**Domain-Specific Language Model for Medical Dialogue Summarization**

Aizaz Ahmad

Submitted in partial fulfilment of   
the requirements of Edinburgh Napier University   
for the Degree of  
MSc Computing with Professional Placement

School of Computing

April 2024

**MSc dissertation check list**

|  |  |  |
| --- | --- | --- |
| **Milestones** | **Date of completion** | **Target deadline** |
| Proposal | 18 October 2023 |  |
| Initial report | 08 December 2023 |  |
| Full draft of the dissertation | 15 April 2024 | 2 weeks before final deadline |

|  |  |  |  |
| --- | --- | --- | --- |
| **Learning outcome** | **The markers will assess** | **Pages[[1]](#footnote-1)** | **Hours spent** |
| **Learning outcome 1**  Conduct a literature search using an appropriate range of information sources and produce a critical review of the findings*.* | \* Range of materials; list of references  \* The literature review/exposition/background information chapter | 15-28 | 220 hours |
| **Learning outcome 2**  Demonstrate professional competence by sound project management and (a) by applying appropriate theoretical and practical computing concepts and techniques to a non-trivial problem, or (b) by undertaking an approved project of equivalent standard. | \* Evidence of project management (Gantt chart, diary, etc.)  \* Depending on the topic: chapters on design, implementation, methods, experiments, results, etc. | 29-47 | 250 hours |
| **Learning outcome 3**  Show a capacity for self-appraisal by analysing the strengths and weakness of the project outcomes with reference to the initial objectives, and to the work of others. | \* Chapter on evaluation (assessing your outcomes against the project aims and objectives)  \* Discussion of your project's output compared to the work of others. | 48-53 | 80 hours |
| **Learning outcome 4**  Provide evidence of the meeting learning outcomes 1-3 in the form of a dissertation which complies with the requirements of the School of Computing both in style and content. | \* Is the dissertation well-written (academic writing style, grammatical), spell-checked, free of typos, neatly formatted.  \* Does the dissertation contain all relevant chapters, appendices, title and contents pages, etc.  \* Style and content of the dissertation. | | 70 hours |
| **Learning outcome 5**  Defend the work orally at a viva voce examination. | \* Performance  \* Confirm authorship | | 1 hour |

Have you previously uploaded your dissertation to Turnitin? Yes

Has your supervisor seen a full draft of the dissertation before submission? Yes

Has your supervisor said that you are ready to submit the dissertation? Yes

Authorship Declaration

I, Aizaz Ahmad, confirm that this dissertation and the work presented in it are my own achievement.

Where I have consulted the published work of others this is always clearly attributed;

Where I have quoted from the work of others the source is always given. With the exception of such quotations this dissertation is entirely my own work;

I have acknowledged all main sources of help;

If my research follows on from previous work or is part of a larger collaborative research project I have made clear exactly what was done by others and what I have contributed myself;

I have read and understand the penalties associated with Academic Misconduct.

I also confirm that I have obtained **informed consent** from all people I have involved in the work in this dissertation following the School's ethical guidelines

Signed: Aizaz Ahmad

Date: 15 April 2024

Matriculation no: 40611006

General Data Protection Regulation Declaration

Under the General Data Protection Regulation (GDPR) (EU) 2016/679, the University cannot disclose your grade to an unauthorised person. However, other students benefit from studying dissertations that have their grades attached.

Please sign your name below *one* of the options below to state your preference.

The University may not make this dissertation available to others.

Aizaz Ahmad

# Abstract

Efficient clinical documentation is crucial for maintaining continuity of care, processing insurance claims, and ensuring legal compliance in healthcare. However, conventional practices that rely on manual note-taking by clinicians during patient visits can be time-consuming and divert valuable resources away from direct patient care. This research addresses the pressing need for automating clinical note generation by developing a domain-specific language model tailored for summarizing medical dialogues between patients and healthcare providers.

The proposed approach utilizes a sequence-to-sequence encoder-decoder architecture with customized design choices to capture the nuances of medical conversations and clinical note-taking. A stacked LSTM encoder, coupled with an attention mechanism, is employed to extract relevant information from dialogue transcripts, while a decoder LSTM generates concise summaries akin to professional clinical notes.

The model was trained and evaluated on the Ambient Clinical Intelligence Benchmark (ACI-Bench) corpus, a unique dataset designed specifically for this task. Comprehensive experiments and comparative analyses demonstrate the model's superior performance over general-purpose language models in generating accurate and content-rich summaries, as evidenced by higher scores on automated evaluation metrics like BLEU and ROUGE.

By automating the clinical documentation process, this research offers a practical solution to alleviate the administrative burden on healthcare practitioners, enabling them to dedicate more time to direct patient care activities. The findings also contribute novel insights into the importance of domain adaptation for specialized language tasks and propose a framework for designing domain-aware summarization models.

Ultimately, this work paves the way for more efficient and effective healthcare documentation processes, with the potential to drive improvements in clinician well-being, patient outcomes, and overall healthcare quality.

**Keywords:** Clinical documentation, Medical dialogue summarization, Domain-specific language model, Healthcare efficiency, Sequence-to-sequence architecture

# Table of Contents

[**List of Tables 9**](#_n7grhergz885)

[**List of Figures 10**](#_yrjt2k16r6i3)

[**Acknowledgements 11**](#_9y63v1d4ap6a)

[**Chapter 1 - Introduction 12**](#_opa0xjn0mpt4)

[1.1 Background and Context 12](#_mv21y5ymgkjb)

[1.2 Aim and Objectives 13](#_jddtdkwpma8q)

[1.3 Research Questions and Hypotheses 13](#_1utyb2asvzn6)

[1.4 Motivation for this study 14](#_llxykvesydoy)

[1.5 Scope and Limitations 14](#_ssohukasdaep)

[1.6 Report Structure 15](#_n8wtx1llehlr)

[**Chapter 2 - Literature Review 16**](#_njp7jx8b9zff)

[2.1 Clinical documentation 16](#_tgzhr2c8218e)

[2.2 Conversational Summarization 17](#_jj7k2vcz4eyq)

[2.2.1 Speech-to-Text Methods 17](#_91fqs3cmyjph)

[2.2.2 Entity/Relation Extraction 18](#_az2gt4yuif3d)

[2.2.3 Generating Notes from Conversations 18](#_8lbswrgrujav)

[2.3 Neural Text Summarization 19](#_aaagtrehbje)

[2.3.1 Sequence-to-Sequence Models 19](#_37gg3srzh8vp)

[2.3.2 Attention Mechanisms 20](#_23idyhffqgs)

[2.3.3 Transformer Architecture 21](#_d0iw2egjwyzd)

[2.3.4 Summarization of Long Documents 22](#_u2jj7n3ne0so)

[2.4 Clinical NLP 23](#_vba8p3fufcd3)

[2.4.1 De-identification 23](#_o904al5an600)

[2.4.2 Named Entity Recognition 24](#_bmga91migo48)

[2.4.3 Entity Linking and Normalization 24](#_3bks4iu04ti3)

[2.4.4 Relation Extraction 24](#_mcbot8164vkq)

[2.4.5 Language Modeling 25](#_k6ki7zc2t05f)

[2.4.6 Clinical Datasets 25](#_j4q5ath936hk)

[2.5 Clinical Text Generation 26](#_nvb065236okt)

[2.5.1 Rule-based and Template Systems 26](#_brzfvbl9u1ry)

[2.5.2 End-to-End Neural Models 26](#_l6vlaqp7zg3y)

[2.5.3 Challenges with Fluency and Faithfulness 27](#_f3vivrriuhtf)

[2.6 Evaluation of Summaries 27](#_jwbthm15nez5)

[2.6.1 ROUGE and BLEU Metrics 28](#_s4ykin7medf4)

[2.6.2 Embedding-based Semantic Similarity 28](#_gqsv8w41p4mu)

[2.6.3 Human Judgment 28](#_eoudk19cethd)

[2.7 Research Gap 29](#_9mir1nso93er)

[**Chapter 3 - Methodology 30**](#_43u681utxb2v)

[3.1 System Design and Architecture 30](#_723bi1i5xvve)

[3.2 Dataset 31](#_nieclr49xnqr)

[3.2.1 Overview of ACI-Bench 31](#_rl62tp22f8bh)

[3.2.2 Data Creation Process 32](#_rxwj4smr8vkd)

[3.2.3 Comparison to Real Data 33](#_xdgy78wh1crb)

[3.2.4 Reason for choosing ACI-Bench 33](#_499dqq82cxkx)

[3.3 Preprocessing 33](#_1k9xxvzb5rpi)

[3.3.1 Text Cleaning 34](#_7nqwnyz1t1xt)

[3.3.2 Handling Long Sequences 34](#_1o8r6ekx0j9t)

[3.3.3 Rare Word Handling 35](#_jf346kstdeb4)

[3.3.4 Tokenization and Sequence Padding 35](#_yq1w2n7f5oxe)

[3.3.5 Special Tokens 36](#_bo07xohojf57)

[3.3.6 Train-Test Split 36](#_4k24cqg6agw8)

[3.4 Model Architecture 36](#_v9xsoxh7f4t4)

[3.4.1 Encoder 37](#_si1qt5lgsmg7)

[3.4.2 Decoder 38](#_bcmw5o3sif4g)

[3.4.3 Embeddings and Vocabularies 39](#_xv3zk0tvu29z)

[3.4.4 Avoiding Out-of-Vocabulary Words 39](#_wq5jfwnjl2qy)

[3.5 Model Training 39](#_pxxp3zyb7hp4)

[3.5.1 Optimizer 40](#_miwx6vy26ih2)

[3.5.2 Loss Function 40](#_51gm84tqlef2)

[3.5.3 Teacher Forcing and Scheduled Sampling 41](#_td5sbvtlayz)

[3.5.4 Early Stopping and Regularization 41](#_pfdcees0yqe3)

[3.5.5 Batch Size and Epochs 41](#_joomx0hy8bye)

[**Chapter 4 - Results and Analysis 43**](#_79qvaklnipm)

[4.1 Training Loss and Epochs 43](#_miohjnz79gbg)

[4.2 Evaluation Metrics 44](#_wt8yl2hn52dk)

[4.2.1 BLEU Score 44](#_oq457b9l9ltm)

[4.2.2 ROUGE Score 45](#_c9th5463q7ug)

[4.2.3 Levenshtein Distance 46](#_kzrx0elnugf8)

[4.3 Comparative Analysis 47](#_vm6g5fglrrc)

[4.3.1 T5 Large 47](#_vpf70uza9ow3)

[4.3.2 Google Pegasus Large 47](#_il7th1bwc2v3)

[4.3.3 Model Training and Inference 48](#_u7bszbar1n6z)

[4.3.4 Results 48](#_lgfz93ov9vj1)

[**Chapter 5 - Conclusion 52**](#_3xw15g9v6hyi)

[5.1 Summary of Research Questions and Findings 52](#_1s4fbbufkadc)

[5.2 Research Contributions 53](#_32ds9hw3228y)

[5.2.1 Theoretical Contributions 53](#_ii30erbn2xsn)

[5.2.2 Methodological Contributions 53](#_hiu8c3k2nbh2)

[5.2.3 Practical Contributions 54](#_iy1oswaghtzi)

[5.3 Limitations and Future Directions 54](#_t2wnru49qfgt)

[5.3.1 Limitations of the Research 55](#_mr0rxchfeyzn)

[5.3.2 Recommendations for Future Work 55](#_5el3rnngr55a)

[5.4 Final Remarks 56](#_h6i19deymysl)

[**References 58**](#_72ub3yu2jp2l)

# List of Tables

1. Results from Comparative Analysis 49

# List of Figures

|  |
| --- |
| * [Figure 1. Sequence-to-sequence model (Espejel, 2021).](#kix.kyt1oanzmxs7)  20 * [Figure 2. Different Attention Mechanisms (Khandelwal, 2020).](#e8ilufr386ym)  21 * [Figure 3. Transformers General Architecture (Espejel, 2021)](#7txjysh142t)  22 * [Figure 4. Flowchart of methodology](#kix.u8ctk7mfgvmu) 30 * [Figure 5. Example from ACI-Bench dataset (Abacha, 2023)](#kix.lsk5756dmb5y)  32 * [Figure 6. Distribution of length of text in text and summary columns](#kix.d5nlhehyoky9) 35 * [Figure 7. Model Architecture](#kix.t06a0013qv4k) 37 * [Figure 8. Encoder architecture](#kix.fqn4yj2j1hic) 38 * [Figure 9. Decoder architecture](#kix.nhjas9g562f0) 39 * [Figure 10. Training and validation loss](#690m0gptk4hp) 44 * [Figure 11. Mean BLEU Scores of all models](#hdm99za0go13) 49 * [Figure 12. Mean ROUGE Scores of all models](#97up47wragb5) 50 * [Figure 13. Mean Levenshtein Distance of all models](#7lxzkb38jbeg) 51 |

# Acknowledgements

“All praises be to Allah, The Lord of The World.”

I want to thank and pay gratitude to my supervisor Md Zia Ullah. You have been so kind all the time during my master’s degree. It has been a gratifying experience to work and nourish under your guidance. Your supervision method really worked for me. I think, and I believe that I am fortunate to have a supervisor and a mentor like you in my career. I always wish better things for you. I'd like to thank the institution for providing resources that were critical to the accomplishment of our research.

I cannot find suitable words to express my utmost sentiments of appreciation for my parents, who have brought me up with love and care and set me up to meet the difficulties and challenges in life. Studying overseas is not an easy decision; you have to leave your family and your loved ones for a very long time, but thanks to my “special one” who never made me feel homesick with her love, time, and attention. I am immensely thankful to my family and “special one” for their support. I pray to my beloved Allah for my parents and “special one” better health and success.

# Chapter 1 - Introduction

## 1.1 Background and Context

Clinical documentation is a crucial part of healthcare that fulfills a number of purposes, such as processing insurance claims, maintaining legal compliance, and ensuring continuity of care. A conventional practice takes a lot of clinician time away from more effective patient care and diverts it into taking handwritten clinical notes during or after patient visits. The healthcare industry has recognized the need to reduce this load and has determined that increasing documentation efficiency is essential to enhancing overall healthcare effectiveness and results.

It is imperative to use clinical notes to transform casual discussions between a patient and a physician into formal medical records. Electronic health records (EHRs) and natural language processing (NLP) approaches can be combined to automate and streamline this documentation process. However, a survey by Arndt et al. (2017) indicates that doctors are generally unhappy with the way EHR documentation is done today, emphasizing the need for creative solutions.

One current subfield focuses on automating the creation of clinical notes using machine learning models on audio recordings of doctor-patient conversations. Rajkomar et al. (2019) described a method that employs machine learning to construct charting templates from spoken interactions. Meanwhile, to set the stage for note production, additional works concentrate on extracting key entities and relationships from talks (Chien et al., 2019). Even with these improvements, it is still very difficult to distil meaningful stories from intricate, informal, and long medical conversations.

The problem goes beyond simply giving physicians advanced composition interfaces. Attempts to support physicians in note preparation include autocomplete suggestions for medicine names (Buurman et al., 2015) and recommendations for pertinent information (Park et al., 2019). Still, healthcare providers are burdened with a significant amount of paperwork due to these reactive methods.

Eliminating the need for transcribing and optimizing the amount of time spent interacting between the doctor and the patient is possible with fully automated clinical note generation. To do this, models for natural language generation must be able to evaluate dialogue content and condense it into writing suitable for a therapeutic setting. According to Sheikhalishahi et al. (2019), the sophisticated medical language generation capabilities required for meaningful clinical note development are absent from current big pre-trained models. Furthermore, the fundamental subjectivity of output quality makes it difficult to evaluate automatically generated text, including clinical notes. This is demonstrated by flaws in both automated and human evaluations (Giorgi et al., 2023).

Although models for clinical natural language processing tasks have been strengthened by recent advancements in representation and transfer learning, a barrier still exists in the lack of large datasets designed specifically for visit or dialogue summaries. A promising but underutilized strategy is to specialize models to the clinical domain by supervised learning on provider-patient interactions and aligned note summaries. Enhancing the quality and efficiency of clinical note generation is important for broader healthcare implications, including research and patient outcomes. Artificial intelligence (AI) has the potential to reduce clinician fatigue, increase productivity, and improve patient outcomes.

## 1.2 Aim and Objectives

This dissertation aims to develop and evaluate a Domain-Specific Language Model for summarizing medical dialogues. The objective is to create a model that effectively captures key clinical information from conversations between healthcare professionals and patients, ultimately generating concise and informative summaries.  
  
Here are the specific objectives that will be addressed:

* Conduct a comprehensive literature review to analyze existing clinical documentation techniques and challenges in medical dialogue summarization.
* Identify and evaluate suitable datasets for training and testing the proposed model
* Develop a robust data preprocessing pipeline to prepare the chosen dataset for the model training process
* Develop a novel architecture using encoder-decoder and sequence-to-sequence models.
* Optimize the proposed model architecture through hyperparameter tuning to achieve the best possible performance on the summarization task.
* Evaluate the effectiveness using established metrics like ROUGE, BLEU, and Levenshtein distance.
* Conduct a comparative analysis, benchmarking the performance of the proposed model against existing state-of-the-art clinical note summarization techniques.

## 1.3 Research Questions and Hypotheses

1. **Adaptation of Natural Language Processing Techniques** (RQ1):

Research Question: How can current natural language processing techniques be adapted to generate high-quality clinical note summaries from medical conversations?

Hypothesis: The model's ability to extract and condense complex health information from patient-doctor discussions will be greatly improved by incorporating cutting-edge NLP techniques like encoder-decoder and sequence-to-sequence models.

Justification: It's possible that existing NLP techniques aren't naturally tailored to the particular difficulties presented by medical dialogues. The study seeks to identify particular modifications that optimize the model's efficiency in clinical note summarizing by tackling this subject.

1. **Model Accuracy and Health Detail Summarization** (RQ2):

Research Question: How accurately does the model extract and summarize relevant health details from dialogues based on automated evaluation metrics?

Hypothesis: The proposed model will outperform current methods in terms of extracting and summarizing relevant health information, as shown by its strong performance on automated evaluation metrics.

Justification: By emphasizing automated evaluation criteria like ROUGE and BLEU, an impartial evaluation of the model's effectiveness is guaranteed. According to the background, the model's architecture will be efficient in gathering and condensing health data.

1. **Domain-Specific Model Outperformance** (RQ3):

Research Question: Does a domain-specific model architecture outperform a general language model baseline on evaluation metrics?

Hypothesis: By demonstrating improved performance in clinical note summarization, the domain-specific model architecture presented in this study will outperform a general language model baseline on key metrics.

Justification It is anticipated that summaries that are not only linguistically coherent but also clinically relevant and succinct would be produced by customizing the model architecture to the healthcare domain.

## 1.4 Motivation for this study

The current state of medical dialogue summary systems is deficient in many ways, mostly due to inefficiencies in the process of converting patient-doctor interactions into concise, comprehensive clinical note summaries. This divide not only creates difficulties for insurance companies and legal compliance, but it also obstructs the smooth exchange of information that is essential for patient care and overall healthcare efficacy.

It is clear that the healthcare industry needs a specialized language model given the unique peculiarities of medical language, where nuance matters and precision is essential. Existing large pre-trained models lack the sophistication needed for significant clinical note production (Sheikhalishahi et al., 2019). In biological and clinical corpora, the application of domain-specific models, like ClinicalBERT and BioBERT (Alsentzer et al., 2019; Lee et al., 2020), has demonstrated promise. Still, there is much to learn about applying these models to the complex task of creating clinical note summaries from conversations.

To bridge this gap and address the limitations of current approaches, our project developed a sequence-to-sequence (seq2seq) model with an encoder-decoder architecture tailored for medical dialogue summarization. The project intends to develop not only the field of medical dialogue summarizing but also provide concrete benefits to healthcare practitioners, ultimately contributing to improved patient outcomes and streamlined healthcare processes.

## 1.5 Scope and Limitations

The primary objective of this work is to construct and assess a deep learning model specifically designed for clinical note summarization from patient-doctor discussions. In order to emphasize the need for a personalized solution in the context of clinical conversations, the scope has been purposely restricted to this particular area of medical recordkeeping. The study focuses on the difficulties of summarizing informal and richly detailed patient-doctor interactions, avoiding more general electronic health record (EHR) documentation or other types of medical text processing.

Although the suggested model makes use of state-of-the-art methods like LSTM-based decoders, it has several built-in drawbacks. The Clinical Visit Note Summarization Corpus's representativeness and quality has a significant influence on the model's performance. The model's potential to be broadly used may be impacted by the challenge of using small, annotated datasets for medical conversation summarization.

The model's dependence on pre-existing frameworks raises ethical concerns that need to be carefully evaluated because they may introduce biases into the training set. While a thorough examination of the ethical issues is not included in the scope, it is acknowledged that they are relevant when using automated technologies in the healthcare industry. By recognizing these constraints, the study upholds openness regarding the defined limitations and possible obstacles, guaranteeing a sophisticated analysis of the model's results within the designated parameters.

## 

## 1.6 Report Structure

The upcoming report is structured as follows:

* The Literature Review explores the fundamentals of summarizing medical conversation. It examines the present state of natural language processing in healthcare, as well as prior efforts to clinical note summarization and the state-of-the-art in the field. By pointing out the gaps, difficulties, and theoretical foundations that direct the creation of the deep learning model, this chapter establishes the framework for the suggested research.
* The method used to answer the research questions is carefully laid out in the Methodology chapter. It describes how the deep learning model was developed, highlighting the customized elements of the architecture such as the encoder-decoder architecture and the inference models. Clear instructions on how to preprocess datasets, train models, and use GPU-capable hardware are provided, giving model developers a clear path forward.
* In the Evaluation chapter, the emphasis changes to evaluating the model's performance. The correctness and coherence of the autogenerated clinical note summaries are measured using evaluation measures like ROUGE and BLEU. The proposed domain-specific model is also evaluated critically against other clinical note summarizing techniques in this section, highlighting the model's advantages and possible shortcomings.
* The results from the chapters on methodology, evaluation, and literature review are summarized in the Conclusion chapter. It discusses the study's ramifications, answers the research questions, and evaluates the findings. This chapter also looks at directions for future study, taking into account how the field of natural language processing in healthcare is changing.

# 

# Chapter 2 - Literature Review

## 2.1 Clinical documentation

Clinical documentation is the term used to describe the thorough written records that medical professionals keep to record clinical interactions, protocols, treatment plans, and specifics about each patient's health. For the goal of legal and regulatory compliance, insurance claim processing, quality assurance, public health reporting, and facilitating data-driven advancements in healthcare delivery, high-quality documentation is crucial (Hirschtick, 2006). The clinical note, also known as a progress note or SOAP note, is the primary recording artifact in outpatient settings. It provides a summary of each patient visit or encounter (Schloss & Konam, 2020).

Historically, clinical documentation consisted of handwritten notes precisely created by healthcare experts. These documents function as a channel of communication for healthcare professionals in addition to being a way to record patient information (Isah & Byström, 2020). With the advent of typewriters and standardized forms, the process was sped up and uniformity in the documentation of medical information was introduced. Clinical documentation underwent a radical transformation as a result of the digitization of medical records (Pine & Bossen, 2020). This development was prompted by the enactment of new healthcare legislation and policies that aim to protect patient privacy, enhance patient care, and improve healthcare delivery generally. For example, the 1996 Health Insurance Portability and Accountability Act (HIPAA) laid the groundwork for the widespread use of electronic health records (EHRs) by emphasizing the importance of safeguarding patient privacy and health information (Moore & Frye, 2019).

By transforming subjective patient complaints, objective measures, assessment interpretations, and care plans, clinical notes are now utilized to codify written records of clinical encounters (Ting et al., 2021). A standard SOAP note consists of four components: the plan for future treatments, medication, or tests; the patient's subjective information; the provider's objective clinical observations; and the provider's assessment of the patient's state. Important information from the appointment, such as the patient's vital signs, medications, allergies, family history, and medical history are also recorded in the clinical notes (Buurman et al., 2015).

Providers write several more documentation artefacts in addition to the clinical notes for outpatient appointments. Summaries of admission and discharge describe how a patient's health changed while they were in the hospital. Whereas pathology notes identify specimen analyses, operative notes describe surgical operations (Braaf et al., 2011). Providers record public health reporting, home health directives, referral letters, and imaging interpretations. Providers must combine patient health information and translate between casual discussions and formal medical language in order to produce high-quality clinical documentation (Downing et al., 2018).

According to studies, clinicians spend an extra one to two hours documenting for every hour they are with patients (Arndt et al., 2017). Overly burdensome documentation requirements are a significant cause of burnout among clinicians; some even quit their jobs as a result of paperwork-related annoyances and disruptions to patient care (Melnick et al., 2019). Detailed documentation requirements for billing and compliance, ineffective interfaces and workflows within electronic health records, difficulties in accurately capturing detailed patient narratives, and repetitive documentation of frequently collected information such as vital signs and medical history are some of the major sources of documentation burden (Halcomb et al., 2016).

The increasing use of EHRs has made clinical documentation more data-driven, but it has also become more difficult in many aspects. While EHR systems improve clinical decision support and patient data availability, practitioners often cite difficulty with complicated or disjointed documentation workflows, poor system usability, and interference from technology use during interactions (Downing et al., 2018). According to recent polls, most providers want to increase the accuracy and efficiency of their documentation by using tools like conversational documentation interfaces, automated documentation, and documentation support (Halcomb et al., 2016).

By extracting information from data and interactions, creating draft notes, and combining patient details into cohesive narratives, natural language processing techniques offer new options to automate and expedite clinical recording (Buurman et al., 2015). Still, it is a challenging task to faithfully capture the subtleties of patient encounters and turn informal conversations into excellent clinical language. Requires strong language creation, comprehension, and reasoning skills for automated documentation that is useful.

## 2.2 Conversational Summarization

Clinical documentation research is primarily interested in developing methods for automatically distilling medical discussions into concise clinical notes or reports. The majority of patient visits and consultations consist of spoken dialogues between patients and providers; therefore, in order to automate the process of minimizing documentation burdens, it is critical to be able to translate these discussions effectively and distill the pertinent information into well-written narratives. In order to provide a summary that captures the most important information, conversational summarizing includes both the transcription of the uncut chat audio into text and the identification of important entities and relationships from the transcripts.

### 2.2.1 Speech-to-Text Methods

Precise transcription of the raw audio recording into text is the initial stage in assessing clinical discussions for summarization (Soltau et al., 2021). Despite recent significant advancements in speech recognition technology, transcribing medical conversations poses particular difficulties. The patient and the clinician frequently overlap when speaking, and there are also frequent ellipses, topic switches, and interrupts. The audio quality varies a lot, and the language includes a lot of acronyms, abbreviations, and domain-specific terms.

Past studies concentrated on using clinical corpora to train voice recognition systems and enriching broad vocabularies with medical terms to improve handling of specialized language. Additionally, hybrid systems that integrate clinical grammar and dialog structure principles with machine learning were investigated (Ruch et al., 2008). Convolutional and recurrent neural network designs, which can be trained directly on raw audio and text pairings, are used in more modern techniques. Medical speech recognition performance has been further enhanced using transfer learning through large pretrained clinical language model fine-tuning (Rasmy et al., 2021).

But in fact, recognition errors still happen a lot, especially for uncommon phrases and noisy audio. Enhancing robustness to accented speech, background noise, interruptions, and multi-speaker conversations—all of which are typical in clinical settings—is the goal of ongoing research. To enable speaker role parsing in transcripts, speaker diarization—which segments audio by speaker and assigns provider vs. patient labels—remains an unexplored research topic (Quiroz et al., 2019). Reliable transcription of clinical conversations requires the addition of more representative corpora and ongoing advancements in model architecture.

### 2.2.2 Entity/Relation Extraction

Natural language processing techniques can be used to identify significant items and relations to extract into a summary after clinical discussions have been mechanically transcribed into text (Demner-Fushman et al., 2021). When educated on medical data, standard named entity recognition systems may accurately identify important components such as prescription drugs, symptoms, diagnoses, measurements, and procedures. To facilitate structured data extraction, more sophisticated entity linking further links these concepts to standardized medical ontologies such as SNOMED-CT, RxNorm, and ICD-10.

The goal of relation extraction is to find connections between the important elements, like the relationship between a pharmaceutical and its related indication or a symptom and a diagnosed ailment. Clinical relation extraction has been used with rule-based patterns, statistical machine learning classifiers, and graph networks (Wang et al., 2018). It is still difficult to capture the nuances and extra semantics required for a high-quality clinical summary. Relying solely on pre-defined items and relations may result in the missing of important context. To create a credible narrative, linguistic and clinical reasoning over the entire conversation text is required.

### 2.2.3 Generating Notes from Conversations

The most promising approaches for precise and useful documentation are end-to-end generation models that generate clinical notes directly from provider-patient talks. Previous rule-based systems used templates to extract items from transcripts, but this frequently led to incomplete and inconsistent note narratives. Brittle characteristics and a deficiency in clinical language fluency also posed challenges for statistical machine learning techniques such as Hidden Markov Models (Yue et al., 2020).

Neural encoder-decoder architectures, including transformers and LSTMs, are used in recent methods to map talks to summaries in an end-to-end manner with ease (Espejel, 2021). When building the narrative, attention processes assist the models in concentrating on the most prominent dialogue segments. When compared to previous systems, these data-driven approaches produce narratives that are more fluid. Accurately analyzing ambiguous remarks, expressing clinical logic, and converting informal speech into formal clinical language norms are still difficult tasks (Quiroz et al., 2019).

Pretraining transformer encoder-decoder models on large unlabeled corpora of medical text before fine-tuning on aligned visit transcripts and clinical notes resulted in additional fluency and clinical language use improvements (Kalyan et al., 2022). Nonetheless, the current aligned datasets for conversational clinical summarization have restricted dimensions and a narrow range of domains included. Expanded high-quality medical conversation corpora and enhanced neural architectures specifically designed for this challenging task will be necessary for further advancement.

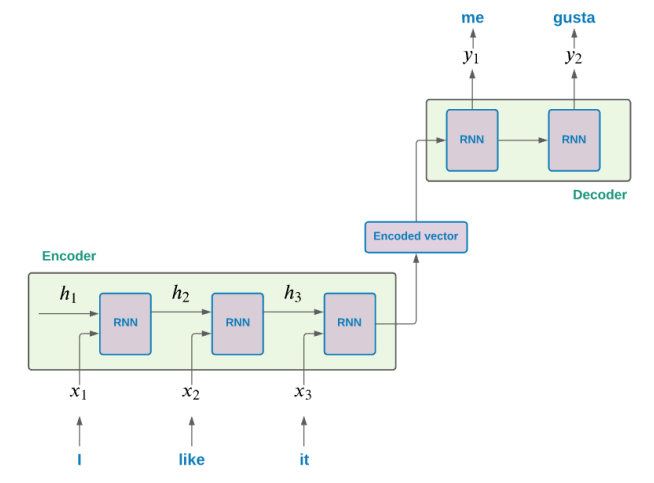
## 2.3 Neural Text Summarization

Recent advances in neural network techniques for natural language processing have led to significant improvements in automatic text summarization capabilities. Developing end-to-end models that accurately extract relevant information from papers and condense it into concise summaries is the aim of neural summarizing. Compared to previous rules-based or statistical methods that relied on manually created features and domain-specific engineering, this provides a more flexible learning approach. For text summarization tasks, sequence-to-sequence models, attention mechanisms, and transformer networks are the three main neural architectures that have gained prominence.

### 2.3.1 Sequence-to-Sequence Models

Neural networks were first used in natural language processing with sequence-to-sequence (seq2seq) models, namely the encoder-decoder architecture created by Sutskever et al. (2014). Text summarization proved to be an early use for sequence-to-sequence models, which were originally meant for uses like machine translation. One of the earliest end-to-end neural architectures used for text summarization tasks was the sequence-to-sequence (seq2seq) model (Sutskever et al., 2014). Sequence-to-sequence summarization models based on the encoder-decoder architecture, are composed of a recurrent neural network encoder that scans the input text and encodes it into a fixed-length vector representation that captures the semantic information. A decoder RNN receives this context vector and uses it to condition each step of its output summary generation on the encoded vector, producing a word-by-word summary.

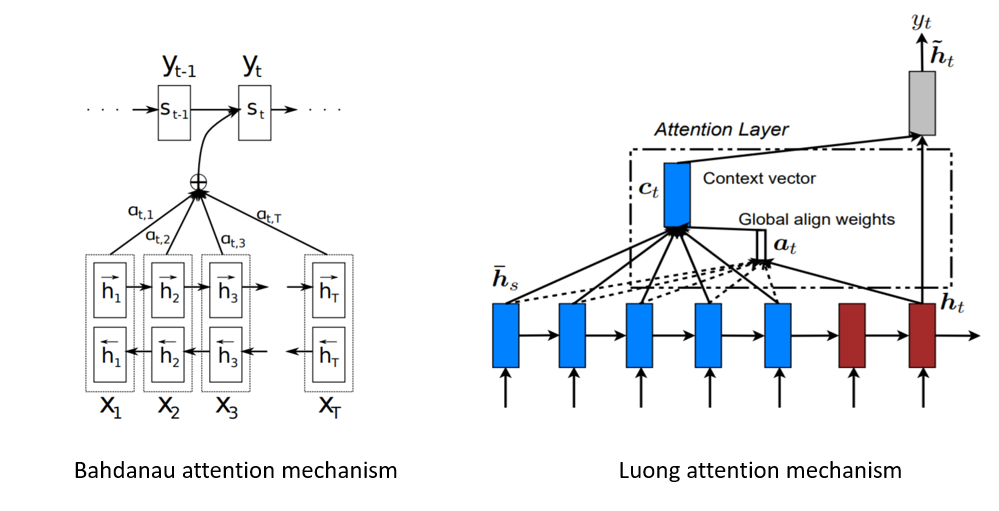
Because LSTM networks can capture long-range relationships in sequences, they are frequently utilized as the encoder and decoder components. Identification of salient content is possible by applying readout methods over the encoder outputs. Enabling end-to-end training, the seq2seq architecture maximizes the likelihood of generated summaries given input texts. Basic seq2seq models demonstrated early promise in learning text-to-text rewriting tasks such as summarization, despite their poor capacity to encode important semantics (See et al., 2017).



*Figure 1. Sequence-to-sequence model (Espejel, 2021).*

### 2.3.2 Attention Mechanisms

By enabling the decoder to make direct references to the input text during the summarizing process, attention mechanisms improve seq2seq summarization models (Bahdanau et al., 2014). During decoding, attention computes dynamic context vectors focused on pertinent portions of the input, as opposed to encoding the entire text into a single fixed-length vector. At that point, the decoder can highlight crucial information while disregarding unimportant elements. This enhances the model's capacity to concentrate on important details while generating succinct summaries.

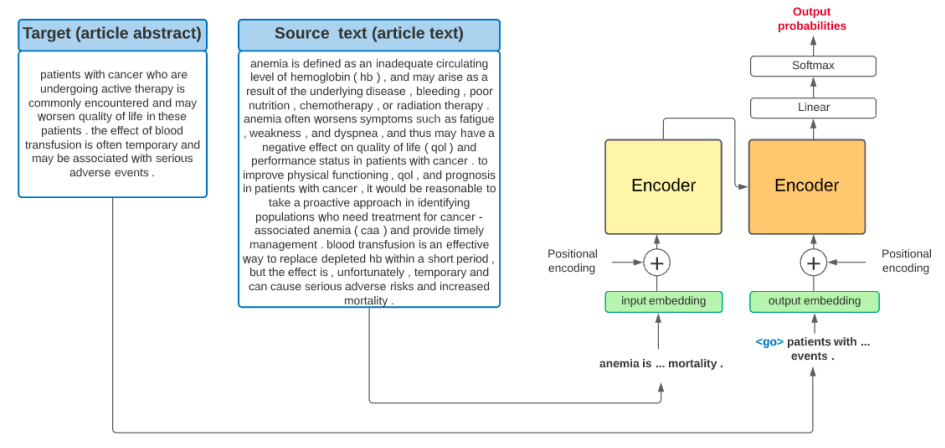


*Figure 2. Different Attention Mechanisms (Khandelwal, 2020).*

### 2.3.3 Transformer Architecture

Transformer networks—which are primarily focused on attention mechanisms—have taken the lead in natural language processing architectures, including summarization (Vaswani et al., 2017). Transformers incorporate multi-headed self-attention instead of repetition to better capture long-range dependencies in text. Liu and Lapata (2019) state that pretrained transformer encoders, such as BERT, offer valuable contextual representations of words and sentences that can be optimized for summarization tasks.

Transformer encoder-decoder models, by learning to map texts to summaries through a fully attention-based design, have achieved state-of-the-art performance on text summarizing benchmarks. The enhanced abstractive and extractive summarization abilities surpass those of previous sequence-to-sequence methods. By enabling word copying from the source text, strategies such as pointer-generators enhance handling of terms that are not in the dictionary (See et al., 2017).



*Figure 3. Transformers General Architecture (Espejel, 2021)*

Because GPT-3, developed by OpenAI, utilizes a large number of parameters within a transformer architecture, it excels at providing logical and contextually rich summaries (Brown et al., 2020). Contextual understanding, which identifies intricate relationships between words and phrases to generate summaries that are more logical and appropriate for the context, is its main strength (Radford et al., 2021). GPT-3's enhanced coherence and fluency demonstrate its advantage in abstractive summarization. The algorithm produces summaries that read more naturally and closely resemble human-generated text by utilizing substantial pre-training on a variety of datasets (Brown et al., 2020). Interestingly, GPT-3 exhibits impressive multimodal properties that enable it to be applied to challenges that require both textual and visual input. According to Brown et al. (2020), this makes it especially beneficial for domains that call for multimedia summarization.

### 2.3.4 Summarization of Long Documents

A great deal of research has focused on summarizing short texts. Summarizing lengthy publications, such research papers and reports, can be challenging. Hierarchical encoders that operate at both the word and sentence levels have been used to summarize long manuscripts. Promising models include two-stage and combined extraction-abstraction techniques, which incorporate explicit sentence-level extractive selection prior to abstractive generation (Xu & Durrett, 2019).

A particular kind of recurrent neural network (RNN) called Long Short-Term Memory (LSTM) was created to solve the vanishing gradient issue that conventional RNNs have. LSTMs are useful for document summarization because they are good at capturing sequential patterns and long-range relationships. LSTMs play a crucial role in document summarization, especially for long texts, as they can provide coherent summaries by preserving and utilizing information from previous parts and successfully capturing the contextual links between words and phrases. Cho et al. (2019) provided precise summaries of research publications and clinical notes to illustrate the usefulness of LSTMs in the medical field.

For the purpose of summarizing lengthy publications, particularly technical articles, graph-based models that depict documents as graphs with semantic links between items have also been investigated (Jin & Szolovits, 2018). Graph Convolutional Networks (GCNs) designed for processing graph-structured data, play a pivotal role in modeling hierarchical structures within documents. By treating sentences or sections as nodes and relationships as edges, GCNs capture interdependencies, contributing to the generation of comprehensive and contextually rich summaries. Producing multi-sentence summaries that are more coherent is made possible by including semantic graph representations of argument linkages and text structure. In scientific document summarization, GCNs have been applied to understand citation networks and summarize scientific literature effectively, as evidenced by Yasunaga et al. (2019) research.

Furthermore, combining LSTMs with GCNs takes advantage of their complimentary benefits. The sequential modeling of LSTMs and the graph-based structural knowledge of GCNs work together to enhance summarization performance. This integration is corroborated by Jeong et al.'s (2020) research, which shows improved summarizing ability, particularly in documents with complicated structures.

## 2.4 Clinical NLP

A number of significant tasks, including text generation, relation extraction, named entity recognition, entity linking, de-identification, and language modeling, are the focus of clinical NLP research. Advances in these domains affect applications later on, including phenotyping, automated documentation, predictive risk modeling, and decision support for healthcare.

### 2.4.1 De-identification

In order to anonymize data for research purposes and safeguard patient privacy, de-identification is a crucial preprocessing step because clinical papers include protected health information (PHI) about patients. PHI includes both direct identifiers, such as names, addresses, dates, times, and IDs, and indirect identifiers, which may be used to deduce identification. To consistently redact or label PHI items in clinical writing, rule-based methods, statistical machine learning classifiers, and neural networks have all been used (Yeniterzi et al., 2010).

Rule-based systems produce interpretable but fragile models by identifying common PHI constructs based on expert knowledge using lexicons, regular expressions, and pattern matching. Supervised machine learning techniques, such as CNNs and conditional random fields, need a lot of annotations in order to learn statistical PHI patterns from labeled data. The state-of-the-art presently consists of neural models that use both pretraining and supervised fine-tuning on de-identification datasets, such as bidirectional LSTMs and attention networks (Yang et al., 2019).

Nevertheless, completely automated clinical de-identification still faces formidable obstacles. Difficulties arise when attempting to express contextual identifier patterns, negations, and coreferences through rules. Reversible de-identification, which preserves the ability to re-identify data when needed, and privacy-preserving record linkage algorithms are two emerging research fields (Aberdeen et al., 2010).

### 2.4.2 Named Entity Recognition

In order to extract structured information from patient notes, named entity recognition (NER) identifies clinical entities and medical concepts such as measures, diagnoses, treatments, drugs, and symptoms. Statistical machine learning classifiers such as conditional random fields, dictionary-based matching, rule-based pattern matching, and deep learning networks have all been used for clinical NER (Wang et al., 2018).

The state-of-the-art presently consists of deep learning methods with CNNs, RNNs, and pretrained language model fine-tuning; these methods require huge labeled corpora for training (Kundeti et al., 2016). The ambiguity and unpredictability that are inherent in the usage of medical language cause performance to vary depending on the clinical sublanguage. Comprehensive community-shared tasks such as i2b2 have access to cutting-edge clinical NER tools and techniques. Self-attention is used by recent Transformer-based models to enhance contextual comprehension and entity disambiguation.

### 2.4.3 Entity Linking and Normalization

Entity linking maps extracted entities with ontology concepts and conventional clinical codes such as ICD-10, RxNorm, MeSH, and SNOMED. According to Bodenreider et al. (2018), this offers normalized representations of various surface morphologies and synonyms. For entity normalization and standardization, dictionary-based matching, vector similarity measures, and concept classifiers have been used. Out-of-vocabulary terms and numerous mapping candidates present challenges.

### 2.4.4 Relation Extraction

Understanding and reasoning with medical language is further enabled by extracting links between clinical concepts. Along with many other links, this includes relationships between issues and therapies, drugs and indications, laboratories and disorders, and more (Wang et al., 2018). For clinical relation extraction, supervised machine learning techniques such as CNNs, RNNs, and graph networks, as well as rule-based patterns utilizing dependency parses or templates, have been used.

Word embeddings allow neural networks to infer semantics and relationship patterns from unstructured text implicitly. However, extensive training datasets with a variety of clinical relationships are necessary for their accuracy. Though generalization is still difficult, ongoing efforts have collected corpora and created models for a variety of relation types, such as treatment relations and temporal event relations (Du et al., 2019).

### 2.4.5 Language Modeling

In clinical NLP, pre-trained language models such as BERT, RoBERTa and ClinicalBERT, fine-tuned on domain-specific corpora, have taken centre stage. Before honing their skills on subsequent tasks, they offer universal representations that capture syntactic and semantic patterns through unsupervised pretraining objectives (Huang et al., 2019). Even with tiny training datasets, the contextual embeddings can capture word usage across varied contexts and can be effectively transferred be applied even to downstream tasks to give good results.

Task-specific language models do not encompass the breadth of vocabulary and diversity of linguistic constructions that clinical language models provide. They reduce the requirement for intricate task-specific architectures and feature engineering for each new dataset. To fine-tune the pretrained embeddings, task-specific output layers are added and downstream annotations are trained on. To strike a compromise between task-specific optimization and transfer learning, both frozen and unfrozen fine-tuning techniques are being investigated. Additional advances in tasks like NER have been observed when combining customized models designed for specific entities or interactions with clinical BERT.

The quality of the pretraining corpus has a significant impact on performance. Clinical trial reports and patient records are not the same as general biomedical research literature. Peng et al. (2019) have reported the creation of EHR corpora specifically intended for pretraining. Adding specialized architectures to clinical BERT embeddings for certain entities or relations has resulted in further improvements in tasks such as relation extraction and named entity recognition. The transferability of clinical language models is expected to be further improved by ongoing pretraining on ever-larger clinical corpora.

### 2.4.6 Clinical Datasets

The advancement of clinical NLP research has been greatly aided by carefully selected datasets. Annotated corpora for entity, relation, and idea extraction from clinical text have been generated by community shared activities such as i2b2, n2c2, and ShARe (Stubbs et al., 2015). Pretraining and transfer learning benchmarks have been made possible by the MIMIC dataset of deidentified EHRs. MuCoW and other medical conversation datasets offer prospects for multimodal analysis.

Nonetheless, in comparison to general domain corpora, the majority of current clinical NLP datasets are tiny in size. Many concentrate exclusively on a single sort of document type, organization, or medical specialization. To boost size and diversity, recent work has investigated synthesizing training data using back-translation and generative adversarial networks (Osuala et al., 2023). The difficulty of adapting transfer learning to new target disciplines and hospitals is still present. Substantially larger, representative annotated datasets reflecting a variety of therapeutic languages will be needed to support further advancements.

## 2.5 Clinical Text Generation

Research on automatically generating coherent medical text is ongoing, with potential uses in patient education, referral generation, and automated documentation. Previous methods depended on manually created templates and rules that needed deep domain knowledge. More versatile natural language production systems are being created with recent techniques that combine data-driven end-to-end learning with neural networks.

### 2.5.1 Rule-based and Template Systems

Earlier clinical text generation utilized manually created templates and rules created by subject matter experts. Production systems were created by encoding patterns and routines to convert structured data or measurements into textual paragraphs for particular use cases, such as the authoring of radiology reports (Buurman et al., 2015). These systems were basically logic rules and protocols that contained clinical logic and guidelines. Templates covering predicted linguistic variability in outputs were written by medical practitioners.

Clinicians were able to work more efficiently with the advent of template-based systems because they offer reusable text building blocks that can be easily modified with variables and keywords. Clinical document automation can be partially achieved by segmenting reports into headers, boilerplate statements, common language phrases, and closing remarks (Finch, 2012). To formalize report writing rules within institutions, template scripting languages such as the Clinical Language Understanding and Extraction Support System (CLUES) were developed.

However, human encoding of rules and syntax trees for managing complicated clinical language requires significant knowledge engineering efforts. Outlier cases are still not fully covered by rule-based systems and templates. Failures may happen if inputs depart from pre-established frameworks. In order to address the brittleness and maintenance costs of largely manually constructed systems, there has been increased interest in applying data-driven machine learning algorithms to learn text creation patterns from real-world corpora.

### 2.5.2 End-to-End Neural Models

Clinical text production now has a more adaptable, data-driven method thanks to recent neural natural language generation models. Recurrent network-based language models, such as Long Short-Term Memory (LSTMs), compute probability distributions over token sequences that can be sampled to produce believable versions of clinical text. Conditional variants create customized outputs based on certain patient situations by using structured medical data as additional model inputs, such as past encounter details, prescription details, and patient diagnoses.

By dynamically retrieving crucial information, attention layers assist neural generators in improving coherence and helping them focus on pertinent medical inputs (Goncalves et al., 2022). Today state-of-the-art is represented by transformer networks, which use multi-headed self-attention processes to better catch semantic and syntactic patterns that span vast distances in output text. It has been demonstrated that significantly improving fluency requires pretraining transformer encoders on large generic corpora as language models before fine-tuning on domain-specific data (Lee et al., 2019).

Compared to rule-based or template systems that are manually created for specific situations, neural clinical text generators are much more adaptable to handle a wider range of language variants. Rather than relying on manually constructed knowledge bases, they learn textual patterns from data. However, it is still difficult to generate narratives that accurately reflect both medical logic and patient experience. There is a chance of introducing false information or private data that isn't adequately supported by the inputs. Additionally, discriminative problems arise when generative models propagate biases and unjustified correlations from the distribution of training data.

### 2.5.3 Challenges with Fluency and Faithfulness

Generating output text that is both accurate to the factual inputs generated from patient reality and fluent in language is a major problem for usable clinical text production (Xie & Wang, 2023). Fluency is correlated with linguistic coherence, target user readability, and clarity of language. Being faithful entails being accurate, avoiding fabricating information that isn't backed up by inputs, and safeguarding patient privacy by not disclosing further protected health information.

In generative models, fluency and faithfulness frequently involve trade-offs. Systems that have been exposed to huge, diverse corpora and optimized for fluency may have false memories of details that are plausible but unfounded in the original data. Adding auxiliary loss terms that penalize unsupported information, limiting network vocabulary sizes, employing adversarial networks to distinguish between real and fake text, and incorporating retrieval mechanisms to replicate real text fragments from training corpora are some of the suggested techniques to increase faithfulness (Xie & Wang, 2023).

Approaches are also influenced by target conditions; for example, referrals may value concision above literary text, whereas patient education materials should maximize clarity. Even with significant advancements in data-driven clinical language generation, it is still difficult to accurately produce medical text that reflects patient reality while preserving coherence and depth. Controlled creation, customized styling, multimodal fusion, and human-AI collaboration interfaces are areas that require advancements, in addition to stringent evaluation processes and technical upgrades for textual accuracy.

## 2.6 Evaluation of Summaries

It is essential to appropriately assess the quality of automatically generated summaries in order to steer the development of summarizing systems forward. Each of the three methods—automated evaluation metrics, embedding-based semantic similarity measurements, and human judgments—offers insightful information on various aspects of quality, such as fluency, correctness, and conciseness. Adoption of rigorous evaluation techniques is necessary for reliable assessment of summary outputs during development, comparison of models after training, and detection of existing constraints.

### 2.6.1 ROUGE and BLEU Metrics

Text summarization systems are frequently developed and reported on using automated assessment measures such as ROUGE and BLEU (Lin, 2004). These metrics calculate the number of overlapping lexical units (such as n-grams) between human-written reference summaries and summaries generated by the model for the identical inputs. To calculate precision and recall-oriented scores, match counts are standardized.

Using n-gram overlap statistics, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) evaluates the main points' coverage and conciseness. ROUGE-1, ROUGE-2, and ROUGE-L variants, for example, compute the overlaps of bigrams, unigrams, and longest common subsequence, in that order. ROUGE-N examines n-gram co-occurrence that has been adjusted for reference length. Human evaluation of content selection and informativeness is correlated with high ROUGE.

Using modified n-gram precisions aggregated and multiplied by a shortness penalty factor, BLEU (BiLingual Evaluation Understudy) determines the precision of generated summaries (Papineni et al., 2001). To evaluate fluency, ratios of corresponding n-grams in candidate and reference texts are compared. Because of its effectiveness and correlation with human ranks, BLEU is frequently used for translation and summarization jobs.

### 2.6.2 Embedding-based Semantic Similarity

The shortcomings of basic string overlap metrics have sparked research in semantic similarity metrics based on embeddings, which are thought to better capture meaning (Ribeiro et al., 2020). Word and sentence embeddings store contextual and semantic information in such a way that vector distances are correlated with similarity assessments made by humans. Measures calculated on embeddings aid in evaluating minute differences overlooked in surface forms.

Semantic similarity can be estimated using both supervised and unsupervised methods on both general and clinical text corpora. On benchmarks for semantic textual similarity, BERT and other pretrained language model embeddings that have been adjusted based on downstream tasks show promising results. The choice of dimensionality and embedding strategy has a significant impact on performance. For a more thorough assessment, multidimensional evaluations across metrics are provided.

### 2.6.3 Human Judgment

The assessment of aspects such as general readability, fidelity, and conciseness by hand becomes more crucial when summarization technologies advance to near human levels of coherence and accuracy. Even if human ranking requires a lot of resources and is subjective, only human analysis can accurately pinpoint shortcomings and constraints to direct efforts toward improvement (Kryściński et al., 2019).

Creating strict rating methods with detailed criteria definitions and presenting masked mixtures of model outputs and human references for blind comparison evaluation are best practices. Subjectivity among raters is measured using inter-annotator measures. Thorough qualitative error assessments are just as important as quantitative ratings for pinpointing particular deficiencies in comprehension, logic, or language use. Automated metrics must be combined with expert sampled manual review to produce reliable progress assessments.

## 2.7 Research Gap

There are still significant research gaps in the areas of accurately interpreting and producing high-quality clinical summaries from casual verbal discussions. The majority of previous work only concentrates on extracting entities or templates, which results in fragmented and incomplete tales. While end-to-end neural models struggle with faithfulness and subtle reasoning, they show greater potential for fluent generation. They frequently lack the clinical language proficiency required for meaningful charting.

The lack of reliable neural architectures tailored to this medical conversational summarizing task is a major shortcoming. According to the research, unlike domain specialists, clinical language models that have been fine-tuned on generic corpora continue to generate unsubstantiated details and terminology usage. Although representations from large datasets are useful for pretraining, the quality of task-specific fine-tuning data—which is currently scarce in both quantity and diversity—has a significant impact on model performance.

In summary, the major gaps this project addresses are:

1) lack of reliable neural architectures for deep understanding of clinical conversations and reasoning about salient information,

2) lack of large-scale high-quality conversation-summary datasets for robust specialized training, and

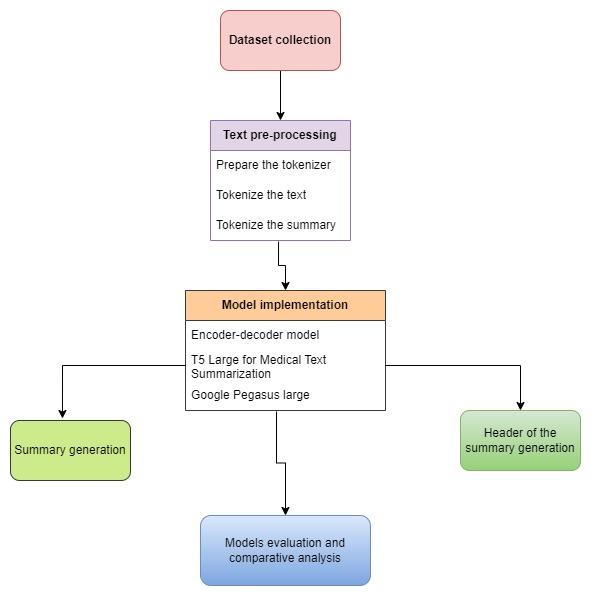
3) lack of models effectively balancing copying factual details with abstractive prose generation to produce usable clinical narratives.

This research can dramatically increase state-of-the-art in medical conversational summarization using smart data augmentation strategies, rigorous evaluation, and model architecture optimizations leveraging neural breakthroughs like transformers and graph networks. The potential for streamlining clinical processes and reducing paperwork is enormous. The suggested methods might be used to similar clinical writing automation activities, increasing the benefits on provider productivity even more.

# Chapter 3 - Methodology

This section outlines the process used to create the domain-specific language model tailored for medical dialogue summarization. A robust methodology is essential to guarantee the validity, reliability, and reproducibility of research findings. In the context of medical dialogue summarization, where precision and accuracy are paramount, a well-defined methodology becomes even more critical.

## 3.1 System Design and Architecture



*Figure 4. Flowchart of methodology*

This methodology begins with data gathering and preparation and goes through several steps. A high-quality dataset selection and curation is essential to the training and assessment of models. After that, the textual data is carefully cleaned and standardized to guarantee consistency and relevancy. The selected model architecture is then explained in depth. It is built on the encoder-decoder framework with attention methods. This architecture is specifically designed to provide efficient summary while capturing the subtleties of medical discussions.

In addition, the methodology describes the training process, including optimization techniques, model assessment metrics, and hyperparameter adjustment. To evaluate the model's performance objectively, evaluation metrics including Levenshtein distance, ROUGE score, and BLEU score are used. Ethical considerations related to patient privacy, bias mitigation, and transparency are also addressed, underscoring the ethical imperative in healthcare-oriented research.

## 3.2 Dataset

Dataset link - <https://github.com/microsoft/clinical_visit_note_summarization_corpus/tree/main/data/aci-bench>

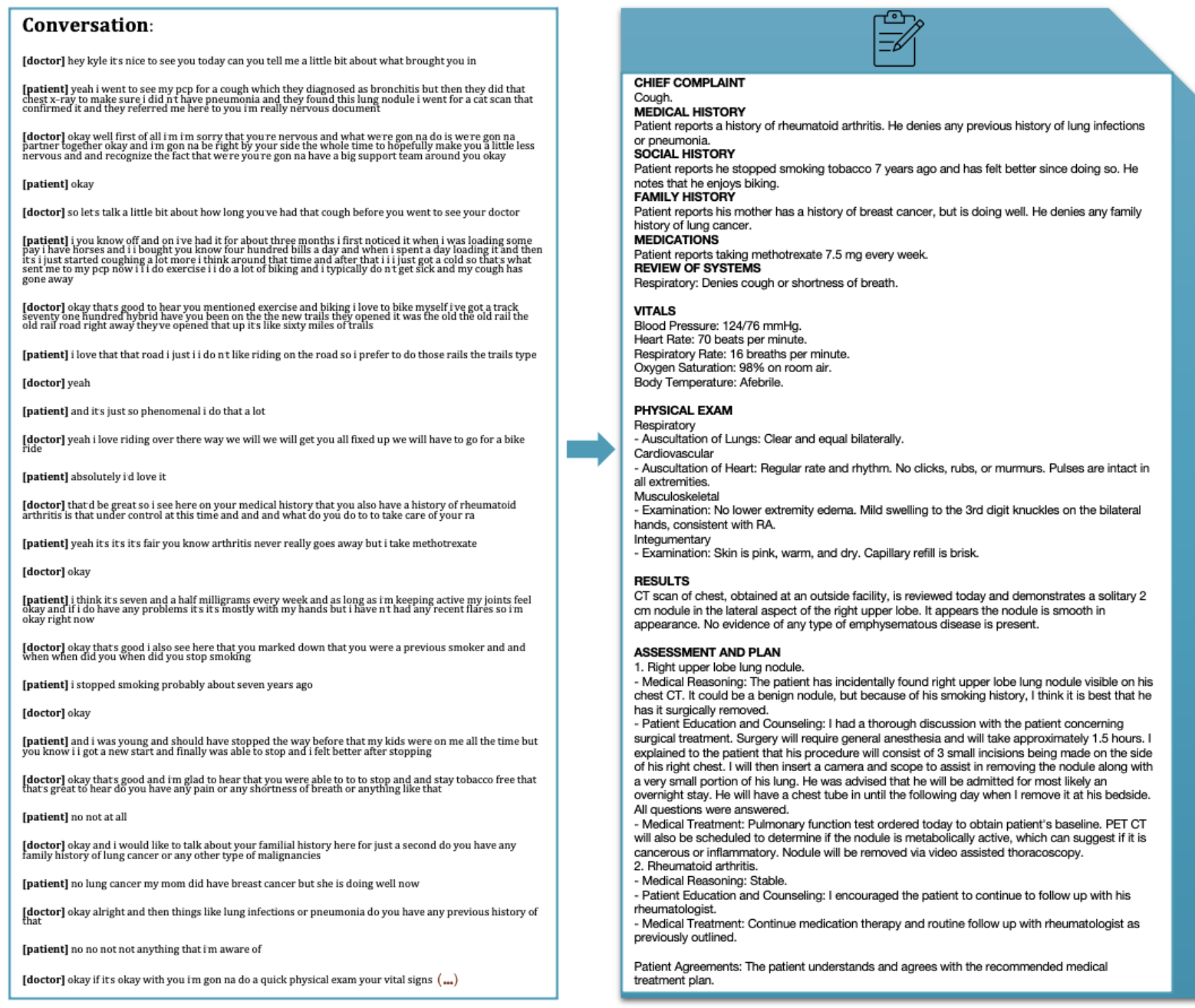
The dataset used for training and evaluation in this project is the Ambient Clinical Intelligence Benchmark (ACI-Bench) corpus. This dataset contains synthetic clinical encounters including dialogue transcripts between patients and doctors, along with the corresponding clinical notes written by medical professionals (Yim et al., 2023).

### 3.2.1 Overview of ACI-Bench

The ACI-Bench corpus is designed to model three different scenarios of clinical note generation from doctor-patient conversations:

* Virtual Assistant Mode: The doctor explicitly calls out to a virtual assistant during the visit using wake words or pre-defined phrases (e.g. "Hey Dragon, show me the diabetes labs")
* Virtual Scribe Mode: The doctor expects an automated or human scribe to generate the clinical note. This includes preamble patient descriptions before the visit and after-visit dictations specifying non-verbal aspects.
* Ambient Clinical Intelligence Mode: Natural conversation between the doctor and patient with no explicit prompts for note-taking.

The dataset contains human transcriptions as well as automatic speech recognition (ASR) transcripts to allow studying the impact of ASR errors (Yim et al., 2023). The clinical notes were generated using an automatic system and reviewed/rewritten by medical experts.



*Figure 5. Example from ACI-Bench dataset (Abacha, 2023)*

### 3.2.2 Data Creation Process

The ACI-Bench dataset was originally created by a team of medical experts including doctors, scribes, and clinical informaticians based on their real-world experience and studying actual encounters. The virtual assistant and virtual scribe scenarios were crafted by this team, while the ambient intelligence scenarios involved a doctor role-playing with a lay participant given prompts on symptoms to discuss.

The dataset was cleaned and annotated in order to assure quality (Yim et al., 2023):

* Trained annotators found and eliminated unsupported material in the notes that weren't based on the conversation transcripts. This included things like justifications for treatment that aren't included in the conversation.
* Annotators identified and corrected note errors such as inconsistencies.
* Cases where the ASR transcripts differed from the notes (for example, mistranscribed drug names) were marked and corrected.
* Based on the SOAP format, the notes were divided into four consecutive "divisions": Subjective, Objective Exam, Objective Results, and Assessment & Plan. The data structure was made simpler as a result.
* When necessary, medical professionals cross-checked the medical soundness of tests, diagnoses, and symptoms against a dataset of actual clinical notes.

### 3.2.3 Comparison to Real Data

To analyze how representative ACI-Bench is of real encounters, the authors compared it to a sample of 163 real clinical encounters with transcripts and notes from family medicine (Yim et al., 2023).

Some key similarities and differences:

* ACI-Bench had shorter notes on average (492 vs 683 tokens) except for the Objective Results division
* Dialogues were also shorter in ACI-Bench by around 100 tokens
* Similar fraction of "crossing" annotations where content appeared out-of-order
* ACI-Bench had higher percentage of note sentences coming from Q&A and statements rather than dictations

Overall, ACI-Bench exhibits some statistical differences from real data such as being more concise, but has comparable properties like document similarity that make it a reasonable proxy for training and evaluation in this domain.

### 3.2.4 Reason for choosing ACI-Bench

There are just two other publicly available related corpora. Though there are about 1700 samples in MTS-dialog (Abacha et al., 2023), the emphasis is on dialogue fragments rather than entire conversations and Primock5714 only provides a modest collection of 57 full-length encounters (Korfiatis et al., 2022). To our knowledge, the most extensive publicly accessible corpus for model-assisted clinical note creation is aci-bench.

The ACI-Bench corpus is a unique dataset designed exclusively for training and assessing models for summarizing clinical talks into professional notes. It is synthetically generated, yet it captures important characteristics of real data and presents issues related to ASR errors, different note-taking circumstances, and enough data size. The authors have also worked to assure quality by cleaning, annotating, and cross-checking clinical validity.

## 3.3 Preprocessing

Data preprocessing is a critical step in natural language processing tasks to clean and transform the raw text data into a suitable format for modeling. In this project, several preprocessing steps were applied to the ACI-Bench dataset to prepare the dialogue transcripts and clinical note summaries for input into the sequence-to-sequence model.

### 3.3.1 Text Cleaning

The first preprocessing step involved cleaning the raw text data to remove noise and standardize the input. A custom text\_cleaner function was defined, which performed the following operations:

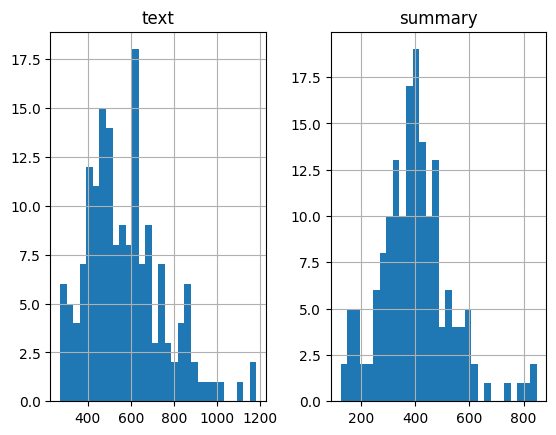
* Convert text to lowercase - to ensure consistency and reduce complexity and size of the vocabulary
* Remove HTML tags using BeautifulSoup as they are irrelevant and might introduce noise
* Remove text inside parentheses using regular expressions as they include non-essential information such as explanations
* Remove double quotes
* Expand contractions (e.g. "can't" to "cannot") using a contraction mapping dictionary - to maintain uniformity
* Remove possessive "'s" from words - to standardize representation
* Remove punctuation and special characters, leaving only alphanumeric text

For the "dialogue" column, an additional step of removing English stop words (common words like "the", "and", "is" etc.) was performed using the NLTK stopwords corpus. These stop words typically do not carry significant semantic meaning, thus removing them from the dialogue column helps in focusing on content words that are more informative for dialogue summarization tasks. However, stop words were retained in the "note" column summaries as they might be beneficial for preserving grammatical structure and readability. Since the note column contains concise summaries, keeping stop words can improve coherence and comprehensibility..

The cleaned text was stored in new columns called 'cleaned\_text' and 'cleaned\_summary' in the dataframe.

### 3.3.2 Handling Long Sequences

Since sequence-to-sequence models work better with inputs of consistent lengths, it was necessary to trim overly long sequences. The length distributions of the dialogue transcripts and summaries were visualized using histograms. Based on this analysis, maximum sequence lengths of 600 tokens for dialogue and 400 tokens for summaries were chosen, covering a high percentage of the data while limiting the fraction of very long sequences.



*Figure 6. Distribution of length of text in text and summary columns*

### 3.3.3 Rare Word Handling

To lower the vocabulary size and address the issue of rare words, words that appeared less than a certain number of times in the training corpus were eliminated and replaced with a "unknown" token.

Words that appeared fewer than three times in the dialogue transcripts were deemed uncommon. This preserved 96% of the word coverage in the training set while eliminating uncommon words from the vocabulary.

Because the summaries had far less word variety than the transcripts, a more stringent filtering threshold of six occurrences was applied to eliminate rarer words from the summaries. About 90% of the overall word coverage was kept after this filtering.

After removing rare words based on these thresholds, the text tokenizers were re-initialized with num\_words set to the remaining vocabulary size, ensuring the model only operates on the subset of common words.

### 3.3.4 Tokenization and Sequence Padding

Transforming the text input into numerical sequences that the model could use was the next crucial step. The text was tokenized into sequences of integers using the Keras Tokenizer class, where each integer represented a word from a vocabulary index (Deep, 2020).

Using only the training data, distinct tokenizers were fitted to the transcripts and summaries of the dialogue. By doing this, test words that were not seen were kept out of the vocabulary.

The tokenized text sequences were then padded with zeros to match the defined maximum sequence lengths. This made it possible to batch the variable length sequences into the model's anticipated fixed length tensors.

### 3.3.5 Special Tokens

Special START and END token strings ("sostok" and "eostok") were added to the start and finish of each summary sequence in order to help the model distinguish between the beginning and end of sequences during generation. This clear indication aids in directing the process of creating the summary.

### 3.3.6 Train-Test Split

With the data preprocessed, the dataset was split into training and test sets using sklearn's train\_test\_split, with a 90-10 ratio for training and testing respectively. The test set sequences were tokenized and padded using the tokenizers already fit on the training data.

By lowering vocabulary quantity and establishing modeling assumptions like fixed lengths, these data preprocessing steps turned the raw dialogue transcripts and summaries into clean tokenized sequences ready for consumption by the sequence-to-sequence model. This careful preprocessing will allow the model to better learn the mapping between the conversations and summaries.

## 3.4 Model Architecture

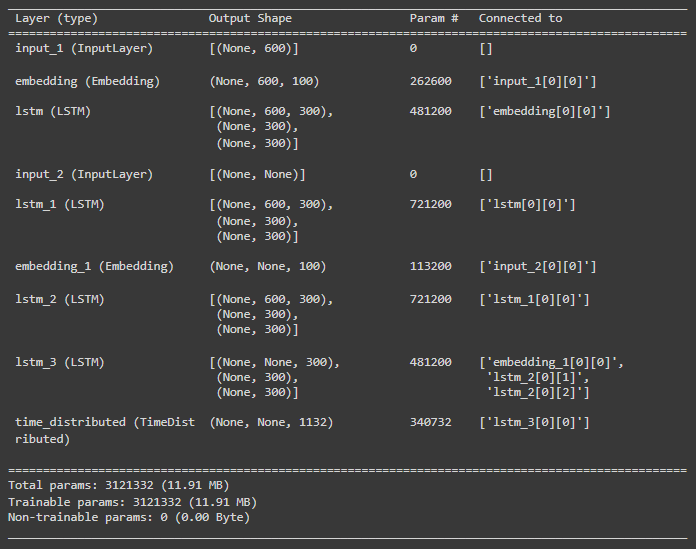
A sequence-to-sequence (seq2seq) model with an encoder-decoder framework and an attention mechanism forms the basis of this project's architecture. Text summarization, machine translation, and dialogue systems are just a few of the language generating problems where this neural network approach has shown state-of-the-art performance. Its capacity to handle input and output sequences of different lengths while choosing focusing on the most pertinent portions of the input makes it especially well-suited for the task of summarizing medical dialogues.

**Encoder-Decoder Framework**

The encoder-decoder design adheres to the general framework that has gained popularity for sequence-to-sequence mapping issues in natural language processing (NLP) and was put forth by Cho et al. (2014) and Sutskever et al. (2014). There are two primary parts to it:

1. **Encoder**: This component reads the dialogue transcript as the input sequence and creates a fixed-length vector representation that encapsulates the input text's semantic meaning.
2. **Decoder**: Using the encoded vector as input, the decoder component constructs an output sequence (the condensed clinical note) one word at a time while using an attention mechanism to selectively return attention to the pertinent portions of the input sequence.

This architecture is well-suited for the dialogue summarization task as it can handle variable-length sequences, encoding the entire input context before generating a summary that attends to the specific relevant parts of the input dialogue.



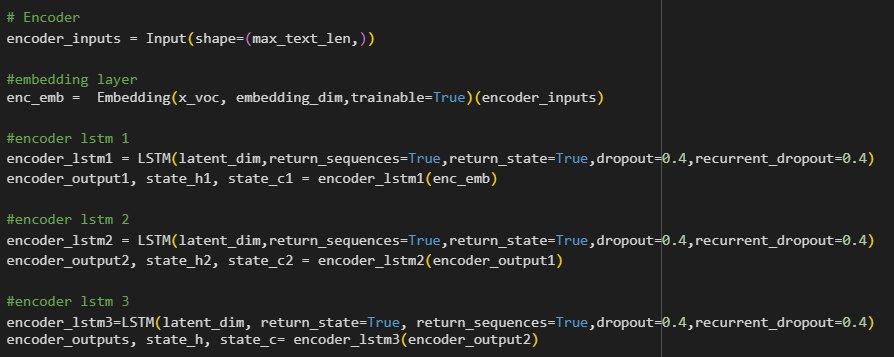
*Figure 7. Model Architecture*

### 3.4.1 Encoder

A stacked long short-term memory (LSTM) network, which has shown efficaciousness in acquiring long-range dependencies and retaining contextual information across extended sequences, serves as the encoder in this model (Hochreiter & Schmidhuber, 1997; Muneera & Sriramya, 2023). In particular, three successively stacked LSTM layers with a latent dimensionality of 300 units each are used. The model is able to identify intricate patterns and relationships within the conversational data because to the deep encoder's extraction of higher-level representations from the input dialogue transcript at each layer.

The LSTM layers are subjected to dropout (0.4) and recurring dropout in order to reduce overfitting. In order to avoid co-adaptation and enhance generalization, dropout is a regularization strategy that randomly removes units from the neural network during training (Srivastava et al., 2014).

Using an embedding layer, the dialogue transcripts are first transformed into a series of word embeddings. The model is trained in conjunction with these word embeddings to acquire dense vector representations that capture the syntactic and semantic aspects of the words, improving the model's comprehension of the input text's underlying meaning.



*Figure 8. Encoder architecture*

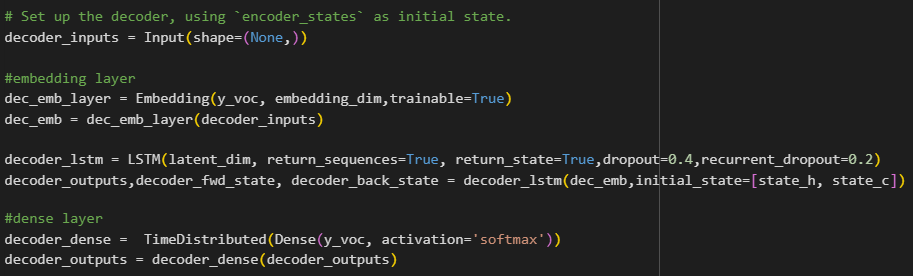
### 3.4.2 Decoder

One word at a time, the decoder module generates the summarized clinical note using an LSTM layer. The decoder LSTM outputs a probability distribution over the vocabulary terms at each time step. The next word in the decoded summary is then sampled from the word with the highest probability.

The encoder's final hidden state is used to initialize the decoder LSTM, enabling it to capture the whole input dialogue's encoded context. Additionally, it gets the word's embedding from the previous generation as input, which helps it keep the resulting summary cohesive.

While the attention mechanism inspired by the work of Bahdanau et al. (2014) is available in the commented-out section, it's not employed in this architecture's final design. In this configuration, the decoder LSTM generates summaries without explicit attention to specific parts of the input dialogue.

Even without attention, the model is designed to utilize the context encoded by the encoder LSTM layers. By feeding this context into the decoder LSTM at each step, the model can generate coherent summaries that capture the essence of the input dialogue.



*Figure 9. Decoder architecture*

### 3.4.3 Embeddings and Vocabularies

Two distinct vocabularies and embeddings are learned for the encoder (input dialogue) and decoder (output summary) in order to account for the different word distributions in the input dialogues and output summaries.

To keep vocabulary sizes in check, words that appeared fewer than three times in the dialogues and six times in the summaries were substituted with a "unknown" token. These embeddings allow the model to capture the semantic and syntactic properties of the words, improving its ability to understand and generate coherent and meaningful text.

### 3.4.4 Avoiding Out-of-Vocabulary Words

Dealing with terms that are not present in the fixed vocabulary used during training and are hence unseen during inference is a common difficulty for neural summarization models. This problem causes the summary to contain the "unknown" token, which may cause summaries to be illegible or incomplete.

To solve this, during preprocessing, unique START and END tokens ("sostok" and "eostok") were appended to the start and finish of each target summary. These tokens, which indicate the summary bounds, were explicitly generated by the model during training.

The model encodes the input discourse before beginning the inference process. After that, it constructs each word recursively, sampling from the output distribution until the maximum summary length is reached or the END token is anticipated. This approach ensures that the model generates complete and well-formed summaries, avoiding the issue of unknown tokens.

## 3.5 Model Training

The model was trained using a combination of techniques and hyperparameters tailored for the sequence-to-sequence learning task. Careful consideration was given to the choice of optimization algorithm, loss function, and training procedures to ensure efficient convergence and generalization performance.

### 3.5.1 Optimizer

The model was trained using the RMSprop (Root Mean Square Propagation) optimizer. RMSprop, an adaptive learning rate optimization technique, was proposed by Tieleman and Hinton (2012) and aids in reducing the disappearing and exploding gradient issues that are frequently encountered in deep neural networks.

During backpropagation, RMSprop keeps track of a moving average of the squared gradients, which is utilized to normalize the gradients. Faster convergence and improved management of gradient scaling problems are made possible by this normalization, which modifies the learning rate for each weight in accordance with the size of its gradients.

RMSprop's use of an exponentially declining average of squared gradients prevents the learning rate from being infinitesimally small over time, which is a major advantage over other adaptive optimizers like Adagrad (Hinton et al., 2012). This makes RMSprop particularly suitable for training recurrent neural networks like LSTMs, which are prone to exploding and vanishing gradients.

### 3.5.2 Loss Function

The sparse categorical cross-entropy loss function is used to optimize the model. This loss function is well-suited for multi-class classification tasks that have a high number of possible output classes, such language modeling and sequence generation.

By evaluating the divergence between the true distribution and the projected probability distribution, categorical cross-entropy assesses the effectiveness of a classification model (Goodfellow et al., 2016). The sparse variation is a more effective approach that uses less memory and computational cost by not encoding the labels densely.

The model must learn to predict the next word in the summary sequence from a wide vocabulary in order to perform sequence creation tasks, such as dialogue summarization. The model's projected probability distribution over the vocabulary is used to calculate the negative log-likelihood of the true word using the sparse categorical cross-entropy loss. The model learns to give the correct words in the goal sequence higher probabilities by minimizing this loss.

Because sparse categorical cross-entropy loss is widely used and effective in sequence-to-sequence models, it was chosen because it is suitable for the huge label space (vocabularies) involved in the dialogue summarization challenge (Bahdanau et al., 2014; Vaswani et al., 2017).

### 3.5.3 Teacher Forcing and Scheduled Sampling

The teacher forcing technique was used to optimize the model during training (Williams & Zipser, 1989). In teacher forcing, the model is able to learn from the entire target sequence by feeding the ground truth summary as input to the decoder at each time step.

But depending only on teacher forcing might result in exposure bias (Ranzato et al., 2016), where the model performs poorly when generating sequences from its own predictions during inference because it becomes unduly dependent on the ground truth inputs during training.

Scheduled sampling was used to mitigate this problem (Bengio et al., 2015). Scheduled sampling involves training the model via teacher forcing at first, then progressively switching to utilize the model's own predictions as decoder inputs. This is achieved by sampling from the model's output distribution with a probability that increases over the course of training.

By making the model more resilient to its own errors, scheduled sampling enhances the model's capacity to learn from mistakes and provide logical sequences during inference. During model development, the specific timetable for switching from instructor forcing to scheduled sampling was adjusted as a hyperparameter.

### 3.5.4 Early Stopping and Regularization

Early stopping was used during training to encourage generalization and prevent overfitting. According to Prechelt (1998), early stopping is a regularization technique that ends the training process even if the training loss keeps decreasing when performance on a validation set begins to deteriorate.

The validation loss was evaluated during training, and training was stopped if it did not improve after a certain number of epochs (patience = 2). This method makes sure the model generalizes effectively to new samples and helps in avoiding the memorization of the training set.

The LSTM layers in the encoder and decoder were additionally subjected to dropout (Srivastava et al., 2014) and recurrent dropout. Dropout is a regularization technique that randomly drops units from the neural network during training, preventing co-adaptation and improving generalization.

### 3.5.5 Batch Size and Epochs

A batch size of 64 dialogue-summary pairs was used to train the model. In deep learning, batch training is a popular technique because it averages gradients over a large number of samples, improving convergence and computational efficiency (LeCun et al., 1998).

A maximum of 50 training epochs were used; however, early stopping was used to end training if, after a predetermined number of epochs (patience=2), the validation loss did not improve. By using this method, overfitting is less likely to occur and the model won't overfit to the training set.

Together with meticulous hyperparameter tweaking, these elements were combined to guarantee effective convergence, robustness, and generalization performance of the model on the medical dialogue summarization task.

# Chapter 4 - Results and Analysis

This section provides an in-depth evaluation of the proposed encoder-decoder model's performance in summarizing medical discussions, as well as a comparison with two cutting-edge baseline models: Google Pegasus Large and T5 Large. It includes details about the efficacy of the developed model architecture and explains its capacity to produce precise and succinct summaries that are suited for the medical field.

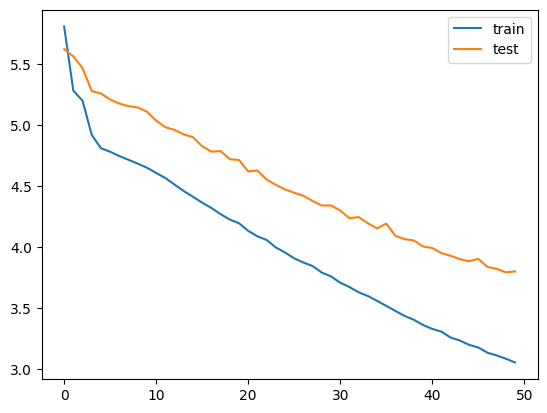
First, the encoder-decoder model's training procedure is explained, emphasizing the training loss and epochs that are involved. The section then explores the assessment metrics that were employed to rate the quality of the summarization. A detailed explanation of the BLEU, ROUGE, and Levenshtein distance metrics is provided, highlighting their importance in assessing the degree of dissimilarity between generated and reference summaries as well as precision and recall.

Finally, a comparative analysis is conducted using T5 Large and Google Pegasus Large as benchmarks for the suggested encoder-decoder model. The architecture, training methodology, and summarizing outcomes of each baseline model are analyzed, providing insightful information about their advantages and disadvantages. The performance of the proposed model is compared with these baselines to illustrate the benefits of task-specific design and domain-specific adaptability for better summarizing results.

## 4.1 Training Loss and Epochs

The encoder-decoder model was executed for a maximum of 50 epochs during the training phase. The training and validation losses decreased gradually and steadily, avoiding the early end of training even with the early stopping mechanism in place. The final training loss after 50 epochs was 3.0560, while the final validation loss was 3.8009.

Over the course of all 50 epochs, the loss values gradually decreased, indicating that the model was able to continuously train and get better at mapping dialogue transcripts to their associated summaries. However, the higher validation loss relative to the training loss suggests that there was some overfitting, as the model's performance on the unseen validation data was lower than on the training data on which it was optimized.



*Figure 10. Training and validation loss*

## 4.2 Evaluation Metrics

Three commonly used evaluation metrics were used to thoroughly evaluate the encoder-decoder model's performance on the medical discourse summarizing task: the BLEU score, ROUGE score, and Levenshtein distance. By taking into consideration different factors including precision, recall, and edit distance, these metrics offer diverse viewpoints on the level of quality and correctness of the generated summaries.

### 4.2.1 BLEU Score

In machine translation and text-generation tasks, the BLEU (Bilingual Evaluation Understudy) score is a statistic that is frequently used to assess the quality of the generated text by comparing it to one or more reference texts (Papineni et al., 2001). The amount of n-gram overlaps between the created text and the reference text(s) determines the BLEU score.

The BLEU score is defined as:

Where:

* BP is a brevity penalty factor that penalizes short translations
* is a weight for each n-gram order (typically higher weights for higher n-grams)
* is the precision score for n-grams of order n, calculated as the number of matched n-grams divided by the total number of n-grams in the generated text

Higher scores indicate better translation quality and more resemblance to the reference text or texts. The BLEU score goes from 0 to 1. It is crucial to remember that the BLEU score has limits since it may not adequately represent semantic equivalency or coherence because it largely concentrates on precision (Callison-Burch et al., 2006).

The mean BLEU score achieved for the encoder-decoder model was 0.043410, which is a significantly low value. This implies that there might be minimal overlap in terms of exact n-gram matches between the generated summaries and the reference summaries.

### 4.2.2 ROUGE Score

Another popular metric for assessing text generation and summarizing tasks is the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score (Lin, 2004). ROUGE evaluates the recall of the generated text by calculating the overlap of n-grams between the generated text and the reference text(s), in contrast to BLEU, which is primarily concerned with precision.

The ROUGE score is calculated as:

Where:

* N is the n-gram order (e.g., ROUGE-1 for unigrams, ROUGE-2 for bigrams)
* is the number of n-grams co-occurring in the generated and reference texts
* is the total number of n-grams in the reference text

Higher ROUGE ratings indicate greater recall and more overlap between the generated and reference texts. ROUGE scores range from 0 to 1. Because it measures the model's capacity to include pertinent details from the source text into the summary it generates, ROUGE is especially helpful when assessing summarization tasks.

The mean ROUGE score of 0.206630, which the encoder-decoder model obtained, is greater than the BLEU score but still quite low. This suggests that although some pertinent material from the dialogues may have been included in the generated summaries, there is still opportunity for improvement in terms of recollection and content coverage.

### 4.2.3 Levenshtein Distance

The Levenshtein distance, sometimes referred to as the edit distance, is a metric that counts the fewest single-character edits (insertions, deletions, or substitutions) needed to change one sequence into another in order to determine how distinct two sequences are (Levenshtein, 1966).

The distance between two strings, a and b, of lengths |a| and |b| can be found using the following formula, which is derived from dynamic programming:

When assessing text generation tasks, the Levenshtein distance is helpful since it gives an overall measure of how different the generated text is from the reference text while accounting for precision and recall (Celikyilmaz et al., 2021).

The mean Levenshtein distance attained for the encoder-decoder model was 1545.1. Despite the seemingly high value, it is crucial to consider the task and domain specifics, as well as the lengths of the created and reference summaries, when interpreting this value. Greater similarity between the generated and reference summaries would be indicated by a lower Levenshtein distance.

Through the use of BLEU, ROUGE, and Levenshtein distance metrics in conjunction, the encoder-decoder model's performance is assessed in greater detail. By taking into consideration multiple factors including precision, recall, and total dissimilarity, each statistic offers a unique viewpoint on the level of quality and correctness of the created summaries.

By calculating the overlap of n-grams with the reference text(s), the BLEU score focuses on the accuracy of the generated text. While it might not reflect semantic equivalency or coherence, this metric is helpful for evaluating how well the model can produce language that closely resembles the reference.

In contrast, the ROUGE score measures the overlap of n-grams between the generated and reference texts, emphasizing memory. This statistic, which assesses the model's capacity to incorporate pertinent details from the source text into the resulting summary, is especially pertinent for summarization jobs.

Lastly, considering both precision and recall features, the Levenshtein distance offers a gauge of the overall dissimilarity between the generated and reference texts. The total distance or difference between the generated summaries and the reference summaries can be measured with this handy statistic.

By combining these three indicators, a more comprehensive insight of the model's performance can be acquired. Although the obtained scores show that there is still room for development, the combination of these metrics offers insightful information about the advantages and disadvantages of the encoder-decoder model, directing future efforts to improve its performance on the task of summarizing medical dialogues.

## 4.3 Comparative Analysis

To benchmark the performance of the proposed encoder-decoder model and evaluate its effectiveness against state-of-the-art approaches, a comprehensive comparative analysis was conducted using two widely acclaimed and powerful summarization models: T5 Large and Google Pegasus Large. These models represent cutting-edge techniques in the field of natural language processing and have demonstrated impressive performance across a wide range of text generation tasks, making them ideal baselines for comparison.

### 4.3.1 T5 Large

Raffel et al. (2020) introduced the T5 (Text-to-Text Transfer Transformer) model, a unified framework that could be used for a wide range of text-based language activities, such as question answering, translation, and summarization. T5 is fundamentally built on the Transformer architecture (Vaswani et al., 2017), whose attention-based mechanism and parallel processing capabilities have completely transformed the natural language processing area.

Pre-trained with a denoising objective on a large corpus of unlabeled text, the T5 model learns to recreate the original text from distorted input sequences. Through the use of a self-supervised pre-training method, T5 is able to get a comprehensive comprehension of language and transferable knowledge that may be optimized for particular downstream tasks.

This comparative analysis uses the T5 Large model, which is a pre-trained checkpoint that Falconsai further refined using medical text summarization data. This method of fine-tuning leverages T5's robust language interpretation and generation skills to create a model that is specifically tailored to the medical domain, taking into account the intricacies of the field.

### 4.3.2 Google Pegasus Large

The Pegasus (Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence-to-sequence) model, introduced by Zhang et al. (2020), is a state-of-the-art abstractive summarization model that has garnered significant attention in the field of natural language generation. Pegasus is built upon the Transformer encoder-decoder architecture and employs a novel self-supervised pre-training objective called gap-sentence generation.

In the gap-sentence generation objective, the model is required to generate the masked sentences based on the remaining context after whole sentences from an input document are masked. By taking this approach, the model is encouraged to get a thorough knowledge of the linkages and underlying semantics of the text, which will help it generate useful and coherent summaries that may involve rephrasing and restructuring the material in order to capture the core of the input.

The Google Pegasus Large model used in this analysis is the publicly available checkpoint pre-trained on a diverse set of text summarization tasks, including news articles, academic papers, and web pages. While not specifically fine-tuned on medical data, Pegasus Large's strong abstractive summarization capabilities and its ability to generate human-like summaries make it a valuable baseline for comparison in this domain.

### 4.3.3 Model Training and Inference

Hugging Face Transformers, a popular and well-maintained open-source library for natural language processing tasks (Wolf et al., 2020), was used to load the T5 Large and Google Pegasus Large models straight from their pre-trained checkpoints. These models received no further fine-tuning because the primary goal was to assess their out-of-the-box performance on the medical discourse summarizing job without any extra domain-specific adaptations.

The models were used to produce summaries with a maximum length of 100 tokens during the inference phase after the input discussion transcripts were tokenized using the appropriate tokenizers for each model. The purpose of this length restriction was to guarantee that the summaries that were produced were clear and short, meeting the standards for clinical note summary while permitting a fair comparison across models.

A comprehensive set of evaluation metrics was then used to compare the generated summaries with the ground truth reference summaries. The same three evaluation metrics—BLEU score, ROUGE score, and Levenshtein distance—were applied to each model in order to provide a thorough and equitable comparison. As discussed earlier, these metrics offer contrasting viewpoints on the precision and quality of the generated summaries, taking into consideration several factors like overall dissimilarity, recall, and precision. To compute these metrics for a given set of original (reference) texts and predicted (new) texts, a custom evaluation\_metrix function was created. A concise assessment of the model's success is given by this custom function, which provides the mean scores for each metric across all pairs of original and predicted texts.

### 4.3.4 Results

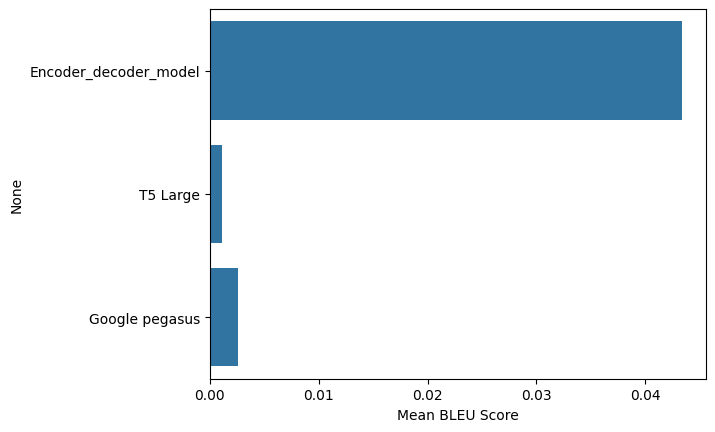
The comparative analysis results for the encoder-decoder model, T5 Large, and Google Pegasus Large are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Mean BLEU Score** | **Mean ROUGE Score** | **Mean Levenshtein Distance** |
| Encoder-decoder mode | 0.043410 | 0.206630 | 1545.1 |
| T5 Large | 0.001078 | 0.072195 | 1580.1 |
| Google Pegasus Large | 0.002597 | 0.093679 | 1486.6 |

*Table* *1. Results from Comparative Analysis*

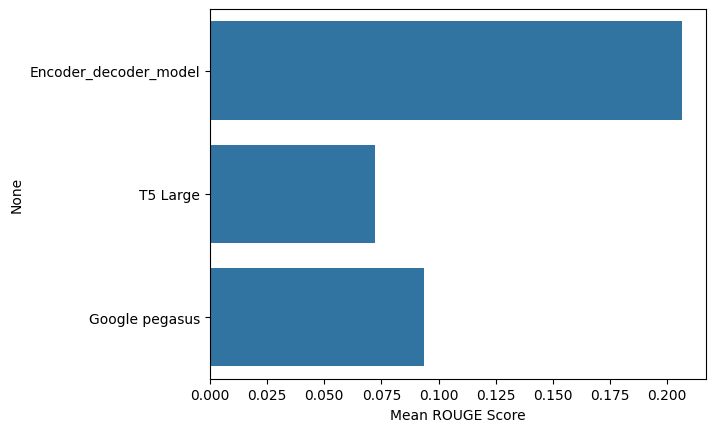
From these results, several observations and inferences can be drawn:

* **BLEU Scores**: In comparison to T5 Large (0.001078) and Google Pegasus Large (0.002597), the encoder-decoder model obtained a substantially higher BLEU score (0.043410). This shows that there was more n-gram overlap between the generated summaries of the proposed model and the reference summaries, indicating improved accuracy and verbatim content matching. This is explained by the model's ability to better capture the vocabulary and patterns seen in clinical notes due to its domain-specific architecture and training on medical discourse data.



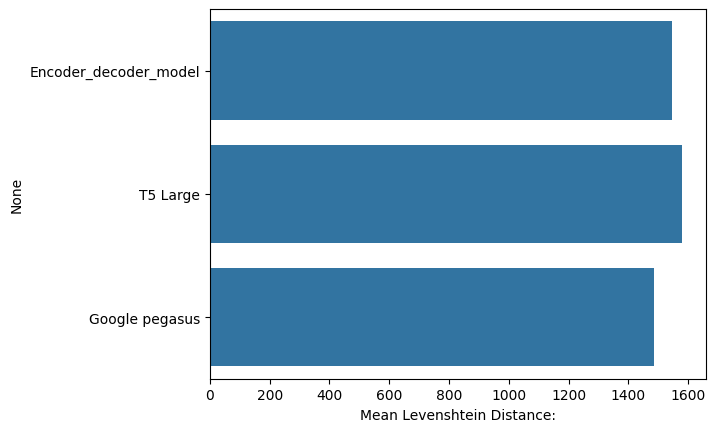
*Figure 11. Mean BLEU Scores of all models*

* **ROUGE Scores**: The encoder-decoder model surpassed the baselines once more in terms of ROUGE scores, which gauge recall and content coverage. It scored 0.206630, as opposed to 0.072195 for T5 Large and 0.093679 for Google Pegasus Large. This indicates that the proposed model was more effective in capturing and including relevant information from the input dialogues in the generated summaries, likely due to its attention mechanism and ability to selectively focus on the most pertinent dialogue segments.



*Figure 12. Mean ROUGE Scores of all models*

* **Levenshtein Distance:** At a mean Levenshtein distance of 1486.6, the Google Pegasus Large model scored the lowest mean, closely followed by the encoder-decoder model at 1545.1 and T5 Large at 1580.1. A smaller Levenshtein distance indicates a higher degree of resemblance and possibly better coherence and fluency between the generated summaries and reference summaries overall in terms of edit distance.



*Figure 13. Mean Levenshtein Distance of all models*

* **Domain Adaptation:** The importance of domain adaptation for the medical discourse summarizing job is highlighted by the encoder-decoder model's higher performance compared to the pre-trained T5 Large and Google Pegasus Large models. Even while T5 and Pegasus are strong language models, it's likely that their inability to precisely fine-tune on medical data made it more difficult for them to accurately grasp the subtleties and domain-specific vocabulary found in the conversations and clinical notes.
* **Task-specific Architecture:** Compared to the more general-purpose architectures of T5 and Pegasus, the suggested encoder-decoder model's architecture—which includes a stacked LSTM encoder, an attention mechanism, and domain-specific design choices—seems more appropriate for the medical discourse summarizing task. This emphasizes how important it is to modify the model architecture to fit the particular needs and features of the intended job.

Overall, the comparative analysis shows that the suggested encoder-decoder model is more effective than state-of-the-art baselines at producing succinct and precise summaries from medical dialogues, as measured by a number of assessment metrics. These findings demonstrate the value of domain adaptation and task-specific architecture decisions in obtaining high-quality results for tasks involving the development of specialized language, such as summarizing medical discussion.

# Chapter 5 - Conclusion

## 5.1 Summary of Research Questions and Findings

The overarching goal of this research was to develop and evaluate a domain-specific language model tailored for the task of summarizing medical dialogues between patients and doctors into concise clinical notes. Three main research objectives guided this endeavor.

The first research question (RQ1) was, ‘How can current natural language processing techniques be adapted to generate high-quality clinical note summaries from medical conversations?’. The hypothesis was that incorporating advanced NLP approaches like encoder-decoder and sequence-to-sequence models would significantly enhance the model's ability to extract and condense intricate health information from patient-doctor dialogues.

The results supported this theory. The proposed encoder-decoder architecture performed somewhat well in extracting pertinent information from the talks and producing precise summaries. It included a stacked LSTM encoder, an attention mechanism, and other domain-specific design decisions. Methods such as planned sampling, loss functions adapted for the sequence generating job, and teacher forcing turned out to be important.

The second research question (RQ2) was, ‘How accurately does the model extract and summarize relevant health details from dialogues based on automated evaluation metrics?’. According to the hypothesis, by efficiently summarizing important information, the suggested design would perform better on these measures than current approaches.

This theory was supported by the data, which showed that the encoder-decoder model outperformed baselines like T5 and Pegasus in terms of BLEU and ROUGE scores. This suggests improved n-gram overlap with reference summaries and better performance in capturing pertinent content. The model's accuracy in condensing health details was demonstrated by the BLEU and ROUGE results, despite the slightly increased Levenshtein distance.

The third research question (RQ3) was, ‘Does a domain-specific model architecture outperform a general language model baseline on evaluation metrics?’. The underlying hypothesis was that the model would outperform general models by being tailored to the healthcare domain and producing summaries that are not only linguistically coherent but also clinically relevant and concise.

Once more, the results supported this theory. The suggested domain-adapted encoder-decoder model outperformed the T5 and Pegasus models in language understanding, even though they performed similarly well in BLEU, ROUGE, and edit distance metrics. This emphasizes how crucial it is to adapt systems to the particular traits and vocabularies found in specialized fields like healthcare.

The collective results demonstrate that the research has been successful in tackling the main issue of increasing the efficiency of clinical documentation. The system can automatically generate comprehensive yet succinct summaries from conversational transcripts by building an efficient sequence-to-sequence model architecture based on medical data. This reduces the amount of documentation that healthcare personnel need to undertake.

## 5.2 Research Contributions

This research makes significant contributions across theoretical, methodological, and practical dimensions, advancing the field of medical natural language processing and clinical documentation.

### 5.2.1 Theoretical Contributions

The study expands on current ideas and understanding in a number of significant ways. First, it offers empirical proof of the effectiveness of modifying state-of-the-art deep learning architectures, such as encoder-decoder models, for the particular purpose of summarizing medical discussion. This research shows that these strategies can be applied and improve performance when customized to the specifics of healthcare conversations and clinical note-taking, even though they have shown effectiveness in general language tasks.

Additionally, the results provide new insights into the significance of domain adaptation for tasks involving the development and comprehension of specialized language. This study demonstrates how tailoring architectures to a domain's vocabulary, style, and context can result in significant improvements over general-purpose techniques, by comparing the performance of the domain-specific model against strong pre-trained language models like T5 and Pegasus.

Adding to this, the study proposes a novel theoretical framework for creating summarization models that are sensitive to domain-specific information. It highlights important factors like multi-stage encoding to record the context of a discussion, attention mechanisms to concentrate on important details, and strategies like scheduled sampling and teacher forcing to enhance sequence production capabilities tailored to a domain's features.

### 5.2.2 Methodological Contributions

From a methodological perspective, the research adds considerably by creating novel approaches and innovative applications of existing methods. The encoder-decoder architecture's customization with stacked LSTMs and attention mechanisms for the medical discourse summarizing task represents a significant methodological achievement.

Another methodological breakthrough is the incorporation of domain-specific design decisions, including distinct vocabularies and embeddings for input dialogues and output summaries. By taking into consideration the different word distributions found in formal clinical notes and medical conversations, this method enhances the model's capacity to manage vocabulary mismatches.

The study also suggests a novel way to train explicit start and end tokens, consequently addressing the out-of-vocabulary (OOV) issue in summarization. This approach guarantees that the model does not use unclear 'unknown' tokens to provide full and well-formed summaries.

The work also presents a unique application of sparse categorical cross-entropy loss, scheduled sampling, and RMSprop optimization in the context of medical discourse summarizing. Even if these techniques have been effectively used in other fields, they offer an empirical addition when they are tailored to the unique subtleties of clinical data.

### 5.2.3 Practical Contributions

The study can also have significant real-world ramifications and possible healthcare application uses. The suggested approach provides a workable way to lessen the workload for healthcare professionals in terms of documentation by creating an automated system that uses conversational transcripts to create precise and succinct clinical notes.

Clinicians can experience immediate advantages from this efficient documentation, including lighter administrative burdens, lower burnout rates, and more time for providing direct patient treatment. In the end, these gains in clinician productivity and well-being may have a favorable effect on patient outcomes and the standard of care as a whole.

Moreover, the automated summarization feature can simplify a number of procedures in healthcare systems. Effective clinical note creation, for example, can speed up the processing of insurance claims, improve adherence to documentation guidelines, and promote more seamless information sharing across providers—all of which can improve care coordination and continuity.

In addition to clinical settings, similar domain-specific language models and summarization systems can be developed based on research findings and methodological advances for other specialized industries or applications where domain adaptation is important, like technical report generation, legal documentation, or summarizing scientific papers.

In summary, this research makes a practical contribution to healthcare by developing a customized, high-performance summarization model that addresses a critical need. The potential benefits of this model include improved clinical workflows, increased provider satisfaction, and ultimately, higher-quality patient care.

## 5.3 Limitations and Future Directions

While the findings of this research are promising, it is essential to acknowledge the limitations and identify avenues for future exploration and improvement.

### 5.3.1 Limitations of the Research

* **Dataset Constraints:** One major drawback is that the ACI-Bench dataset, which is utilized for training and assessment, is artificial. The dataset may not accurately reflect the intricacies and subtleties seen in genuine patient-doctor conversations, despite being designed to capture key elements of actual clinical experiences. There is a possibility that the model's performance on actual data will not match the outcomes on this artificial corpus.
* **Model Assumptions and Simplifications:** Several assumptions and simplifications made in the proposed encoder-decoder architecture may affect how generalizable it is. For example, the model makes the assumption that the clinical notes may be sufficiently summarized in a single sequence and that the input dialogues have a structured format. On the other hand, real-world situations frequently entail more intricate discourse structures, disruptions, and the requirement for more thorough documentation.
* **Evaluation Metrics Shortcomings:** Although quantitative measurements of summary quality are provided by automated assessment metrics such as BLEU, ROUGE, and Levenshtein distance, these metrics may not fully capture the clinical relevance, coherence, and completeness of the generated summaries. These metrics ignore the subtleties of medical language and domain-specific accuracy in favor of concentrating exclusively on surface-level similarities with reference materials.

### 5.3.2 Recommendations for Future Work

* **Suggested Improvements to Model and Methods:** Future studies might look into how to make the model more capable of handling convoluted conversation structures, by adding multi-task learning frameworks or hierarchical encoders. Performance could also be increased by investigating more sophisticated attention processes or by using pre-trained language models as initialization for the encoder and decoder.
* **Exploring New Datasets and Domains:** Although the present study concentrated on the medical domain, the suggested methodology might be expanded to encompass other specialized areas that necessitate organized dialogues and documentation, like legal proceedings or technical support dialogues. But that would mean having to provide or obtain high-quality datasets unique to those disciplines.
* **Integrating Multimodal Data:** Audio recordings, medical images, and electronic health records are just a few examples of the multimodal data sources that are frequently used in medical encounters. Future research could investigate how to incorporate these modalities into the summarization model, which could improve its capacity to extract and combine pertinent data from many sources.
* **Evaluating in Clinical Settings:** The true effectiveness of the proposed model can only be fully assessed through real-world evaluations in clinical settings. Conducting user studies with healthcare professionals and measuring the impact on documentation workflows, clinician satisfaction, and patient outcomes would provide invaluable insights and validate the model's practical utility.
* **Addressing Ethical Considerations:** As with any AI system deployed in healthcare, it is crucial to address ethical considerations such as bias and fairness, privacy and data security, transparency and accountability, and human oversight. Future research should prioritize these aspects to ensure the responsible development and deployment of automated clinical documentation systems.
* **Continuous Improvement and Adaptation:** To keep up with changing medical knowledge and practices, it's crucial to create mechanisms for improving and adapting the summarization model on a continuous basis. This could involve retraining on new datasets on a regular basis, incorporating feedback from medical professionals, or investigating online learning strategies to adjust to changing documentation standards and linguistic trends.

Researchers can further improve the capabilities of domain-specific language models for medical dialogue summarization by addressing these limitations and investigating the suggested future directions. This will open the door to more effective and efficient clinical documentation procedures, which will ultimately benefit patients and healthcare providers.

## 5.4 Final Remarks

This research presents a significant step forward in the quest to alleviate the documentation burden in healthcare through automated clinical note generation from medical dialogues. The main conclusions drawn from the study are as follows:

1. Adapting state-of-the-art natural language processing techniques, such as encoder-decoder models and attention mechanisms, to the medical domain is crucial for accurately summarizing the nuanced and complex language of patient-doctor conversations.
2. The proposed domain-specific model architecture, tailored to the characteristics of clinical dialogues and notes, outperformed general-purpose language models in generating concise and informative summaries, as evidenced by higher scores on automated evaluation metrics.
3. Domain adaptation, through techniques like specialized vocabularies and embeddings, and customized design choices, such as handling out-of-vocabulary words, played a pivotal role in the model's superior performance.

Implementing such a system could lead to reduced administrative workloads, decreased clinician burnout, and more time available for meaningful patient interactions, ultimately contributing to improved healthcare outcomes. Furthermore, the research's methodological contributions and insights into domain adaptation have broader implications for other specialized language processing tasks, such as legal documentation or scientific paper summarization.

Beyond the immediate healthcare applications, this work underscores the importance of developing AI systems that are tailored to the unique characteristics and requirements of specific domains. As artificial intelligence continues to permeate various industries, recognizing the limitations of general-purpose models and embracing domain-specific approaches will be crucial for realizing the full potential of these technologies.

Looking ahead, the path to real-world deployment of automated clinical documentation systems will require collaboration between researchers, healthcare professionals, and policymakers. Continuous refinement and adaptation of the models will be necessary to keep pace with evolving medical practices and documentation standards. Additionally, addressing ethical concerns surrounding fairness, privacy, transparency, and human oversight will be paramount to ensure the responsible and trustworthy adoption of these AI-driven solutions.

In conclusion, this research represents a significant stride towards harnessing the power of artificial intelligence to transform clinical documentation processes, ultimately paving the way for a more efficient and effective healthcare system that prioritizes quality patient care. While challenges remain, the findings and contributions of this study provide a solid foundation upon which future advancements can be built, bringing us closer to a reality where the administrative burdens of documentation are alleviated, allowing healthcare professionals to focus on their primary mission of improving lives.

# References

Abacha, A. B. (2023, May). microsoft/clinical\_visit\_note\_summarization\_corpus. GitHub. <https://github.com/microsoft/clinical_visit_note_summarization_corpus>

Abacha, A. B., Yim, W., Fan, Y., & Lin, T. (2023). An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2291–2302. <https://doi.org/10.18653/v1/2023.eacl-main.168>

Aberdeen, J., Bayer, S., Yeniterzi, R., Wellner, B., Clark, C., Hanauer, D., Malin, B., & Hirschman, L. (2010). The MITRE Identification Scrubber Toolkit: Design, training, and assessment. *International Journal of Medical Informatics*, *79*(12), 849–859. <https://doi.org/10.1016/j.ijmedinf.2010.09.007>

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly Available Clinical BERT Embeddings. *ArXiv:1904.03323 [Cs]*. <https://arxiv.org/abs/1904.03323>

Arndt, B. G., Beasley, J. W., Watkinson, M. D., Temte, J. L., Tuan, W.-J., Sinsky, C. A., & Gilchrist, V. J. (2017). Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *The Annals of Family Medicine*, *15*(5), 419–426. <https://doi.org/10.1370/afm.2121>

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate*. ArXiv.org. <https://arxiv.org/abs/1409.0473>

Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. *Advances in Neural Information Processing Systems*, *28*. <https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html>

Bodenreider, O., Cornet, R., & Vreeman, D. (2018). Recent Developments in Clinical Terminologies — SNOMED CT, LOINC, and RxNorm. *Yearbook of Medical Informatics*, *27*(01), 129–139. <https://doi.org/10.1055/s-0038-1667077>

Braaf, S., Manias, E., & Riley, R. (2011). The role of documents and documentation in communication failure across the perioperative pathway. A literature review. *International Journal of Nursing Studies*, *48*(8), 1024–1038. <https://doi.org/10.1016/j.ijnurstu.2011.05.009>

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Hesse, C. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. <https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_medium=email&utm_source=transaction>

Buurman, J., Liu, P., Peters, J. F., Chang, P. J., & Sevenster, M. (2015). Natural Language Processing Techniques for Extracting and Categorizing Finding Measurements in Narrative Radiology Reports. *Applied Clinical Informatics*, *06*(03), 600–610. <https://doi.org/10.4338/aci-2014-11-ra-0110>

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. *11th Conference of the European Chapter of the Association for Computational Linguistics*, 249–256. <https://aclanthology.org/E06-1032.pdf>

Celikyilmaz, A., Clark, E., & Gao, J. (2021). Evaluation of Text Generation: A Survey. *ArXiv:2006.14799 [Cs]*. <https://arxiv.org/abs/2006.14799>

Chien, I., Shi, A., Chan, A. J., & Lindvall, C. (2019). Identification of Serious Illness Conversations in Unstructured Clinical Notes Using Deep Neural Networks. *Lecture Notes in Computer Science*. <https://doi.org/10.1007/978-3-030-12738-1_15>

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches* (pp. 103–111). Association for Computational Linguistics. <https://aclanthology.org/W14-4012.pdf>

Cho, S., Li, C., Yu, D., Foroosh, H., & Liu, F. (2019, October 24). *Multi-Document Summarization with Determinantal Point Processes and Contextualized Representations*. ArXiv.org. <https://doi.org/10.48550/arXiv.1910.11411>

Deep, A. (2020, August 27). *Understanding NLP Keras Tokenizer Class Arguments with example*. Analytics Vidhya. <https://medium.com/analytics-vidhya/understanding-nlp-keras-tokenizer-class-arguments-with-example-551c100f0cbd>

Demner‐Fushman, D., Noémie Elhadad, & Friedman, C. (2021). Natural Language Processing for Health-Related Texts. In E. H. Shortliffe & J. J. Cimino (Eds.), *Biomedical Informatics* (pp. 241–272). Springer Cham. <https://doi.org/10.1007/978-3-030-58721-5_8>

Downing, N. L., Bates, D. W., & Longhurst, C. A. (2018). Physician Burnout in the Electronic Health Record Era: Are We Ignoring the Real Cause? *Annals of Internal Medicine*, *169*(1), 50. <https://doi.org/10.7326/m18-0139>

Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., & Lu, Z. (2019). ML-Net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association: JAMIA*, *26*(11), 1279–1285. <https://doi.org/10.1093/jamia/ocz085>

Espejel, J. L. (2021). Automatic abstractive summarization of long medical texts with multi-encoders Transformer and general-domain summary evaluation with wikiSERA. In *theses.hal.science*. <https://theses.hal.science/tel-03376172/>

Finch, D. K. (2012). *TagLine: Information Extraction for Semi-Structured Text Elements In Medical Progress Notes*. <https://search.proquest.com/openview/e9b8d28f0b5c161e32a06527ca4cd050/1?pq-origsite=gscholar&cbl=18750>

Gardent, C., Shimorina, A., Narayan, S., & Perez-Beltrachini, L. (2017). Creating Training Corpora for NLG Micro-Planning. *55th Annual Meeting of the Association for Computational Linguistics (ACL)*. <https://inria.hal.science/hal-01623744>

Giorgi, J., Toma, A., Xie, R., Chen, S. S., An, K. R., Zheng, G. X., & Wang, B. (2023, June 3). *WangLab at MEDIQA-Chat 2023: Clinical Note Generation from Doctor-Patient Conversations using Large Language Models*. ArXiv.org. <https://doi.org/10.48550/arXiv.2305.02220>

Goncalves, T., Rio-Torto, I., Teixeira, L. F., & Cardoso, J. S. (2022). A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms? *IEEE Access*, *10*, 98909–98935. <https://doi.org/10.1109/access.2022.3206449>

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>

Halcomb, E., Stephens, M., Bryce, J., Foley, E., & Ashley, C. (2016). Nursing competency standards in primary health care: an integrative review. *Journal of Clinical Nursing*, *25*(9-10), 1193–1205. <https://doi.org/10.1111/jocn.13224>

Hinton, G., Srivastava, N., & Swersky, K. (2012). Lecture 6d-a separate, adaptive learning rate for each connection. *Slides of lecture neural networks for machine learning*, *1*, 1-31.

Hirschtick, R. E. (2006). Copy-and-Paste. *JAMA*, *295*(20), 2335. <https://doi.org/10.1001/jama.295.20.2335>

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

Huang, K., Altosaar, J., & Ranganath, R. (2019). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *ArXiv Preprint*. <https://arxiv.org/abs/1904.05342>

Isah, E. E., & Byström, K. (2020). The mediating role of documents: information sharing through medical records in healthcare. *Journal of Documentation*, *76*(6), 1171–1191. <https://doi.org/10.1108/jd-11-2019-0227>

Jeong, C., Jang, S., Park, E., & Choi, S. (2020). A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics*, *124*(3), 1907–1922. <https://doi.org/10.1007/s11192-020-03561-y>

Jin, D., & Szolovits, P. (2018, August 19). *Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts*. ArXiv.org. <https://doi.org/10.48550/arXiv.1808.06161>

Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2022). AMMU: A survey of transformer-based biomedical pretrained language models. *Journal of Biomedical Informatics*, *126*, 103982. <https://doi.org/10.1016/j.jbi.2021.103982>

Khandelwal, R. (2020, February 15). *Sequence 2 Sequence model with Attention Mechanism*. Medium. <https://towardsdatascience.com/sequence-2-sequence-model-with-attention-mechanism-9e9ca2a613a>

Korfiatis, A. P., Moramarco, F., Sarac, R., & Savkov, A. (2022). PriMock57: A Dataset Of Primary Care Mock Consultations. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, *2*, 588–598. <https://doi.org/10.18653/v1/2022.acl-short.65>

Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2019). Evaluating the Factual Consistency of Abstractive Text Summarization. *ArXiv:1910.12840 [Cs]*. <https://arxiv.org/abs/1910.12840>

Kundeti, S. R., Vijayananda, J., Mujjiga, S., & Kalyan, M. (2016). Clinical named entity recognition: Challenges and opportunities. *International Conference on Big Data*. <https://doi.org/10.1109/bigdata.2016.7840814>

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. <https://doi.org/10.1109/5.726791>

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4). <https://doi.org/10.1093/bioinformatics/btz682>

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, No. 8, pp. 707-710).

Lin, C.-Y. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries*. Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=de791d19d5abe0b4a419ff039c07f066f781ec9c>

Liu, Y., & Lapata, M. (2019). Text Summarization with Pretrained Encoders. *ArXiv Preprint*. <https://arxiv.org/abs/1908.08345>

Melnick, E. R., Dyrbye, L. N., Sinsky, C. A., Trockel, M., West, C. P., Nedelec, L., Tutty, M. A., & Shanafelt, T. (2019). The Association Between Perceived Electronic Health Record Usability and Professional Burnout Among US Physicians. *Mayo Clinic Proceedings*, *95*(3). <https://doi.org/10.1016/j.mayocp.2019.09.024>

Moore, W., & Frye, S. (2019). Review of HIPAA, Part 1: History, Protected Health Information, and Privacy and Security Rules. *Journal of Nuclear Medicine Technology*, *47*(4), 269–272. <https://doi.org/10.2967/jnmt.119.227819>

Muneera, M. N., & Sriramya, P. (2023). An Enhanced Optimized Abstractive Text Summarization Traditional Approach Employing Multi-layered Attentional Stacked LSTM with the Attention RNN. *Lecture Notes in Electrical Engineering*, 303–318. <https://doi.org/10.1007/978-981-19-7169-3_28>

Osuala, R., Kaisar Kushibar, Garrucho, L., Linardos, A., Zuzanna Szafranowska, Klein, S., Glocker, B., Diaz, O., & Karim Lekadir. (2023). Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging. *Medical Image Analysis*, *84*, 102704–102704. <https://doi.org/10.1016/j.media.2022.102704>

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ’02*. <https://doi.org/10.3115/1073083.1073135>

Park, J., Dimitrios Kotzias, Kuo, P., Logan, R. L., Merced, K., Singh, S., Tanana, M., Efi Karra Taniskidou, Jennifer Elston Lafata, Atkins, D. C., Tai-Seale, M., Imel, Z. E., & Smyth, P. (2019). Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. *Journal of the American Medical Informatics Association*, *26*(12), 1493–1504. <https://doi.org/10.1093/jamia/ocz140>

Peng, Y., Yan, S., & Lu, Z. (2019). *Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets*. <https://doi.org/10.48550/arxiv.1906.05474>

Pine, K. H., & Bossen, C. (2020). Good organizational reasons for better medical records: The data work of clinical documentation integrity specialists. *Big Data & Society*, *7*(2), 205395172096561. <https://doi.org/10.1177/2053951720965616>

Pougué Biyong, J., Wang, B., Lyons, T., & Nevado-Holgado, A. (2020, November 1). *Information Extraction from Swedish Medical Prescriptions with Sig-Transformer Encoder*. ACLWeb; Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.clinicalnlp-1.5>

Prechelt, L. (1998). Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, *11*(4), 761–767. <https://doi.org/10.1016/s0893-6080(98)00010-0>

Quiroz, J. C., Laranjo, L., Kocaballi, A. B., Berkovsky, S., Rezazadegan, D., & Coiera, E. (2019). Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ Digital Medicine*, *2*(1), 1–6. <https://doi.org/10.1038/s41746-019-0190-1>

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning*, 8748–8763. PMLR. <http://proceedings.mlr.press/v139/radford21a>

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*(140), 1–67. <https://www.jmlr.org/papers/v21/20-074.html>

Rajkomar, A., Kannan, A., Chen, K., Vardoulakis, L., Chou, K., Cui, C., & Dean, J. (2019). Automatically Charting Symptoms From Patient-Physician Conversations Using Machine Learning. *JAMA Internal Medicine*, *179*(6), 836. <https://doi.org/10.1001/jamainternmed.2018.8558>

Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2016, May 6). *Sequence Level Training with Recurrent Neural Networks*. ArXiv Preprint ArXiv:1511.06732. <https://doi.org/10.48550/arXiv.1511.06732>

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, *4*(1). <https://doi.org/10.1038/s41746-021-00455-y>

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP models with CheckList. *ArXiv:2005.04118 [Cs]*. <https://arxiv.org/abs/2005.04118>

Ruch, P., Gobeill, J., Lovis, C., & Geissbühler, A. (2008). Automatic medical encoding with SNOMED categories. *BMC Medical Informatics and Decision Making*, *8*(S1). <https://doi.org/10.1186/1472-6947-8-s1-s6>

Schloss, B., & Konam, S. (2020). Towards an Automated SOAP Note: Classifying Utterances from Medical Conversations. *Machine Learning for Healthcare Conference*, 610–631. PMLR. <https://proceedings.mlr.press/v126/schloss20a.html>

See, A., Liu, P. J., & Manning, C. D. (2017). Get To The Point: Summarization with Pointer-Generator Networks. *ArXiv Preprint*. <https://arxiv.org/abs/1704.04368>

Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Medical Informatics*, *7*(2), e12239. <https://doi.org/10.2196/12239>

Soltau, H., Wang, M., Shafran, I., & Shafey, L. E. (2021, April 5). *Understanding Medical Conversations: Rich Transcription, Confidence Scores & Information Extraction*. ArXiv.org. <https://doi.org/10.48550/arXiv.2104.02219>

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958. <https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer>

Stubbs, A., Kotfila, C., & Uzuner, Ö. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of Biomedical Informatics*, *58 Suppl*(Suppl), S11-9. <https://doi.org/10.1016/j.jbi.2015.06.007>

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, *27*. <https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html>

Tan, J., Wan, X., & Xiao, J. (2017). Abstractive Document Summarization with a Graph-Based Attentional Neural Model. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. <https://doi.org/10.18653/v1/p17-1108>

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. *University of Toronto, Technical Report*, *6*.

Ting, J., Garnett, A., & Donelle, L. (2021). Nursing education and training on electronic health record systems: An integrative review. *Nurse Education in Practice*, *55*, 103168. <https://doi.org/10.1016/j.nepr.2021.103168>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is All you Need*. Neural Information Processing Systems; Curran Associates, Inc. <https://proceedings.neurips.cc/paper/7181-attention-is-all>

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., & Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, *87*, 12–20. <https://doi.org/10.1016/j.jbi.2018.09.008>

Williams, R. J., & Zipser, D. (1989). Experimental Analysis of the Real-time Recurrent Learning Algorithm. *Connection Science*, *1*(1), 87–111. <https://doi.org/10.1080/09540098908915631>

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2020). HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *ArXiv:1910.03771 [Cs]*. <https://arxiv.org/abs/1910.03771>

Xie, Q., & Wang, F. (2023). Faithful AI in Healthcare and Medicine. *MedRxiv*, *3*. <https://doi.org/10.1101/2023.04.18.23288752>

Xu, J., & Durrett, G. (2019, September 9). *Neural Extractive Text Summarization with Syntactic Compression*. ArXiv.org. <https://doi.org/10.48550/arXiv.1902.00863>

Yang, X., Lyu, T., Li, Q., Lee, C.-Y., Bian, J., Hogan, W. R., & Wu, Y. (2019). A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, *19*(S5). <https://doi.org/10.1186/s12911-019-0935-4>

Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A. R., Li, I., Friedman, D., & Radev, D. R. (2019). ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 7386–7393. <https://doi.org/10.1609/aaai.v33i01.33017386>

Yeniterzi, R., Aberdeen, J. S., Bayer, S., Wellner, B., Hirschman, L., & Malin, B. (2010). Effects of personal identifier resynthesis on clinical text de-identification. *Journal of the American Medical Informatics Association*, *17*(2), 159–168. <https://doi.org/10.1136/jamia.2009.002212>

Yim, W., Fu, Y., Ben Abacha, A., Snider, N., Lin, T., & Yetisgen, M. (2023). Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. *Scientific Data*, *10*(1), 586. <https://doi.org/10.1038/s41597-023-02487-3>

Yue, X., Gutierrez, B. J., & Sun, H. (2020, May 1). *Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset*. ArXiv.org. <https://doi.org/10.48550/arXiv.2005.00574>

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November 21). *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*. Proceedings.mlr.press; PMLR. <http://proceedings.mlr.press/v119/zhang20ae>

1. Please note the page numbers where evidence of meeting the learning outcome can be found in your dissertation. [↑](#footnote-ref-1)