natural Language Processing
- LSTM: Long Short-Term Memory
- CNN: Convolutional Neural Network
- ROUGE: Recall-Oriented Understudy for Gisting Evaluation
- RNN: Recurrent Neural Network
- BERT: Bidirectional Encoder Representations from Transformers
- GPT: Generative Pre-trained Transformer
- ELMo: Embeddings from Language Models
- Bi-LSTM: Bidirectional Long Short-Term Memory
- PICO: Population, Intervention, Comparison, Outcome
- BART: Bidirectional and Auto-Regressive Transformers
- MoCo: Momentum Contrast
- CHQ: Consumer Health Question
- ECL: Entity-driven Contrastive Learning
- TESLEA: Text Simplification using Reinforcement Learning
- SVM: Support Vector Machine
- LDA: Latent Dirichlet Allocation
- OKAM-CGAN: Orthogonal K-means Aligned Multiscale – Conditional Generative Adversarial Network

Abstract

This study systematically reviews the current state of deep learning methods in summarising medical texts. The review examines a total of 20 papers that were published from 2019 to 2024. The main focus of the analysis is on the types of input data, pre-processing methods, algorithms, and evaluation metrics utilised in the previous studies. The studies utilise a range of deep learning architectures, such as LSTM, transformer, and hybrid models, and employ datasets such as clinical notes, discharge summaries, and biomedical literature. The papers showcase the efficacy of deep learning methods in producing concise and informative summaries, yielding promising outcomes regarding ROUGE scores and human evaluation.

Nevertheless, there are still obstacles to overcome when it comes to managing vocabulary specific to a particular field, guaranteeing the correctness of information, and adjusting to various clinical environments. The review emphasises the capacity of deep learning-based summarisation to improve information retrieval and decision-making in healthcare systems. Potential areas for future research encompass integrating domain expertise, enhancing model interpretability, and creating more extensive evaluation criteria.

Keywords:
medical text summarisation, deep learning, natural language processing, clinical decision support, information retrieval, systematic review

Introduction

In recent times, the healthcare sector has seen a notable upsurge in the quantity of patient data, mostly influenced by reasons such as the expanding patient population and the escalating intricacy of medical ailments. The proliferation of data has brought advantageous prospects and formidable obstacles for healthcare practitioners, especially within the framework of national healthcare systems. The abundance of patient data is a tremendous opportunity to gain insights that have the potential to enhance patient outcomes and propel advancements in medical research. Nevertheless, the substantial data has also resulted in operational limitations, technological complexities, and extended waiting periods for patients and healthcare practitioners. The healthcare system's significant obstacle is the need for interconnection across extensive databases. Every patient visits the hospital, creating a fresh record and adding to the continuously growing patient data collection. According to Digital Health, the estimated number of primary care records exceeds 55 million (Owen Hughes, 2024). The lack of a cohesive system often leads to establishing isolated data repositories, posing challenges for healthcare practitioners in efficiently accessing and using patient information. Fragmentation in healthcare systems provides obstacles to the seamless provision of treatment and presents technological complexities in data storage, retrieval, and analysis.

Managing the increasing amount of patient data relies heavily on the technological architecture of healthcare systems. Using conventional database indexing techniques may result in increased loading durations for information retrieval and the possibility of system unavailability, intensifying the capacity challenges healthcare providers encounter. Clinicians often allocate significant effort to locating pertinent information within extensive databases.

This task becomes notably arduous when confronted with unstructured textual material, such as clinical notes and medical reports. The quality of patient data significantly influences the effectiveness of healthcare systems. Untidy textual data often exposes data quality concerns, such as inaccuracies or omissions, which might impede the precise analysis of patient records. It is crucial to tackle these difficulties related to data quality to guarantee that healthcare providers may get dependable and thorough patient information. To address these difficulties, there is an increasing inclination towards using sophisticated natural language processing (NLP) technologies to summarise medical content. These technologies can optimise information retrieval and allow doctors to swiftly access the most relevant data points by automatically providing succinct and useful summaries of patient records. Condensed text may greatly decrease healthcare workers' time looking through lengthy patient records, enabling them to make better-informed judgements and provide prompt treatment, even in current backlogs. This study comprehensively analyses existing research on deep learning methods for summarising medical content and their possible implications for healthcare systems. The aim is to ascertain the primary obstacles, approaches, and prospective avenues within this domain through systematically analysing the current state of the research. The study will use deep learning models to condense several forms of medical textual data, including clinical notes, discharge summaries, and medical reports.

The primary objective of doing a thorough literature study is to provide a full understanding of the efficacy, constraints, and pragmatic factors associated with the use of deep learning-based text summarisation in healthcare environments. Additionally, an examination will be conducted on the prospective advantages of employing these methodologies in expanding data accessibility, mitigating waiting periods, and augmenting the quality of patient treatment inside the framework of national healthcare systems.

Proposal

The review aims to identify the types of medical text data used as input for deep learning-based summarisation models, such as clinical notes, discharge summaries, and medical reports. It will investigate the pre-processing techniques, algorithms, and deep learning architectures employed in the existing research on medical text summarisation. Furthermore, the review will examine the performance evaluation metrics and methodologies used to assess the effectiveness of deep learning-based summarisation models in the medical domain. Lastly, it will highlight the challenges, limitations, and future research directions in deep learning for medical text summarisation.

The proposed systematic literature review holds significant importance in healthcare systems facing the challenges of growing patient data, capacity issues, and the need for efficient information retrieval. The review will contribute to the existing body of knowledge by providing an overview of the current landscape of deep learning techniques applied to medical text summarisation, enabling researchers and practitioners to understand this field's state of the art. It will identify the most effective and promising deep learning approaches for generating concise and informative summaries of medical text data, which can aid in clinical decision-making and improve patient care. The review will also highlight the potential benefits of deep learning-based summarisation in reducing waiting times, enhancing data accessibility, and streamlining information retrieval within healthcare systems.
Moreover, it will inform future research directions and guide the development of advanced deep-learning models tailored to medical text summarisation's specific needs and challenges.

The systematic literature review will follow a rigorous methodology to ensure a comprehensive and unbiased analysis of the existing research. It will involve defining the search strategy, which includes identifying relevant databases and search engines, such as PubMed, IEEE Xplore, ACM Digital Library, and Google Scholar. The search queries will be formulated using keywords related to deep learning, medical text summarisation, and relevant synonyms. Inclusion and exclusion criteria will be set based on publication date and study design. Next, it will conduct the literature search, which involves executing the search queries in the selected databases and search engines, screening the retrieved articles based on the defined inclusion and exclusion criteria, and removing duplicates and irrelevant studies. The next step will focus on data extraction and synthesis. The extracted data will be structured and synthesised to identify patterns, trends, and key findings related to the research questions.

Finally, it will involve assessing the quality and methodological rigour of the included studies using established quality assessment tools or checklists. The risk of bias and the reliability of the reported results will be evaluated. Finally, the findings of the systematic literature review will be summarised clearly and concisely. The results will be presented as tables, graphs, and narrative summaries.

Methodology

The primary objective of this systematic literature review is to extensively analyse and synthesise the current state of medical text summarisation. This review will be conducted in order to achieve this objective. To answer the following research questions, the evaluation will be performed:

Research Question 1: What input data types and summarisation techniques have been employed in the existing research on deep learning for clinical text summarisation?

Research Question 2: What algorithms have been utilised in the studies?

Research Question 3: How has the performance of the text summarisation models been validated and evaluated?

By addressing these research questions, the review seeks to provide insights into the current landscape of deep learning techniques applied to medical text summarisation, identify promising approaches and challenges, and guide future research efforts.
The proposed technique encompasses essential stages, including literature review and data selection, data extraction, quality evaluation, and data synthesis. The literature search will use prominent bibliographic databases, such as IEEE Xplore, ACM Digital Library, Web of Science (Primary Collection), and Google Scholar. These databases are selected for their extensive coverage of research in computer science, engineering, and healthcare informatics. The search strategy will use keywords and Boolean operators to identify relevant studies. The search query will be constructed as follows:

((deep learning) OR (machine learning) AND (clinical text) OR (medical text) AND (summariz) OR (summaris*) NOT (literature) NOT (review) NOT (comparative) NOT (review))*

The performed search uses each database's advanced search features, allowing for the targeting of specific fields such as paper titles and abstracts. This approach helps narrow the search results to studies most relevant to the research questions. A set of inclusion and exclusion criteria is defined to ensure the selection of appropriate studies for the systematic review.

The inclusion criteria species the characteristics that studies must possess to be considered for the review, while the exclusion criteria outline the attributes that disqualify studies from being included. The inclusion criteria consist of studies that apply deep learning techniques for medical text summarisation published in English.

Furthermore, studies must include the dataset source, key findings, limitations, and future directions. Studies that did not involve deep learning techniques for medical or clinical text were excluded. Additionally, literature reviews, comparative analyses, or survey papers were part of the exclusion criteria, including those needing more methodology, input, performance evaluation or languages other than English. The assessment considers factors such as the clarity of research objectives, the appropriateness of the method, the data analysis's robustness, and the reported results' reliability.

The extracted data synthesis uses a narrative approach, identifying the findings across studies, patterns, trends, and key themes related to the research questions. Furthermore, it will also highlight the strengths and limitations of the included studies and discuss the implications of the findings for healthcare systems and future research directions.

Literature Review

The rapid advancement of deep learning techniques has revolutionised various domains, including natural language processing (NLP) and healthcare. One area that has garnered significant attention is the application of deep learning methods for medical text summarisation. Medical text summarisation aims to generate concise and informative summaries from vast amounts of medical text data, such as clinical notes, electronic health records (EHRs), biomedical literature, and medical reports [(Guo et al., 2022).

Deep learning approaches have shown promising results in medical text summarisation tasks, outperforming traditional rule-based and statistical methods (Sun & Platoš, 2024a).

Extractive approaches focus on selecting relevant sentences or phrases from the original text to form a summary. At the same time, abstractive methods generate novel sentences that capture the essence of the input text (Guo et al., 2022). Commonly used evaluation metrics include ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores, which measure the overlap between the generated summary and reference summaries (Xu et al., 2022). Other metrics, such as accuracy and human evaluation, are also employed to assess the quality and readability of the generated summaries (Sibunruang & Polpinij, 2018).

This literature review aims to provide a comprehensive overview of the current state of research in deep learning techniques for medical text summarisation. By categorising and analysing selected papers based on their methodologies, application domains, and evaluation metrics, we seek to identify this field's strengths, limitations, and future directions. The insights gained from this review can inform future research efforts and contribute to developing more effective and efficient medical text summarisation models. This critical analysis aims to provide a valuable resource for researchers and practitioners working on the intersection of deep learning and medical text summarisation.

Table 1. Summary of the published studies that included Clinical Text Summarisation

Study Reference	Year	Dataset	Method	ROUGE-1	ROUGE-2	ROUGE-L	Accuracy
Sun, Y. & Platoš, J.	2024	Gigaword, CNN/Daily Mail	Hierarchical attention, Pointer-generator, RL	0.4036	0.1914	0.3855	
Tang, Gao, Zhang, & Wang	2023	Newsroom, CNN/DM, XSum	URLComSum	0.3087	0.1397	0.252	
Chen et al.	2020	NTUH-iMD	AlphaBERT	0.712	0.533	0.693	
Phatak et al.	2022	Medical	TESLEA	0.39	0.11		
Xu et al.	2022	Medical	COSSUM	0.57193	0.45936	0.56696	0.8256
Lu et al.	2024	MeQSum, CHQ-Summ, iCliniq, HealthCareMagic	ECL	0.4521	0.26585	0.42045	
Afsharizadeh et al.	2020	COVID-19 Corpus	RNN-based Model	0.53383	0.25357	0.34383	
Guo et al.	2021	Cochrane Database of Systematic Reviews	BART + CNN/DM + PubMed	0.5302	0.2206	0.5024	
Balouch & Hussain	2023	MIMIC III	Biobart-V2	0.6634	0.612	0.6942	
Sibunruang & Polpinij	2024	MEDLINE	Text Summarization with NLP & SVM				0.9541
Kedzie, McKeown, & Daume III	2019	CNN, DailyMail, New York Times, DUC, Reddit, AMI, PubMed	Seq2Seq		0.256		
Lin, Guo, Dong, Lyu, Xu, & Chen	2023	LCSTS	Improved Transformer	0.3003	0.1789	0.2785	
Tan,Kieuvongngam, & Niu	2020	CORD-19	BERT and GPT-2				
Nafees Muneera M. & Sriramya P.	2023	CNN/Daily Mail, Edmunds	OKAM-CGAN	0.4765	0.46	0.422	0.22
Srivastava, R., Singh, P., Rana,	2022	Wikihow, CNN/Daily Mail, DUC2002 Corpus	LDA-based topic modelling, K-Medoids clustering	0.4268	0.1989	0.3850	
Searle et al.	2023	MIMIC-III, KCH	Ensemble model with clinical concept guidance		0.1255	0.3105	
Afzal et al., 2020	2020	Intracranial aneurysm data	Deep neural network, Word embedding, Semantic similarity, Bi-LSTM				0.94205
Bedi et al.	2023	Medical Transcripts from MT Samples	Deep Dense LSTM-CNN Framework	0.935			
Sarker et al.	2020	EBM Corpus, COVID-19 Articles	Word Embedding-based Maximal Marginal Relevance (MMR), Sentence Length, and Position–			0.166	
Afzal et al.	2020	Intracranial Aneurysm Studies	Biomed-Summarizer (Deep Neural Network, LSTM, AdaBoost-MLP)				0.94205

As shown in Table 1, this literature review covers twenty recent papers that propose and evaluate various text summarisation approaches for biomedical and clinical applications.

The study papers are published between 2019 and 2024. Therefore, the analysis consists of five years' worth of research. Several papers focus on improving abstractive summarisation using deep learning techniques. Sun Platoš (Sun & Platoš, 2024b) combine hierarchical attention, pointer-generator networks, and reinforcement learning to generate more semantically consistent summaries. (Xu et al., 2022) developed COSSUM, a prompting-based unified approach to generating structured summaries from medical conversations. (Balouch & Hussain, 2023) fine-tuned the Biobart-V2 transformer model and pre-trained on biomedical text to summarise radiology reports.

Other work explores extractive summarisation, identifying the most relevant sentences. (Chen et al., 2020) propose AlphaBERT, a modified BERT model using character-level tokens, to summarise hospital discharge diagnoses efficiently. (Kedzie et al., 2019) analysed position bias in news summarisation and found that more complex models do not necessarily outperform simpler ones. (Lin et al., 2023) incorporate keyword information into transformer models to improve extractive summaries.

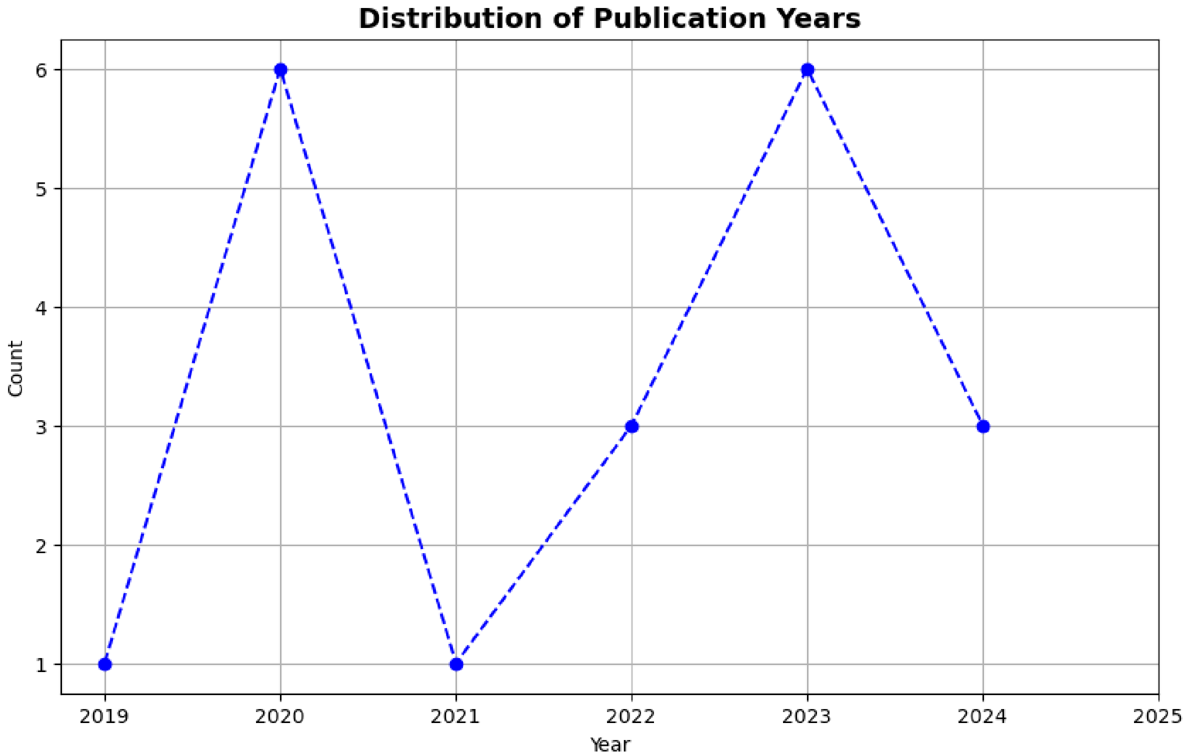
Some studies target specific domains like COVID-19 research. (Tang et al., 2023) apply BERT for extractive and GPT-2 for abstractive summarisation of COVID-19 articles. (Afsharizadeh et al., 2021) developed an RNN-based model for abstractive summaries of this rapidly expanding literature. Readability and accessibility are another focus. (Phatak et al., 2022) Introduces TESLEA, a reinforcement learning approach that simplifies complex medical texts while preserving quality. (Guo et al., 2021) generate lay language summaries of biomedical reviews using BART models pre-trained on CNN/DailyMail and PubMed.

Other approaches include contrastive learning with medical entities (W. Lu et al., 2024), unsupervised dual-agent reinforcement learning for efficient extractive/compressive summaries (Tang et al. 2023), clinical context modelling and evidence quality assessment (Afzal et al. 2020), and lightweight heuristic-based methods for low-resource settings (Sarker et al. 2020). Key evaluation metrics are ROUGE scores measuring n-gram overlap with reference summaries and readability measures like Flesch-Kincaid grade level. The studies demonstrate statistically significant improvements over the common-sense baselines.

Critical Analysis and Synthesis

The field of medical text summarisation has witnessed significant advancements in recent years, driven by the application of deep learning techniques. By categorising and analysing selected papers based on their methodologies, application domains, and evaluation metrics, the goal is to identify trends, gaps, and future research directions in this rapidly evolving field.

Figure 1. Line Graph of Publication Year Distribution



In Figure 2 above, The line graph shows the distribution of publication years for the selected papers in the literature review, ranging from 2019 to 2024. The graph indicates an increasing trend in the number of publications over the years, with a notable peak in 2023.

Method-based Analysis

Afsharizadeh et al. developed an RNN-based deep learning model with a coreference resolution procedure that embedded vectors, resulting in improved performance. The dataset used for this research is the Covid Open Research Dataset, containing articles related to Covid-19. The proposed method achieves the highest ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4 (as shown in Table 1)scores compared to all other methods, indicating better performance in capturing unigram, bigram, longest common subsequence, and skip-bigram statistics.
Future research should explore applicability to diverse clinical text sources, incorporate domain knowledge, and evaluate using human evaluations. The findings have implications for the rapid dissemination of COVID-19 research findings. The synthesis will also highlight the included studies' strengths and limitations and discuss the findings' impact on healthcare systems and future research directions.

LSTM-based models automatically condense lengthy text documents into shorter summaries while retaining the core meaning. Chen et al. (2020) built a model that used 258,050 discharge diagnoses from the NTUH-iMD as input data and employed an extractive summarisation technique based on a BERT-like model with character-level tokenisation. The performance of the text summarisation models was evaluated using ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L) and critique scores from physicians. The models were compared against each other, and experienced doctors labelled reference summaries. Bedi et al. (2023) use biomedical algorithms to capture semantic nuances and relevant features for summarisation. The input data types utilised in the paper include LSTM and CNN frameworks, often combined with ELMo for contextualised word embeddings. Model performance has been predominantly validated using ROUGE scores, with one study reaching a high average of 93.5%. This indicates close alignment with human-generated summaries and reflects the models' accuracy in extracting key information. The paper by Afzal et al. (2020) introduces Biomed-Summarizer, a novel framework for biomedical text summarisation focused on intracranial aneurysms. It employs quality-aware PICO-based summarisation, semantic enrichment via medical ontologies, and contextual parameters. Algorithms include a multilayer perceptron for quality recognition and a Bi-LSTM network for PICO classification. The evaluation shows high accuracy rates (95.41% for quality recognition, 93% for PICO classification) and significant improvements in semantic similarity.

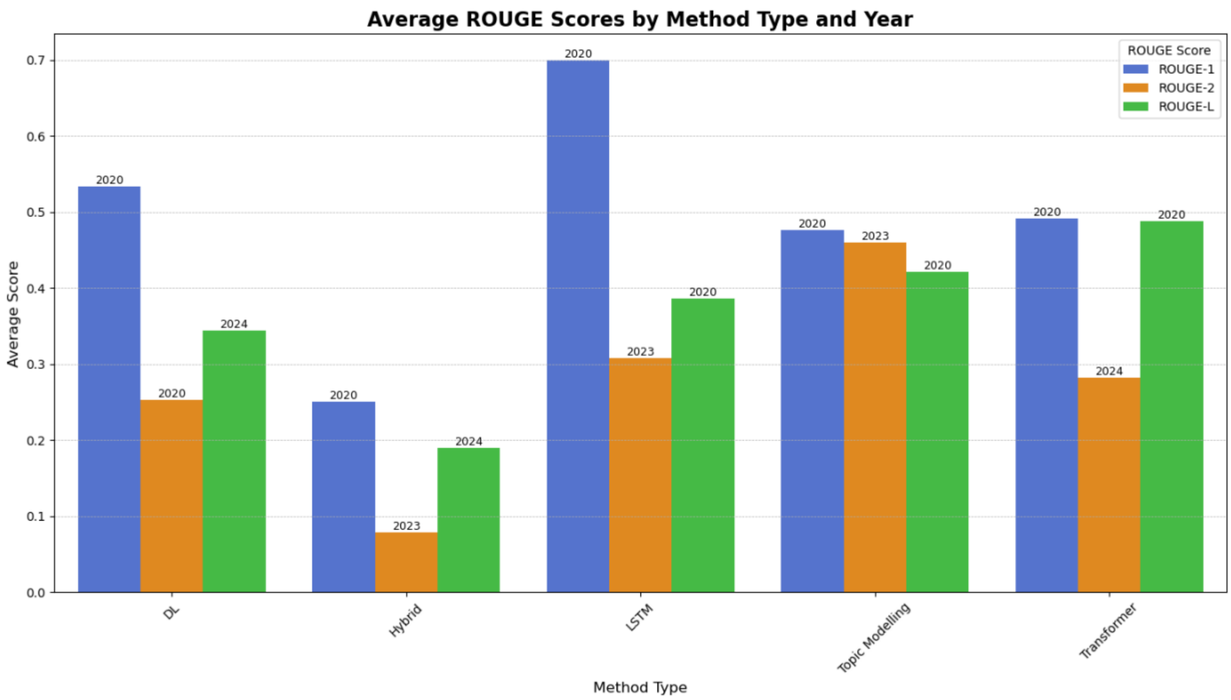
Transformer-based models have emerged as a powerful approach to text summarisation. Their encoder-decoder architecture and self-attention mechanism effectively capture long-range dependencies within text, leading to more comprehensive and informative summaries. (Vaswani et al., 2023) The data types used by Guo et al. include high-quality abstract pairs from biomedical scientific reviews, combining professional and plain language versions. Summarisation techniques involve state-of-the-art extractive and abstractive models with data augmentation strategies. Algorithms utilised include Oracle extractive, BERT extractive, pointer-generator, and BART models, with intermediate pre-training on CNN/DM and PubMed abstracts for domain-specific language exposure. Performance validation involved automated metrics (ROUGE scores, readability tests) and human assessment, focusing on grammaticality, meaning preservation, understandability, and correctness of key information. Balouch & Hussain (2023) This paper used the MIMIC-III dataset for radiology reports summarisation, employing NLP pre-processing and transformer models. Algorithms utilised include transformer models like Biobart-V2, with fine-tuning on the MIMIC III dataset for summarisation tasks. Lin et al. (2023) used Chinese short texts from the LCSTS dataset. Techniques used are deep learning and an improved Transformer model incorporating keyword information. The study utilised the Transformer model enhanced with a keyword information module for text summarisation, employing algorithms like TextRank for keyword extraction. Performance validated using the ROUGE metric on the LCSTS dataset, showing the model improved summarisation quality when incorporating keyword information.

Hybrid summarisation methods combine the strengths of both extractive and abstractive summarisation techniques to produce summaries that are faithful to the original text while still being fluent and concise. (G. Lu et al., 2023) Sun & Platoš (2024) input data consists of English sentences from the Gigaword dataset and news-summary pairs from CNN/Daily Mail. The summarisation technique uses an improved encoder-decoder model with hierarchical attention, a pointer-generator network, and reinforcement learning. The study utilised seq2seq models combined with a hierarchical attention mechanism, pointer-generator networks for handling out-of-vocabulary words, and multi-objective reinforcement learning to optimise summary generation. The performance of summarisation models was validated using ROUGE, METEOR, and BERTScore metrics.

Additionally, a human evaluation assessed fluency, informativeness, faithfulness, and conciseness, with the model achieving scores close to GPT-3. Tang et al. (2023) The paper employs an unsupervised dual-agent reinforcement learning method for compressive text summarisation, including extractive and compressive techniques.

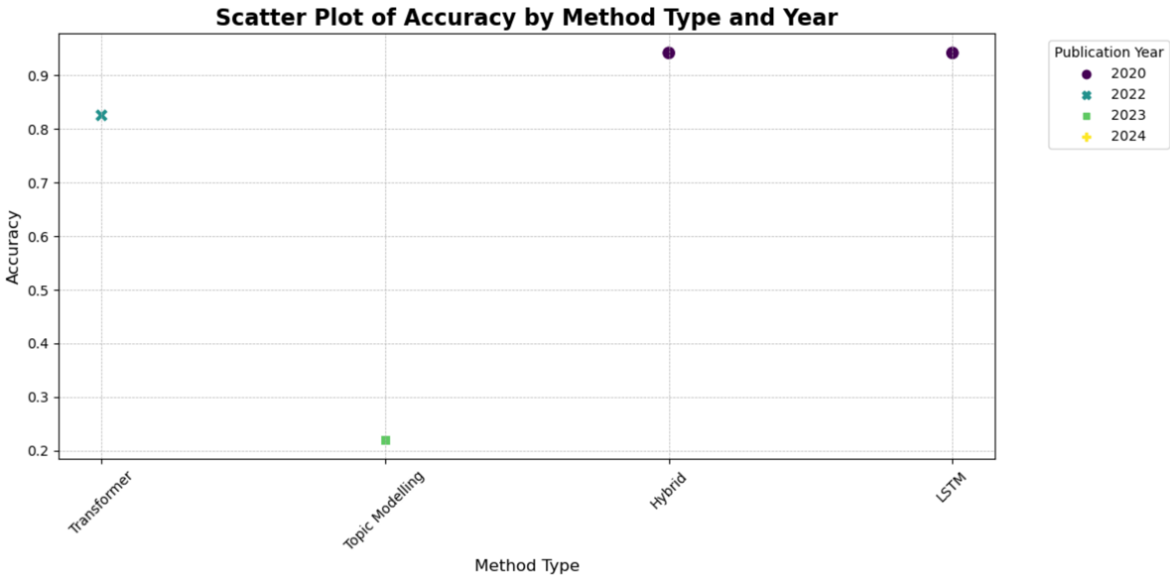
Input data includes documents from which the extractor agent selects salient sentences, and then the compressor agent compresses these sentences by selecting salient words. The algorithms include a hierarchical multi-head attentional Bi-LSTM model for sentence representation and a pointer network for extracting salient sentences and words. Xu et al. (2022) The paper focuses on consumer health questions (CHQs) as input data. It employs an entity-driven contrastive learning (ECL) technique for summarisation, which leverages medical entities within CHQs to generate concise summaries. Algorithms utilised include BART for basic summarisation and MoCo for contrastive learning, enabling the generation of hard negative samples by altering medical entities in summaries. : Model performance is validated and evaluated using ROUGE metrics, with the ECL method outperforming existing approaches across multiple datasets, demonstrating improvements in summarisation accuracy. Model performance is validated and evaluated using ROUGE metrics, with the ECL method outperforming existing approaches across various datasets, demonstrating improvements in summarisation accuracy.

Figure 2. Bar chart of the ROUGE Evaluation metrics based on the methods



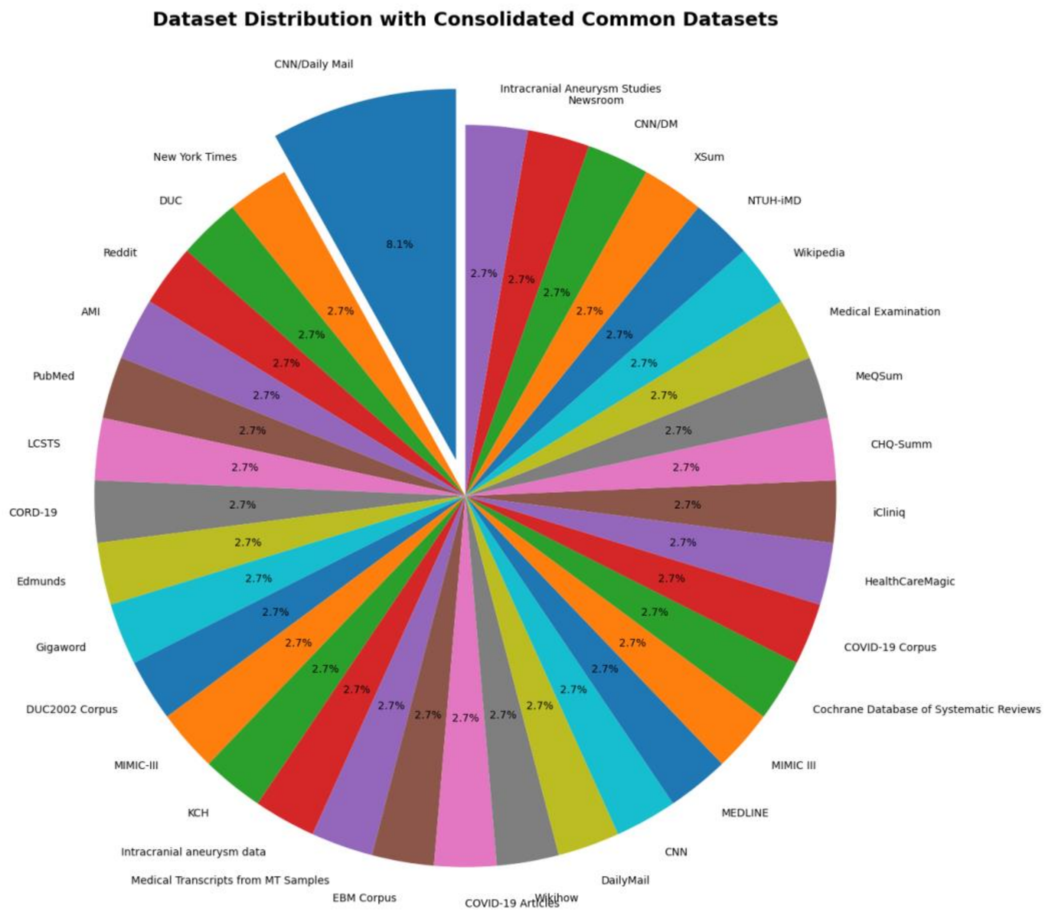
This graph represents the evolution of text summarisation methods over time, comparing average ROUGE scores across different method types. Each bar denotes the mean ROUGE-1, ROUGE-2, and ROUGE-L scores achieved by method types in studies published in years. LSTM produced the highest ROUGE-1 (0.712), the Topic Modelling method produced the highest Rouge-2 (0.46), and Transformers produced the highest ROUGE-L (0.56696). As some papers had multiple models evaluated, the mean of each ROUGE value was calculated.

Figure 3. Scatter Plot of The Accuracy results



The scatter plot in Figure 3 maps the accuracy of different text summarisation methods by type and annotated by publication year. The hybrid (0.94205) and LSTM (0.94205) methods achieved higher accuracy than the transformers (0.8256) and topic modelling (0.22) methods.

Figure 4. Pie Chart of the Dataset Distribution amongst the selected papers



The pie chart in Figure 4 shows a diverse range of datasets employed in text summarisation studies, with each slice indicating the dataset's frequency of use. This signifies the breadth of sources used to enhance summarisation models. The most commonly used dataset is CNN/DailyMail (8.1%).

The current trajectory in the domain of text summarisation is focused on tackling the intricate issues related to semantic representation, domain adaptation, and model efficiency. A significant trend is observed in adopting unsupervised learning and reinforcement learning techniques, which aim to decrease dependence on extensive annotated datasets and enhance the ability to replicate human judgement. Furthermore, there is a notable emphasis on improving the interpretability and explainability of models, specifically within the domain of medical information, where precision holds utmost importance. Furthermore, hybrid models incorporate extractive and abstractive summarisation techniques to enhance the equilibrium between content relevance and the summary's brevity. Incorporating domain-specific knowledge, such as medical ontologies, to manage specialised vocabularies and effectively enhance summaries' accuracy is becoming increasingly significant. In addition, a growing body of research is focused on using character-level and token-level representations to improve computational efficiency in resource-limited settings, such as hospital information systems. Furthermore, there is a growing demand for more comprehensive assessment criteria that surpass ROUGE scores and consider factors such as factual precision, legibility, and even user engagement feedback to enhance iterative model enhancement. The primary objective is to create summarisation systems that can adapt to different domains, effectively handle linguistic intricacies, and efficiently handle substantial amounts of text while upholding high accuracy and user confidence.

Conclusion

The review examined twenty recent papers published from 2019 to 2024, encompassing various methodologies, datasets, and evaluation metrics. The results illustrate the notable progress achieved in this domain, where diverse deep learning structures, including LSTM-based models, transformer-based models, and hybrid approaches, are used to produce precise and informative summaries of medical text data. The review emphasises deep learning's capacity to tackle the difficulties linked to increasing medical text data, including clinical notes, discharge summaries, and biomedical literature. These techniques can help healthcare professionals access and use important information more efficiently by creating short and relevant summaries. This ultimately leads to better patient care and decision-making. Nevertheless, the review highlights various constraints and difficulties that must be resolved in future investigations. A significant drawback of this review is the crucial need for a structured screening process for the included studies, which could lead to incorporating studies with differing degrees of quality and relevance. Due to time constraints, the quality assessments of the selected papers were not completed. However, it could have prevented non-medical datasets from leaking into the review. Future evaluations should incorporate a stringent vetting procedure to guarantee the inclusion of exclusively top-notch and pertinent research.

The review highlights the necessity for broader evaluation criteria beyond the commonly employed ROUGE scores. However, ROUGE scores offer a quantitative assessment of the similarity between generated summaries and the ground truth.

Subsequent investigations should examine supplementary assessment criteria considering factual precision, comprehensibility, and user involvement to evaluate summarisation models' performance comprehensively.
One important challenge highlighted in the review is the requirement for successful domain adaptation and integrating domain-specific knowledge in medical text summarisation. Medical text data frequently includes specialised terminologies, abbreviations, and intricate associations among medical ideas. By incorporating medical ontologies and utilising domain-specific knowledge, the accuracy and relevance of the generated summaries can be enhanced. Future research should prioritise the development of techniques that can efficiently incorporate domain knowledge into the summarisation process.

The review also emphasises the shift towards unsupervised learning and reinforcement learning methods in medical text summarisation. These approaches seek to decrease the dependence on extensively annotated datasets and empower the models to acquire knowledge from unlabelled data and user feedback. Future research should conduct in-depth investigations into these techniques to develop more efficient and adaptable summarisation models capable of managing the continuously increasing amount of medical text data.

This systematic literature review offers valuable insights into the present condition and future prospects of deep learning methods for summarising medical texts. Although there has been notable advancement, there is still scope for enhancement in evaluating metrics, adapting to different domains, and integrating domain-specific knowledge. Future research should prioritise tackling these challenges to create summarisation systems that are more precise, efficient, and user-friendly. These systems should effectively assist healthcare professionals in their daily practice.

Acknowledgements

- During the systematic review, the author used ChatGPT v4 for the data synthesis: The research papers were uploaded onto chat.openai.com, and the chatbot was prompted to do the following:

    - Summarise the paper's key takeaways
    - Provide details on the future scope in bullet points

  The author's ownership of the Systematic Literature Review of Deep Learning Medical Text Summarisation remains.

- The author sincerely appreciates all researchers and participants whose studies have contributed to this comprehensive literature review.

References

Afsharizadeh, M., Ebrahimpour-Komleh, H., & Bagheri, A. (2021). Automatic Text Summarization of COVID-19 Research Articles Using Recurrent Neural Networks and Coreference Resolution. *Frontiers in Biomedical Technologies*. https://doi.org/10.18502/fbt.v7i4.5321

Balouch, B. A. K., & Hussain, F. (2023). A Transformer based approach for Abstractive Text Summarization of Radiology Reports. *International Conference on Applied Engineering and Natural Sciences*, *1*(1), 476–486. https://doi.org/10.59287/icaens.1042

[Bedi2023] Bedi, J., Bahl, A., & Natu, A. (2023). (n.d.). *Extractive text summarization for biomedical transcripts using deep dense LSTM-CNN framework. Multimedia Tools and Applications, 82(2), 2873-2889.*

Chen, Y.-P., Chen, Y.-Y., Lin, J.-J., Huang, C.-H., & Lai, F. (2020). Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT):

Development and Performance Evaluation. *JMIR Medical Informatics*, *8*(4), e17787. https://doi.org/10.2196/17787

Guo, Y., Qiu, W., Wang, Y., & Cohen, T. (2021). Automated Lay Language Summarization of Biomedical Scientific Reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(1), 160–168. https://doi.org/10.1609/aaai.v35i1.16089

Guo, Y., Qiu, W., Wang, Y., & Cohen, T. (2022). *Automated Lay Language Summarization of Biomedical Scientific Reviews* (arXiv:2012.12573). arXiv. http://arxiv.org/abs/2012.12573

Kedzie, C., McKeown, K., & Daume III, H. (2019). *Content Selection in Deep Learning Models of Summarization* (arXiv:1810.12343). arXiv. http://arxiv.org/abs/1810.12343

Lin, J., Guo, X., Dong, C., Lyu, C., Xu, L., & Chen, Z. (2023). Text Summarization Generation Based on Improved Transformer Model. *2023 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 0831–0836. https://doi.org/10.1109/DASC/PiCom/CBDCom/Cy59711.2023.10361507

Lu, G., Larcher, S. B., & Tran, T. (2023). *Hybrid Long Document Summarization using C2F-FAR and ChatGPT: A Practical Study* (arXiv:2306.01169). arXiv. http://arxiv.org/abs/2306.01169

Lu, W., Wei, S., Peng, X., Wang, Y.-F., Naseem, U., & Wang, S. (2024). Medical Question Summarization with Entity-driven Contrastive Learning. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 3652160. https://doi.org/10.1145/3652160

Owen Hughes. (2024, March 29). *NHS data worth £9.6bn per year, says Ernst & Young*. https://www.digitalhealth.net/2019/07/nhs-data-worth-9-6bn-per-year-says-ernst-young/

Phatak, A., Savage, D. W., Ohle, R., Smith, J., & Mago, V. (2022). Medical Text Simplification Using Reinforcement Learning (TESLEA): Deep Learning–Based Text Simplification Approach. *JMIR Medical Informatics*, *10*(11), e38095. https://doi.org/10.2196/38095

Sibunruang, C., & Polpinij, J. (2018). Finding Clinical Knowledge from MEDLINE Abstracts by Text Summarization Technique. *2018 International Conference on Information Technology (InCIT)*, 1–6. https://doi.org/10.23919/INCIT.2018.8584867

Sun, Y., & Platoš, J. (2024a). Abstractive text summarization model combining a hierarchical attention mechanism and multiobjective reinforcement learning. *Expert Systems with Applications*, *248*, 123356. https://doi.org/10.1016/j.eswa.2024.123356

Sun, Y., & Platoš, J. (2024b). Abstractive text summarization model combining a hierarchical attention mechanism and multiobjective reinforcement learning. *Expert Systems with Applications*, *248*, 123356. https://doi.org/10.1016/j.eswa.2024.123356

Tang, P., Gao, J., Zhang, L., & Wang, Z. (2023). *Efficient and Interpretable Compressive Text Summarisation with Unsupervised Dual-Agent Reinforcement Learning* (arXiv:2306.03415). arXiv. http://arxiv.org/abs/2306.03415

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Xu, S., Wan, X., Hu, S., Zhou, M., Xu, T., Wang, H., & Mi, H. (2022). COSSUM: Towards Conversation-Oriented Structured Summarization for Automatic Medical Insurance

Assessment. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4248–4256. https://doi.org/10.1145/3534678.3539116