2022

Semantics-driven Abstractive Document Summarization

Amanuel Alambo
Wright State University - Main Campus

SEMANTICS-DRIVEN ABSTRACTIVE DOCUMENT SUMMARIZATION

A Dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

by

AMANUEL ALAMBO
M.S., Wright State University, 2020
B.S., Jimma University, Ethiopia, 2007

2022
Wright State University

Wright State University
GRADUATE SCHOOL

May 23, 2022

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Amanuel Alambo ENTITLED Semantics-driven Abstractive Document Summarization BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Tanvi Banerjee, Ph.D.
Dissertation Co-Director

Krishnaprasad Thirunarayan, Ph.D.
Dissertation Co-Director

Thomas Wischgoll, Ph.D.
Director, Computer Science and Engineering Ph.D Program

Barry Milligan, Ph.D.
Dean of the Graduate School

Committee on
Final Examination

Tanvi Banerjee, Ph.D.

Krishnaprasad Thirunarayan, Ph.D.

Michael Raymer, Ph.D.

Vijayan Asari, Ph.D.

ABSTRACT

Alambo, Amanuel. Ph.D., Department of Computer Science and Engineering, Wright State University, 2022. *Semantics-driven Abstractive Document Summarization.*

The evolution of the Web over the last three decades has led to a deluge of scientific and news articles on the Internet. Harnessing these publications in different fields of study is critical to effective end user information consumption. Similarly, in the domain of healthcare, one of the key challenges with the adoption of Electronic Health Records (EHRs) for clinical practice has been the tremendous amount of clinical notes generated that can be summarized without which clinical decision making and communication will be inefficient and costly. In spite of the rapid advances in information retrieval and deep learning techniques towards abstractive document summarization, the results of these efforts continue to resemble extractive summaries, achieving promising results predominantly on lexical metrics but performing poorly on semantic metrics. Thus, abstractive summarization that is driven by intrinsic and extrinsic semantics of documents is not adequately explored. Resources that can be used for generating semantics-driven abstractive summaries include:

- Abstracts of multiple scientific articles published in a given technical field of study to generate an abstractive summary for topically-related abstracts within the field, thus reducing the load of having to read semantically duplicate abstracts on a given topic.
- Citation contexts from different authoritative papers citing a reference paper can be used to generate utility-oriented abstractive summary for a scientific article.
- Biomedical articles and the named entities characterizing the biomedical articles along with background knowledge bases to generate entity and fact-aware abstractive summaries.
- Clinical notes of patients and clinical knowledge bases for abstractive clinical text summarization using knowledge-driven multi-objective optimization.

In this dissertation, we develop semantics-driven abstractive models based on intra-document and inter-document semantic analyses along with facts of named entities retrieved from domain-specific knowledge bases to produce summaries. Concretely, we propose a sequence of frameworks leveraging semantics at various granularity (e.g., word, sentence, document, topic, citations, and named entities) levels, by utilizing external resources. The proposed frameworks have been applied to a range of tasks including:

1. Abstractive summarization of topic-centric multi-document scientific articles and news articles.

2. Abstractive summarization of scientific articles using crowd-sourced citation contexts.

3. Abstractive summarization of biomedical articles clustered based on entity-relatedness.

4. Abstractive summarization of clinical notes of patients with heart failure and Chest X-Rays recordings.

The proposed approaches achieve impressive performance in terms of preserving semantics in abstractive summarization while paraphrasing. For summarization of topic-centric multiple scientific/news articles, we propose a three-stage approach where abstracts of scientific articles or news articles are clustered based on their topical similarity determined from topics generated using Latent Dirichlet Allocation (LDA), followed by extractive phase and abstractive phase. Then, in the next stage, we focus on abstractive summarization of biomedical literature where we leverage named entities in biomedical articles to 1) cluster related articles; and 2) leverage the named entities towards guiding abstractive summarization. Finally, in the last stage, we turn to external resources such as citation contexts pointing to a scientific article to generate a comprehensive and utility-centric abstractive summary of a scientific article, domain-specific knowledge bases to fill gaps in information about entities in a biomedical article to summarize and clinical notes to guide

abstractive summarization of clinical text. Thus, the bottom-up progression of exploring semantics towards abstractive summarization in this dissertation starts with (*i*) Semantic Analysis of Latent Topics; builds on (*ii*) Internal and External Knowledge-I (gleaned from abstracts and Citation Contexts); and extends it to make it comprehensive using (*iii*) Internal and External Knowledge-II (Named Entities and Knowledge Bases).

Contents

List of Figures

List of Tables

Acknowledgment

Praise the Lord Almighty, I would have never thought I would come this far. God is good. I would like to express my deepest gratitude to my advisers Professor Tanvi Banerjee and Professor Krishnaprasad Thirunarayan. Your precious time, support, encouragement, inspiration, and advice kept me going; thank you so much for being fully invested in my work and the constant enthusiasm you have shown through my PhD.

Next, I am immensely grateful to my PhD committee members Professor Michael Raymer, and Professor Vijayan Asari for guiding me through my journey, for always being my inspiration and for crafting me into a researcher. Your insights and wisdom played a key role as I advanced in my research.

I am so thankful to the faculty in the Department of Computer Science and Engineering at Wright State University. I would like to particularly say thank you very much to Professor Mateen Rizki, Professor Amit Sheth, Professor William Romine, and Professor Yong Pei. Your words of wisdom and advice guided me through my journey.

My gratitude extends to the administration of the Department of Computer Science and Engineering at Wright State University; specifically, I would like to say thank you so much to Jennifer Limoli, Wendy Chetcuti, and Tonya Davis; from the first day I started my PhD, your support and cooperation have been profound; this would have never been possible without your guidance and cooperation. I am also thankful to Wright State University Administration - the University Center for International Education (UCIE), Graduate School, Bursar Office, and Raider Connect.

I am so grateful to the directors, supervisors, and colleagues I had during my research internship with the Search and Core Ranking team at eBay Inc. I learned the crucial lesson of conducting an industry-scale research through your mentorship and guidance.

Thanks a lot to the funding agencies for their generous support towards my research - the United States Air Force (USAF), the National Institutes of Health (NIH), and the National Science Foundation (NSF).

My gratitude goes to my student mentors and collaborators Dr. Manas Gaur, Dr. Ugur Kursuncu, Daniel Foose, Swati Padhee, Dr. Riddhi Doshi, Usha Lokala, Alan Smith, Joy Prakash Sain, Andrew Young, Farahnaz Golroo, Ankita Agarwal, and Mike Partin. It has always been an absolute delight working with you.

Thank you so much to my father, Fanta Alambo, who showered me with his abundant love and for being a role model of fatherhood and a good husband. To my brother, Dr. Habtamu Alambo, for showing me true brotherhood and for being around whenever I needed. To my cousins Tsegaye Zewdu, Afework Yilma, Eskinder Yilma, Aberash Yilma, and Bizuayehu Anbessie.

To my amazing friends Addisu Guddissa, Worede Gebremariam, Feysel Ahmed, Ayalew Tigro, Yohannes Bogale, Bisrat Tsegaye, Delk and family, Jay Finley, John Fleming, Nicely and family, Akshay Hira, Mamay Yohannes, Reason Alemayehu, Solomon Bekele, Henok Fisseha, and Daniel Kebede. Thank you so much for your unwavering friendship and support. Thanks a lot to International Friendships (IFI) for your hospitality and friendship. I am thankful to the Habesha Christian community in Dayton, Ohio, for being a family.

Dedicated to

My departed mother Woletebirhan Anbessie (Etete), the epitome of love, faith and strength!

Introduction

"Perhaps the best test of a man's intelligence is his capacity for making a summary."

—Lytton Strachey, 1880 – 1932

1.1 Motivation

The birth of the World Wide Web [3] has enabled massive information generation and exchange among users. Within this spectrum of information, text is the most widely shared form of information on the Web with varying types, size, and diversity. Scientific publications and news articles constitute a major proportion of textual data on the Web. To overcome the consequent information overload and redundancy in different fields of study, summarization of documents about a topic has become critically important for the users to keep abreast of the changing landscape.

Based on the number of documents to summarize, document summarization can be classified into single document summarization and multi-document summarization. Similarly, based on the approach used, summarization can also be classified as extractive or abstractive. While extractive summarization is mainly focused on extracting salient sentences from source document(s) verbatim, abstractive summarization deals with abstracting and paraphrasing the sentences in source documents while preserving semantics. Extractive summarization is relatively well studied, particularly in the news domain. However, abstractive document summarization has untapped potential because we are yet to explore

semantically aligned abstraction techniques.

At the level of leveraging intrinsic semantics, clustering of scientific articles based on semantic analysis of the topics in the articles and generating a topic-aware abstractive summary for a cluster of articles is key to acquiring a summary of a set of documents on a given topic. Further, another source of intrinsic semantics that can be used for abstractive summarization constitutes named entities in a scientific article which give a high level view of the "aboutness" [4, 5] of the article which can be leveraged for the task of abstractive summary generation. Next, extrinsic sources of semantics such as citation contexts pointing to a given scientific article, and facts curated by domain experts and stored in domain-specific knowledge bases can be used to enhance abstractive summaries.

Named entity recognition has advanced natural language processing and text understanding, particularly in the biomedical domain [6, 7, 8, 9]. However, the use of named entities and their semantics (i.e., information about related entities and relationships) for abstractive summarization of documents has not been adequately explored. While there have been recent efforts to leverage entity information for single document abstractive summarization, these efforts are focused towards single document summarization of news articles (e.g., NYT and CNN/Daily Mail corpora) and do not leverage entity semantics. Multi-document abstractive summarization of biomedical literature driven by entity information, while not studied yet, is crucial for the practice of a biomedical researcher.

In addition to producing an abstractive summary of a document or a set of documents that are composed by the original authors driven by topical or entity information, abstractive summary generation using the views of authoritative sources is critical in the scientific domain. To this end, citation sentences in scientometrics from multiple citing papers (authors) have led to the applications of extractive summarization to produce a community-based summary of a reference paper from citing papers. Citation-based summarization is focused on producing a comprehensive and concise summary of a reference paper from the perspectives of others. While the abstract of a scientific article provides the author(s)'

perspectives, abstractive summaries synthesized from citation sentences citing a reference paper provide a utility-driven highlight and terse perspective of the reference paper in view of authoritative sources. A series of CL-SciSumm tasks in the past decade led to various approaches to generating a summary of a reference paper from citation sentences. However, these approaches have been primarily extractive and hinge on the notion of identifying target text spans in a reference paper corresponding to citing sentences and fusing the target text spans to generate an extractive summary. Thus, abstractive summarization of a scientific article by leveraging the metadata in citation networks and the semantics of citation sentences has not been studied.

Although recent progress in natural language processing and deep learning has enabled novel approaches for abstractive document summarization in the contexts discussed above, these approaches are limited to showing improvements in lexical metrics, and evaluations against semantic metrics have not been investigated.

In this dissertation, we set out four aims and develop a series of frameworks to conduct abstractive document summarization at different levels of semantics for various tasks and domains. Particularly, we apply and evaluate our proposed frameworks on four types of datasets: 1) 20 fields of study from the Microsoft Academic Graph (MAG) [10]; 2) News articles from the Document Understanding Conference DUC-2004 task-2 [11]; 3) Medline abstracts we curate from PubMed, and a benchmark dataset on biomedical literature summarization; and 4) clinical notes of patients with heart failure and benchmark datasets on Chest X-Ray radiology reports. Figure 1.1 and Figure 1.2 show the rapidly increasing trend of scientific publications over the last 27 years and the adoption of Electronic Health Records (EHRs) from 2004 - 2017, necessitating the need for automatic summarization. Figure 1.3 illustrates the types of semantics, the domains, and the tasks used for the four research aims in this dissertation characterizing our four-pronged approach to semantics-driven abstractive summarization.

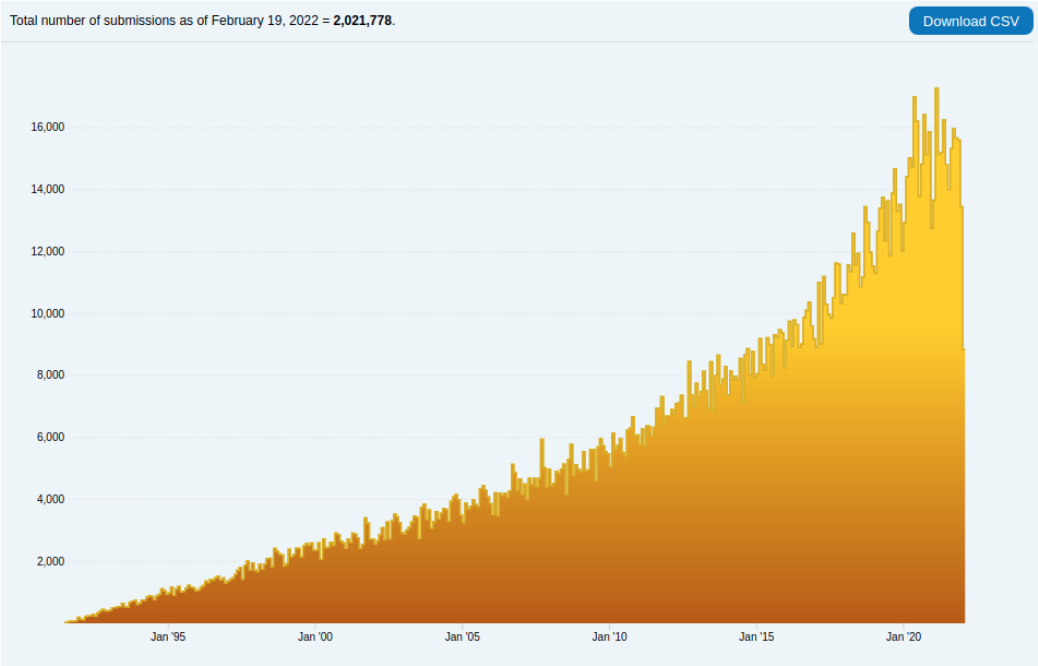

Figure 1.1: Monthly data of scientific articles submitted to arXiv from Jan 1995 till Jan 2022.
Source: https://arxiv.org/stats/monthly_submissions

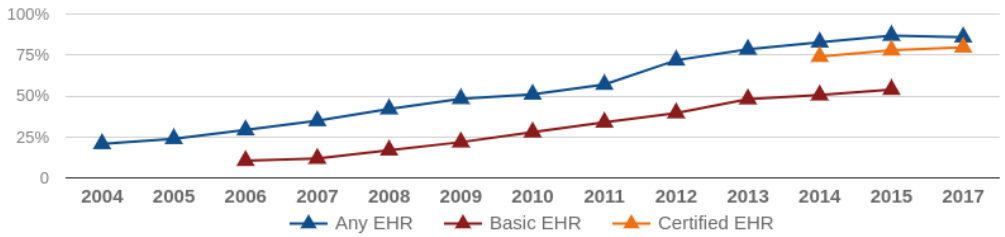

Figure 1.2: Rate of adoption of EHR from 2004 - 2017.
Source: https://bit.ly/3scspHs

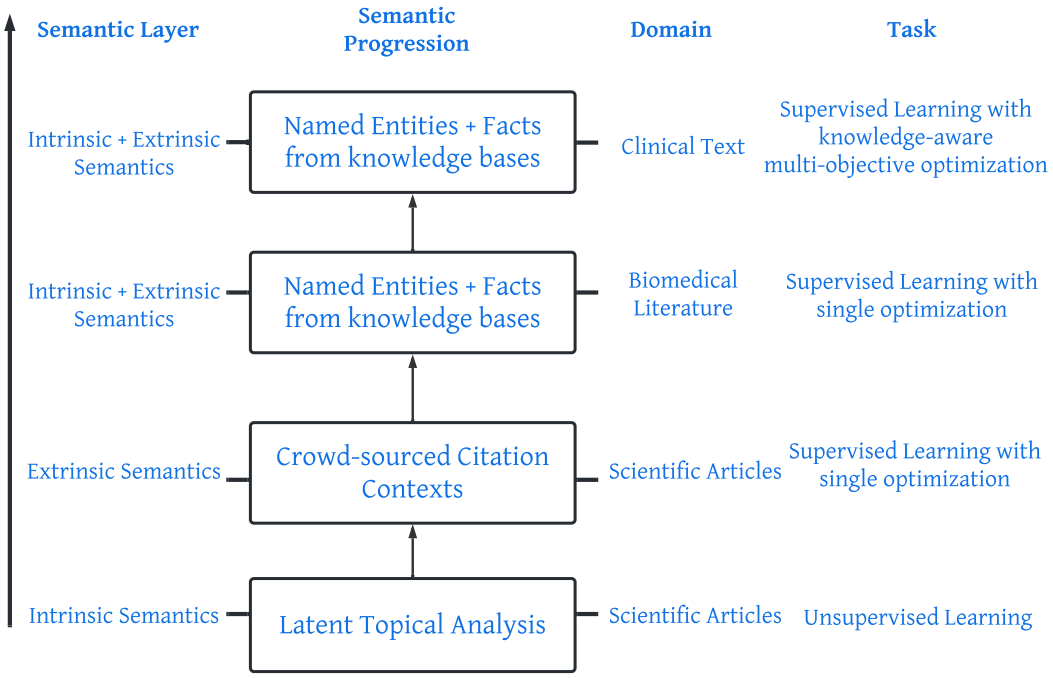

Figure 1.3: Exploring semantics at various levels for different domains and tasks.

Figure 1.4 illustrates the historical evolution of the task of text summarization in general along with the most seminal works and situate our dissertation with respect to what has been achieved and what gaps we have attempted to fill in.

The following are the research aims of this dissertation.

1.2 Research Aim I

We propose to build a topic-centric unsupervised multi-document abstractive summarization framework for scientific and news articles. The framework comprises a topical clustering module for clustering topically related documents followed by a hybrid model that consists of an extractive and an abstractive phase where salient language unit selection is performed using the extractive phase, and the content abstraction and generation are performed by the abstractive phase.

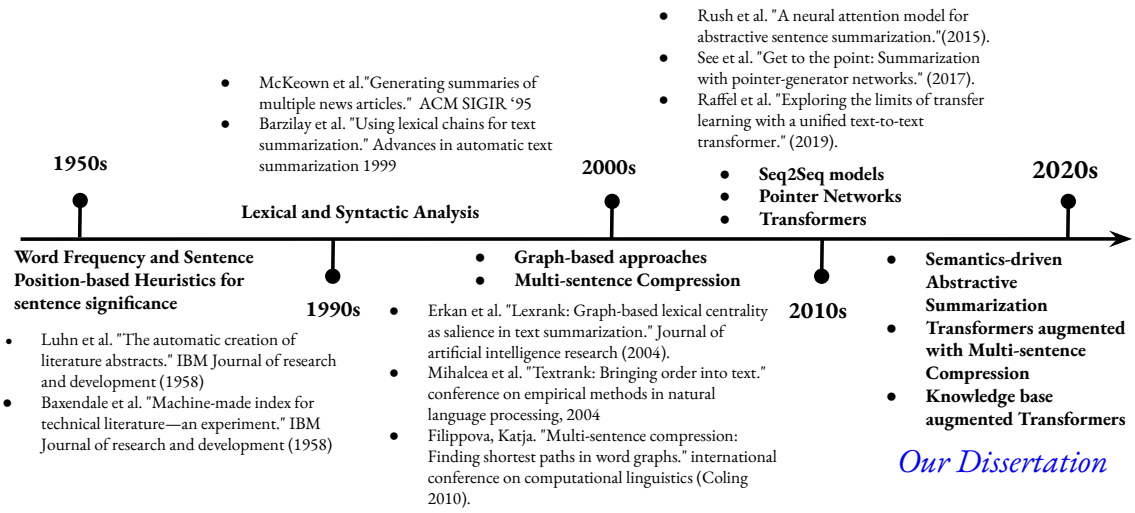

Figure 1.4: Historical Evolution of Text Summarization and the direction of our dissertation.

1.3 Research Aim II

Next, we propose to build a citation-driven abstractive summarization model to generate a summary of a scientific article using citation sentences from citing papers. These summaries are called community-based abstractive summaries where authoritative citing sources are used to compose a summary of a target scientific article based on the views of other researchers.

1.4 Research Aim III

We utilize named entities in biomedical literature and facts retrieved from domain-specific knowledge bases to build a multi-document abstractive summarization framework. The documents are clustered based on entity-relatedness and content selection is determined based on entity-informativeness to provide a comprehensive and coherent summary using background knowledge.

1.5 Research Aim IV

We build an abstractive summarization framework for clinical text using Multi-Objective Optimization by jointly optimizing cost functions to minimize generative loss with respect to ground truth summaries, named entity chain in ground truth summaries, and facts in ground truth summaries.

1.6 Thesis Statement

Semantics-driven abstractive summarization of scientific articles, news articles, biomedical literature, and clinical text is crucial for efficient information consumption and effective decision making. This can be attained through: i) Semantic analysis of topics in articles followed by topic-centric multi-document abstractive summarization; ii) Leveraging crowd-sourced knowledge in the form of citation contexts for abstractive summarization of a scientific article; iii) Entity-driven fact-aware abstractive summarization of biomedical literature; and iv) Abstractive summarization of clinical text using knowledge-aware multi-objective optimization.

1.7 Structure of the Dissertation

This dissertation is structured into six chapters.

Chapter 2 introduces a brief historical context on document summarization and the various techniques proposed by researchers over the last three decades. Then, it goes onto discuss related works corresponding to each of the research aims.

Chapters 3, 4, 5, and 6 present the motivation, problem statement and the proposed approach for Research Aims I, II, III, and IV.

Chapter 7 wraps up the dissertation with conclusion, summary of contributions, future

directions and major insights.

Table 1.1 outlines the Research Questions we address under each Research Aim.

Research Aim	Chapter	Research Questions
Topic-Centric Unsupervised Multi-document Summarization	3	**RQ1**: Can latent topics in document corpus be automatically discovered and be used to 1) cluster articles based on topical relatedness; and 2) guide abstractive summarization? **RQ2**: How can large language models such as GPT-2 be used to improve the task of abstractive summarization? **RQ3**: How does the proposed framework whose building blocks are GPT-2 and a novel Multi-Sentence Compression algorithm perform with respect to human summary evaluation metrics?
TransFuse for Abstractive Summarization of Scientific Article from Citation Contexts	4	**RQ1**: Can multiple citation contexts citing a given scientific article be used along with the content of the article to generate a hybrid (i.e., integrating citation contexts and the content of the article) summary for the article? **RQ2**: How can we amalgamate transformer-based models and sentence fusion technique to improve abstractive summarization and address neural text degeneration (a phenomenon where generated text is repetitive and non-sensical)? **RQ3**: Will the amalgamation of a transformer-based model and sentence fusion lead to better abstractive summaries in terms of semantic equivalence and paraphrasing?
Entity-driven Fact Aware Abstractive Summarization of Biomedical Literature	5	**RQ1**: How can named entities be leveraged for abstractive summarization of biomedical literature and address entity hallucination? **RQ2**: Can we use named entities to mine facts from biomedical knowledge bases and use these facts along with the named entities to guide abstractive summarization? **RQ3**: How does an abstractive summarization framework augmented with knowledge bases perform with respect to entity-level factual accuracy and semantic equivalence?
Improving the Factual Accuracy of Abstractive Clinical Text Summarization using Multi-Objective Optimization	6	**RQ1**: Can we model the clinical practice of writing an impression from a set of findings as the task of abstractive summary generation? **RQ2**: How can we use named entities and facts retrieved from domain-specific knowledge bases to define a multi-objective optimization cost function and train end-to-end abstractive summarization models? **RQ3**: Will a multi-objective optimization based abstractive summarization model perform better than single optimization based model on factual accuracy metrics?

Literature Review

We discuss the related research efforts in five segments: *i)* a historical overview of research in document summarization and its progress over the last 70 years; *ii)* document summarization in the context of summarizing multiple topically related scientific and news articles which Research Aim I is focused on; *iii)* summarization of a scientific article using crowd-sourced citation contexts which Research Aim II attempts to address; *iv)* the role of named entities in summarization of scientific and news articles which is the focus of Research Aim III; and *v)* research in summarization of clinical text which is the objective of Research Aim IV.

2.1 A brief historical context

Early work on document summarization as applied to scientific literature dates back to the 1950's [12, 13]. [12] proposed automatic abstract (auto-abstract) generation by scanning a machine-readable scientific article for "significant" sentences based on sentence "significance" criteria. [13] introduced the notion of determining sentence saliency based on sentence position in a document. While there have been few efforts [14, 5, 15, 16] towards automatic summarization of documents in the decades following the 1950s, recent advances in document summarization are primarily or partly influenced by pioneering research in the 1990s.

Document summarization in the 1990s mainly focused on lexical and syntactic anal-

yses of documents [17, 18, 19, 20, 21, 4, 22]. The resurgence of document summarization in the 1990s was mainly influenced by the invention of the Web and progress in the field of information retrieval. While efforts in the 1990s saw a dramatic progress in document summarization, they were mainly extractive.

The invention of Pagerank in the late 1990s [23] for web search engines brought graph-based insights to document summarization. Graph-based approaches to document summarization came to the spotlight in the 2000s with the introduction of LexRank [24], and TextRank [[25] algorithms which proved successful at extractive summarization. Techniques to fuse multiple sentences into one sentential unit were later proposed with sentence fusion [26] and multi-sentence compression algorithms [27] paving the way for efforts to generate paraphrased sentences as opposed to extracting sentences from source document(s). The invention of multi-sentence compression (MSC) and word graph based approaches left a longstanding impact on abstractive summarization including the works of [28, 29, 30].

With the advent of sequence to sequence deep learning models with or without attention [31, 32], new techniques to abstractive summarization emerged in the 2010s [33, 34, 35]. While sequence to sequence models achieved significant improvements in abstractive summarization, their application to long documents was limited as seq2seq models suffer from long range dependency issues. The introduction of transformer models [36] helped address the challenge of summarizing long documents while simultaneously capturing deeper semantics of documents at different levels of granularity (words, entities, sentences, and paragraphs).

The sections below discuss the related works appropriate to each of the four aims proposed in this dissertation: topic-centric unsupervised multi-document abstractive summarization, citation-driven abstractive single document summarization, entity-driven fact-aware abstractive summarization, and abstractive summarization using knowledge-aware multi-objective optimization.

2.2 Research Aim I

This section discusses recent works related to Research Aim I: Topic-centric unsupervised abstractive multi-document summarization of scientfic articles and news articles.

Broadly, approaches used for unsupervised abstractive document summarization can be organized based on the techniques employed including sequence to sequence models [37, 38], neural attentive models [39, 40, 41, 42, 35, 43, 44], Abstract Meaning Representation (AMR) [45, 46, 47], and centroid-based summarization [30, 48].

Recent advances in deep learning [36, 49, 50] enabled abstractive summarization of multiple documents. [39] propose MeanSum which consists of two components: 1) an autoencoder, which learns representations (for Yelp and Amazon reviews) followed by 2) a summarization module that produces abstractive summaries. The autoencoder and summarization module components are based on LSTM encoder and decoder networks and the summary generation is achieved via straight-through-gumbel-softmax trick. [48] introduce ILPSumm which is based on Integer Linear Programming (ILP) and includes modules for 1) Identification of informative content across documents using LexRank; 2) Clustering similar sentences from the documents; and 3) Generating informative and linguistically grounded sentences from different clusters using word-graph. [30] improved upon the multi-document summarization approach proposed in [48] by introducing a paraphrastic fusion model, which they call ParaFuse, based on context-aware paraphrasing of words using PPDB 2.0 database [51] and a deep representation learning of sentences using Gated Recurrent Units (GRUs). The main improvement of ParaFuse over ILPSumm is in paraphrasing of words in source documents using lexical substitution. While lexical substitution enables the generation of novel words, it is limited when it comes to capturing the context in the source document. [35] propose a Pointer Generator Network which they evaluated on the CNN/Dailymail dataset. While the pointer does the task of preserving the information in source articles, the generator performs the task of generating novel words that do not appear in the source documents. [43] extended the pointer generator

network proposed in [35] by introducing a Hierarchical MMR-Attention Pointer-generator (Hi-MAP) model for multi-document neural abstractive summarization. They used a Bidirectional Long Short Term Memory (Bi-LSTM) network for sentence level encoding. In addition to their Hi-MAP model, they also introduced Multi-News dataset, which consists of 56k articles-summary pairs.

Further, there is a good deal of research in topic-oriented abstractive summarization. [44] propose a neural encoder-decoder framework that takes an article and a topic of interest and generates a summary specific to the topic. They conducted the summary generation for all the topics a document discusses by training their neural network in such a way that it gives more weight (attention) to parts of the input text that are deemed to belong to the topic in question. They create a synthetic topic-centric training corpus where each document is associated with a set of topics. They use the dataset of news articles tagged with topics like politics, sports, and education released at the 2017 KDD Data Science + Journalism Workshop. Their work, however, is supervised and thus relies on availability of human generated training corpus to train their model, unlike our approach which is unsupervised.

The first aim of this dissertation builds upon the techniques on centroid-based summarization [30, 48] by employing language unit identification using coreference dependencies among sentences in an article and a novel bidirectional encoder and autoregressive text generation model [52].

2.3 Research Aim II

This section discusses the related works on Research Aim II: *TransFuse* for Abstractive Summarization of Scientific Articles using Citation Contexts.

While citations harbor valuable insights from authoritative sources, the use of citations for the task of abstractive scientific article summarization is less explored. [53] address the three sub-tasks of CL-SciSumm 2017 using structural correspondence learning (SCL),

positional language modeling (PLM), and textual entailment (TE) for target span identification of a reference paper for a citance [54]. Having identified the target spans, [53] ran LambdaRank for ranking spans in an article, followed by picking the top three spans, and sorting them to appear in the document to produce a summary. [55] conducted identification of cited text spans in a reference paper corresponding to a citance in a citing paper using a pre-trained BERT language model and then generated extractive summaries by combining the identified cited text spans. They experimented with two training configurations for their summarization model on the CL-SciSumm 2019 shared task and Sci-SummNet. In their first configuration, they used full paper sentences as input, and in their second configuration, they used a combination of the abstract and cited text spans. They observed that using cited text spans along with the abstract yields better ROUGE-2 scores.

[56] propose an approach to improve the cohesion and readability of citation-based summaries using a sequence of three steps: *the preprocessing step* that rules out noisy and irrelevant fragments of sentences, *the extraction step* that selects citation sentences based on coverage, and *the postprocessing step* that produces summaries by maximizing readability. They built an SVM binary sentence classifier to classify sentences as *suitable* or *unsuitable* in the preprocessing stage. [57] propose an approach to address inconsistency in citation-based summaries by leveraging citation contexts (cited text spans) in a reference paper and document discourse structure of citations in a citing paper on the TAC2014 dataset. They extract citation contexts from the reference paper for each citation in a citing paper using n-gram based vector space model similarity measure and rank sentences in the citation contexts by maximizing novelty and informativeness to produce final summaries. They report an improvement over baseline approaches by 30% in ROUGE scores. In this dissertation, we use *citation contexts* to refer to the span of sentences including citances and context sentences in a citing paper unlike [57]. The reason for our use of *citation contexts* in this sense is because it directly and naturally refers to what is summarized by the citing author in a citation, and hence *citation contexts*.

[58] propose two approaches based on Graph Convolutional Networks to generate a hybrid summary that integrates a reference paper's abstract and the other researchers' viewpoints of the reference paper. Just as in [57], [58] employ a technique to identify cited text spans given a citation sentence. Their first approach combines the cited text spans and the reference paper's abstract to generate a summary while their second approach augments the reference paper's abstract, which the model takes as a clean summary, with the salient cited text spans (i.e., the community's views not covered in the abstract). They consider citation counts of the reference paper and citing paper as an additional feature to better reflect the authority of each work. They have released 1000 manually annotated scientific documents along with their summaries which is useful for building supervised methods. They report better ROUGE scores over existing work on the CL-SciSumm-2016 shared task. Their work, however, is still extractive and entirely based on data-driven neural models.

[59] propose a dataset of over 2M cited papers from arXiv and over 29M citation contexts from the Microsoft Academic Graph (MAG) spread over 1M citing papers. While most datasets for citation-based summarization are limited in size (less than 100), the dataset provided by [59] can be used to train supervised machine learning models for citation-based summarization. [60] propose an unsupervised technique that leverages word embeddings and domain knowledge (specifically, MeSH and Protein ontologies) to improve identification of cited text spans given a citance in a citing paper. They enrich the citation texts using the cited text spans and the domain knowledge for the purpose of contextualizing the citation texts. For injecting domain knowledge in the embedding-based representation of words, they experiment with retrofitting [61] (a technique where word representations are enhanced using lexicons) and language model interpolation. They conduct two sets of experimental evaluations: 1) they evaluate the relevance of the extracted cited text spans using intrinsic measures; and 2) they evaluate the impact of citation contexts on citation-based summarization using external evaluation. They outperform strong

baselines on the TAC 2014 benchmark dataset in intrinsic and external evaluation metrics.

Although all the aforementioned approaches propose different techniques for citation-based summarization, they are 1) extractive, where sentences are lexically copied into final summary without their semantic understanding or without performing abstraction; 2) do not systematically capture the sentences surrounding a citance in a citing paper to scope the boundary of a citation context, which is critical to understand the semantics of a citance where similar motivation is used in the inverse cloze task (ICT) [62, 63]; and 3) they do not leverage the topics of a reference paper's abstract and introduction or the title of the reference paper in their experiments, while topics or titles of a scientific article offer a high-level overview of the article that can be used to improve abstraction of summaries. Our approach differs in that we propose abstraction of a group of related citation contexts using a coupling of text generation, clustering and a variation of multi-sentence compression [27] algorithm. Further, we use the reference paper topic to identify the span of a citation context in a citing paper. This enables to capture a citation context that is on-topic with respect to the reference paper, generate more relevant words, and preserve the semantics of the citation contexts, eventually leading to more abstractive community-based summaries. Further, we conduct bi-directional evaluation of summaries against 1) the abstract of the reference paper; and 2) the group of topic-aware citation contexts our citation-based summaries are generated from.

The framework proposed in the second aim of this dissertation is evaluated against the following four state of the art baseline abstractive summarization models.

- *Text-to-Text Transfer Transformer (T5)* [64] is a unified framework that casts every natural language processing problem as a text-to-text problem. The framework is pre-trained with the "Colossal Clean Crawled Corpus" and is tested on downstream tasks including machine translation, question answering, and abstractive summarization. The unified framework follows the same training procedures (e.g., teacher forcing [65] hyperparameters, loss functions (e.g., denoising objectives during pre-training)

and decoding strategies (e.g., greedy decoding) for each of the NLP tasks.

- *BART* [66] is a denoising autoencoder that consists of a bidirectional encoder to encode a document and a left-to-right autoregressive decoder (GPT) to generate an abstractive summary. The pre-training has two stages: 1) corruption of text with an arbitrary noising function; and 2) a sequence-to-sequence model trained to reconstruct the original document.

- *Pegasus* [67] is a transformer-based encoder-decoder model that proposes a new pre-training objective to mask a certain number of tokens and important sentences in an input document and learn to generate the important sentences from the context of the remaining sentences. It is trained based on two objectives: 1) Gap Sentence Generation (GSG); and 2) Masked Language Modeling (MLM). The Gap Sentence Generation (GSG) pretraining objective aims at generating a set of sentences given other context sentences in a document and is experimented using three variants: a) random masking of *n* sentences; b) masking of first *n* sentences; and c) masking of *m* high scoring sentences on ROUGE-1 metric with respect to the rest of the document.

- *ProphetNet* [68] is a sequence-to-sequence model that is pre-trained with a self-supervised objective of simultaneous future n-gram prediction using n-stream self-attention and mask-based autoencoder denoising task. The future n-gram prediction enables the model to plan n-step ahead to future tokens, preventing the possibility of overfitting to local n-grams correlations.

2.4 Research Aim III

This section discusses the related works on Research Aim III: Entity-driven Fact-Aware Multi-document Abstractive Summarization of Biomedical Literature.

While abstractive summarization is well studied for summarization of news articles

with success attributed to the availability of a massive amount of training data, their applicability to scholarly articles, particularly, in the biomedical domain is limited. Further, although named entities have been extensively studied to convey the semantics of an article (news, scientific, social media) and the saliency of individual sentences [69] within an article, they have not been widely used as part of modeling abstractive summarization. [70] performed entity-aware single-document abstractive summarization using reinforcement learning for training. Their pipeline-based approach consists of an entity-aware content selection module and abstract generation module. They evaluate their approach on the CNN/Daily Mail and NYT corpora. [69] perform entity-driven multi-document abstractive summarization of news articles (WikiSum, and Multi-News) using an encoder-decoder framework augmented with Graph Attention Network (GAT). [71] proposed EntityRank, an extension of the LexRank [24] graph-based algorithm, for entity-supported summarization of biomedical abstracts.

There have been a few recent efforts towards knowledge/fact-aware abstractive summarization in different domains. [72] introduced a Fact-aware Abstractive Summarization model called FaSum for improving the factual consistency of summaries in the domain of news articles. However, their approach does not leverage named entities for fact retrieval. [73] extended a transformer-based abstractive summarization model using entities disambiguated and linked to Wikidata knowledge graph and attending to the entities for summarization of news articles. Their approach, however, does not perform named-entity based fact retrieval from the knowledge base constrained by the article to be summarized and the named entities. [74] developed an unsupervised pipeline-based approach for knowledge-infused abstractive summarization for condensing patient-to-clinician diagnostic interviews based on Multi-Sentence Compression [27] and Integer Linear Programming [75]. Nevertheless, their approach uses domain-specific lexicons as knowledge source for filtering irrelevant utterances and for retroffiting language models [61] and, does not leverage named entities or facts as part of an end-to-end training of models. [76] proposed

Biomed-Summarizer, a framework for extractive summarization of biomedical literature in a multi-document setting and evaluated on PubMed abstracts. [77] built a model for abstractive summarization of long documents using a discourse-aware encoder-decoder framework and experimented on two large scale datasets including research articles collected from PubMed. To address the challenge associated with the scarcity of large-scale training data in the biomedical domain, [78] released MS2 (Multi-Document Summarization of Medical Studies). They experimented with BART [66] for abstractive summary generation on the dataset they introduced in a traditional multi-doc-to-summary setting.

Though all the aforementioned studies conduct abstractive summarization of biomedical literature or the use of facts mined from knowledge bases for a different domain, they follow the well-established paradigm of source-document vs summary pairing during training/inference of models. Our approach is different in that we augment the state-of-the-art abstractive summarization models with additional contextual signals during training/inference and apply them to the biomedical domain.

2.5 Research Aim IV

This section discusses the related works on Research Aim IV: Improving the Factual Accuracy of Clinical Text Summarization using Multi-Objective Optimization.

We start with a motivation to demonstrate that multi-optimization of different cost functions is important in the context of clinical text summary generation since the impression corresponding to a given set of findings may not necessarily be directly inferred from the findings, partly because clinicians use their domain knowledge while composing an impression for a set of findings. Consequently, it is imperative that findings-to-impression mapping be tightly optimized. Figure 2.1 shows a case where there is not a significant overlap between the named entities in findings and impression, and thus the single optimization objective investigated in the previous chapter will not have enough guidance to predict the

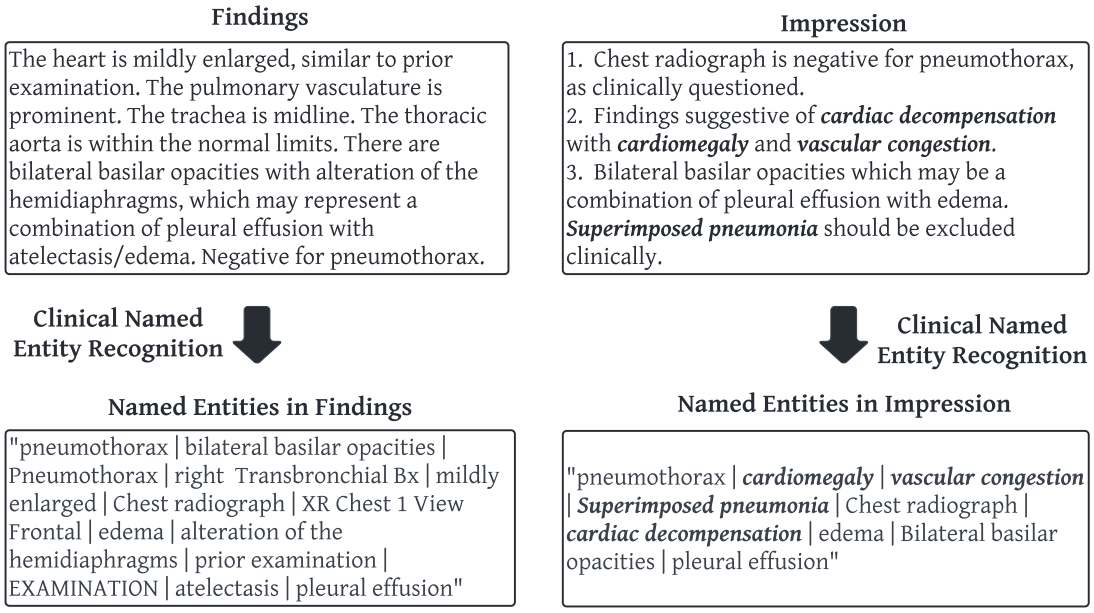

Figure 2.1: Example findings-to-impression pair. It can be seen that there are named entities that appear in the impression but not in the findings. Single optimization with predicting impression does not put more weights to these named entities since these entities are treated just like other tokens in the impression. We show through experiments that transformer models in their vanilla training setting perform poorly in terms of recall of named entities in ground truth summary/impression and semantic equivalencec.

target entities/facts. This is mainly because the training objectives are based on Maximum Likelihood Estimation (MLE) and optimizing a single cost function will not be enough, since every token in the target summary (whether it be a semantically important or not) is equally considered from the vocabulary. We believe optimizing named entity generation and facts describing named entities will boost recall with respect to named entities in impressions. Further, since these named entities that appear in an impression, but not in the findings can have significant semantics, we capture their related facts from the knowledge bases. Our multi-objective optimization training approach enables to give more weights to these named entities and the related facts rather than the MLE w.r.t just the impression. The reason is because clinicians look at radiology image, in addition to the findings while writing impression.

The birth of Transformer encoder-decoder models [36, 79, 66, 67, 80] has led to significant advances in abstractive summarization in the domains of news articles [81, 82, 35, 41] and scientific articles [77, 83, 58]. Nevertheless, their application to the summarization of clinical notes has not been adequately explored. [84] proposed a model based on Pointer-Generator-Networks [35] for abstractive summarization of radiology reports by linking entities in a clinical note to domain-specific ontology from UMLS [85] and RadLex [86]. They use pairings of findings and impressions for the abstractive summarization task where findings form the input sequences and impressions form the target summaries for training. [87] propose a two-stage model consisting of a content selector and abstractive summarizer for clinical abstractive summarization. The content selector identifies ontological terms from the findings using a medical ontology (RadLex) and the summarizer is trained to generate summaries (impressions). They use Bi-LSTMs to encode findings and use LSTMs to encode the ontological terms followed by an LSTM-based decoder to generate a summary. [88] built a model for extractive summarization of clinical notes of patients with diabetes and hypertension to generate disease-specific summaries. They framed the extractive summary generation problem as a sentence classification problem and experimented on a clinical dataset consisting of 3,453 clinical notes collected for 762 patients. [89] proposed a model comprised of syntax-based negation detection and semantic clinical concept recognition for extractive summarization of clinical text. They conducted their experiments on the MIMIC-III [90] dataset. While the aforementioned approaches employ different techniques for clinical text summarization, we show experimentally that our proposed knowledge-aware Multi-Objective Optimization (MOO) improves the factuality of the generated summaries when compared to strong state-of-the-art transformer-based abstractive summarization models.

Topic-Centric Unsupervised Multi-document Summarization

"Everything in this world has a hidden meaning."

—Nikos Kazantzakis, 1883 – 1957

In this chapter, we introduce an unsupervised framework for discovering intrinsic semantics of a corpus of documents using latent topical analysis and use the latent topics discovered to guide extractive and abstractive summarization pipelines. While the primary domain of discourse is scientific articles, we also test the proposed approach on a benchmark news articles dataset.

3.1 Why (Motivation)

There is an increasing number of scientific articles in technical fields that share common latent topics that describe the internal semantic structure of the articles. Automatically identifying these hidden topics across scientific articles and clustering the scientific articles based on their topical relevance and generating a topic-centric summary improves the process of harvesting scientific information.

3.2 What (Problem Statement)

We propose extractive and abstractive approaches to topic-centric multi-document summarization. Specifically, we devise unsupervised multi-document extractive and abstractive summarization frameworks and apply to abstractive summarization of topically-clustered scientific and news articles. The abstractive approach follows a sequence of extractive phase and abstractive phase. We performed evaluation of the extractive summaries using the Recall Oriented Understudy for Gisting Evaluation (ROUGE) metrics [91] and the abstractive summaries by humans on five evaluation metrics (Coherence, Readability, Entailment, Conciseness, and Grammar) and copy rate (paraphrasing). The proposed frameworks are evaluated on two datasets: 1) MAG-20 (the 100-most cited articles across 20 fields of study from the Microsoft Academic Graph); and 2) DUC-2004 (Document Understanding Conference of 2004 - Task-2) benchmark dataset of news articles.

3.3 How (Approach)

3.3.1 Data Curation

For this aim, we queried the Microsoft Academic Graph (MAG) for the 100 most-cited abstracts for each of the 20 Fields of Study (FoS) for scientific papers published in the years 2016 - 2020. We refer to the dataset we build from MAG, the MAG-20 dataset, and is made publicly available [1]. In addition to MAG-20 dataset, we also use a benchmark DUC-2004 dataset for comparing our proposed approach with two baseline approaches on unsupervised multi-document abstractive summarization.

The 20 FoS we are focused on are:

- Artificial Intelligence

[1] https://github.com/AmanuelF/MAG-20-Abstractive_Summarization

- Artificial Neural Network
- Big Data
- Case-Based Reasoning
- Cybernetics
- Cyberwarfare
- Data Mining
- Data Science
- Decision Support System
- Electronic Warfare
- Expert System
- Human-Machine Interaction
- Intelligent Agent
- Knowledge-Based Systems
- Machine Learning
- Multi-Agent System
- Prediction Algorithms
- Predictive Analytics
- Predictive Modeling
- Sensor Fusion

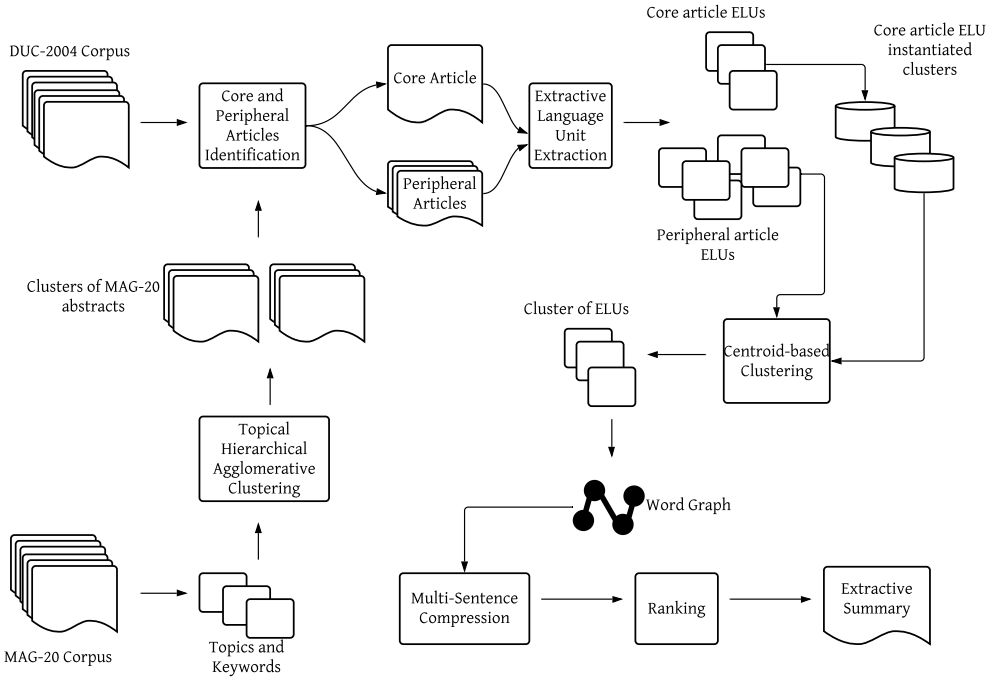

Figure 3.1: Extractive Phase

3.3.2 Proposed Framework

Extractive Phase

Figure 3.1 shows the extractive phase of the proposed framework where extractive language units are extracted for MAG-20 and DUC-2004 datasets. DUC-2004 comes with news articles that are topically clustered. Since MAG-20 is a dataset we build for this research aim, we perform clustering of the 100 most cited scientific articles in each of the 20 fields of study by first discovering latent topics using Latent Dirichlet Allocation (LDA) [92] followed by clustering of topics and then clustering of the articles driven by their topic membership.

Topic Modeling For each field of study (FoS) in the MAG-20 dataset, our first task is to identify the dominant topics in the 100 most cited articles and perform clustering of the

topics. To identify topics in the abstracts, we build an ensemble of Latent Dirichlet Allocation (LDA) topic models by specifying the number of topics in the range of 2 to 92 (with increments of 10) and generate the most dominant topic for each article. The reason for building different topic models corresponding to different number of topics is to determine the optimal number of topics from an ensemble of the LDA models that maximizes topic coherence score [93]. Topical hierarchical clustering is conducted on the dominant topics for all articles associated with the corresponding LDA models that give the highest coherence score. A vector representation of a topic is generated by first embedding each keyword in a topic using SciBERT [94] and concatenating the representations of the individual keywords followed by dimensionality reduction using t-SNE [95]. Figure 3.2 shows plots of number of topics vs coherence scores for selected fields of study. Higher topic coherence is associated with better interpretability of topics generated. The motivation for using HAC over embedded topics is that standard HAC is syntactic; consequently, in order to enable semantics-based clustering, we form topics using similarity metric based on semantics captured through SciBERT.

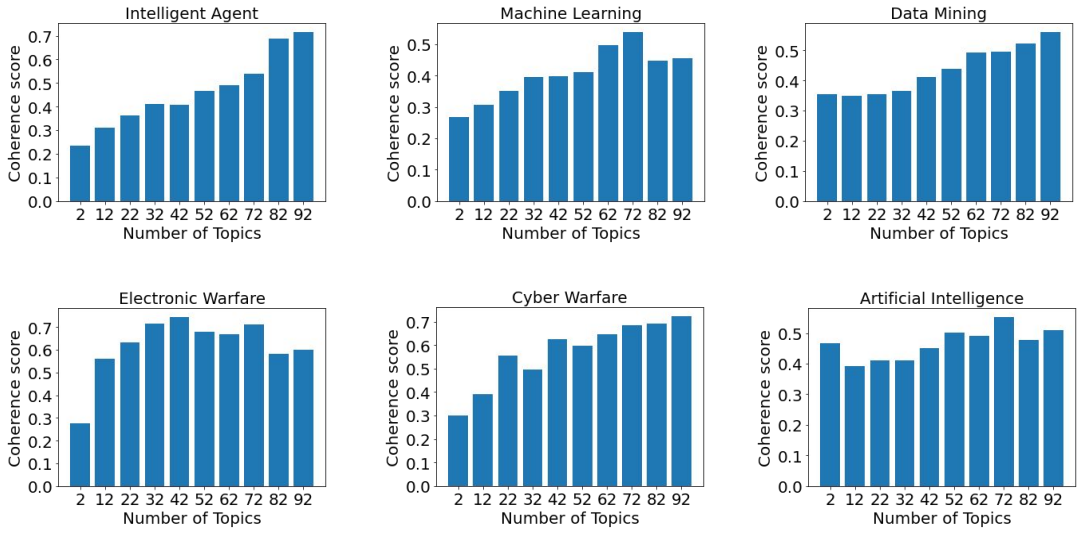

Figure 3.2: Number of topics vs Coherence Scores for selected sample fields of study.

Topical Hierarchical Agglomerative Clustering Since keywords in different topics can be semantically redundant, we cluster topics having high similarity among their keywords using hierarchical clustering. For this, we encode a topic using the concatenation of the SciBERT encodings of its constituent keywords. We then perform topical hierarchical agglomerative clustering of the topics. To determine the optimal number of clusters to cluster topics into, we run hierarchical clustering for several clusters ranging from 2 to the total number of topics and the number of clusters that gives the highest Silhouette coefficient is set as the optimal number of clusters.

We introduce a topical similarity metric (Equation 3.1) for measuring the similarity between a pair of topics. As can be seen in the equation, each keyword in a topic is compared with all the keywords in another topic, and the sum of highest similarity scores is preserved. This similarity metric is inspired by Word Mover's Distance proposed in [96].

$$sim(\text{Topic-i, Topic-j}) = \sum_{i \in \text{Topic-i}} maxcos(i, \text{Topic-j}) \tag{3.1}$$

where

$maxcos(i, \text{Topic-j}) =$ maximum of cosine similarities between term i and terms in Topic-j

A MAG-20 abstract is associated with a topic that is the most dominant among all topics the abstract addresses. Table 3.1 shows topical distribution among abstracts for a field of study.

Figure 3.3 shows topics clustered together using agglomerative hierarchical clustering applied to the topics discovered.

Core and Peripheral Articles Identification Since our proposed approach follows a centroid-based summarization paradigm, we identify the article that is semantically the closest to other articles in a cluster and designate it as the core article; other articles in the cluster are similarly designated as peripheral articles. Equation 3.2 computes the Cross-

Abstract ID	**Dominant Topic**	**Dominant Topic Contribution(%)**	**Topic Keywords**
9	12	0.99	radar, signal, communication, base, data, challenge, scenario, model, detection, nature
18	12	0.99	radar, signal, communication, base, data, challenge, scenario, model, detection, nature
17	14	0.95	assessment, processing, assess, forecast, bandwidth, receiver, physical, accuracy, technology, electronic_warfare
11	14	0.88	assessment, processing, assess, forecast, bandwidth, receiver, physical, accuracy, technology, electronic_warfare
5	18	0.99	classification, require, vehicle, overlap, environment, application, snr, wacr, illustrate, commercial
23	18	0.32	classification, require, vehicle, overlap, environment, application, snr, wacr, illustrate, commercial
6	19	0.87	inspire, state, device, accelerator, small, size, high, power, ved, advantage
4	19	0.59	inspire, state, device, accelerator, small, size, high, power, ved, advantage

Table 3.1: Topical distribution of abstracts. Abstract with ID 6 is about Topic-19 in 87% of its content while Abstract with ID 4 is about the same topic in 59% of its content.

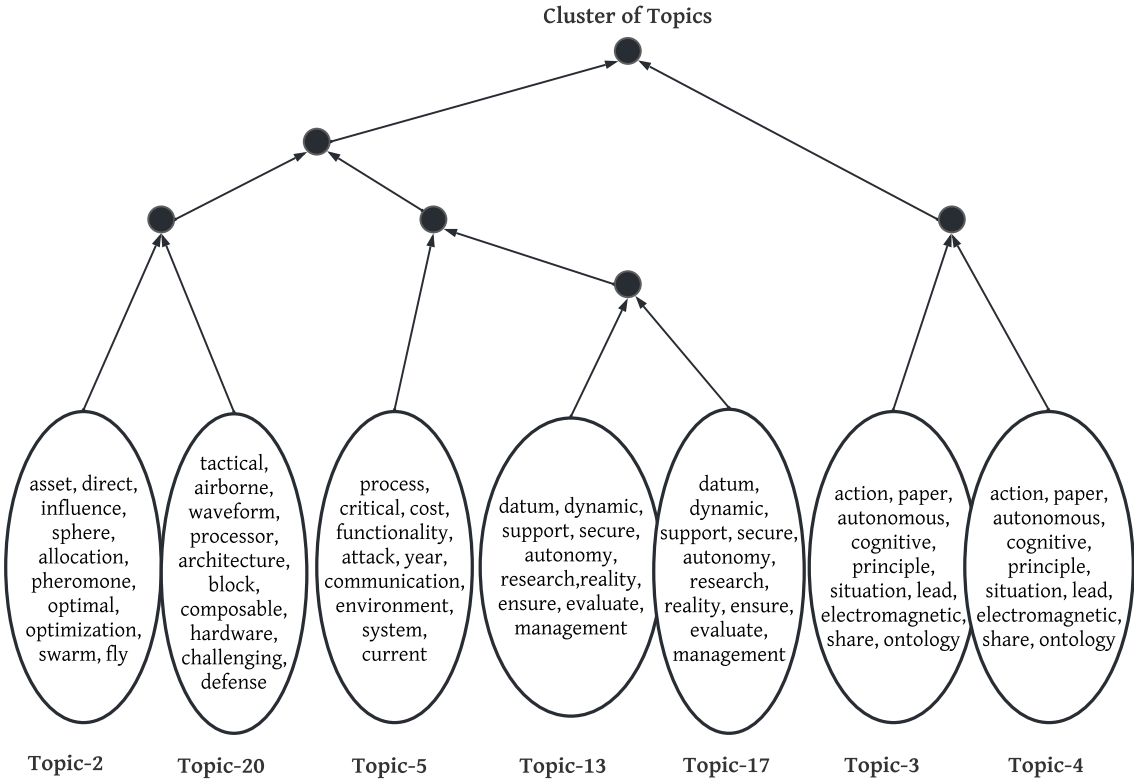

Figure 3.3: Sample topics clustered using hierarchical agglomerative clustering based on semantics-based similarity metric. As can be seen, some topics have same top keywords (e.g., topic-13 and topic-17 as well as topic-3 and topic-4. This is because each keyword contributes different weights to different topics. For an intuitive description of this behavior of LDA (i.e., topic-word distributions), refer to [1] §17.2

.

Article Similarity Score of an article in a cluster. An article with the highest cumulative semantic similarity with other articles in a cluster is chosen as the core article. We consider the rest of the articles in the cluster as peripheral articles. We use gensim's implementation of doc2vec[2] to encode an article. Figure 3.4 shows a heatmap of cross-article similarities among news articles in a DUC-2004 topic.

$$CAS_i = \frac{\sum_{i,j \in C} doc2vec\_sim(i,j)}{N} \tag{3.2}$$

[2] https://radimrehurek.com/gensim/models/doc2vec.html

where $i \neq j$

N- Number of articles in the cluster

C- The cluster of articles

$doc2vec\_sim$ - doc2vec-based cosine similarity

$article_{core} = argmax\{CAS_i; i \in C\}$

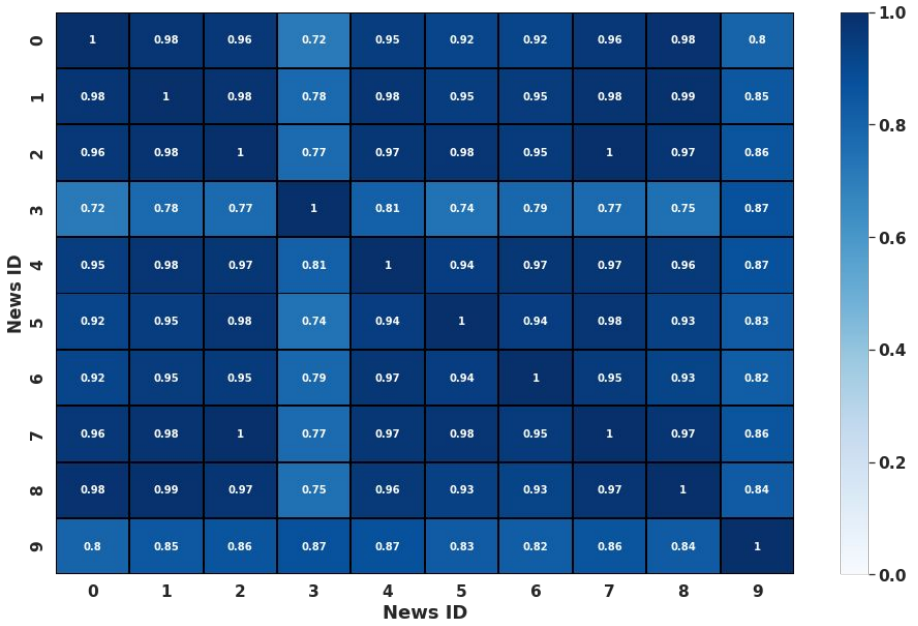

Figure 3.4: Cross-Article Similarity among articles in a topic in the DUC-2004 task. Momentary glance reveals news article 3 or 9 cannot be a core article. In this heatmap, it can be seen that article 2 is the core article across all the articles in this topic. (Similarity scale legend on the right - the darker the color, the more similar the documents).

Centroid based Clustering After core and peripheral articles in a cluster are identified, we generate extractive language units from the core and each of the peripheral articles. Recent studies on centroid based clustering approaches to summarization utilized sentences in documents as standalone language units to initiate clusters and to quantify semantic relatedness with sentences in other documents [30, 48]. This approach, however, breaks the interdependence among sentences in a document and eventually leads to incoherent summaries. We address this limitation by identifying the sentences that are interdependent and

preserve them as one extractive language unit [97, 98]. For this, we use neural coreference resolution [99] to identify coreferents across sentences and to group these sentences into one extractive language unit. We use an implementation of neural coreference resolution from hugging face [3] for our study. Figure 3.5 shows an example of identifying sentences that have coreference dependency and are grouped into one language unit. Sentences which do not exhibit any dependency with other sentences form individual standalone language units. The language units from the core article are used to instantiate (seed) clusters.

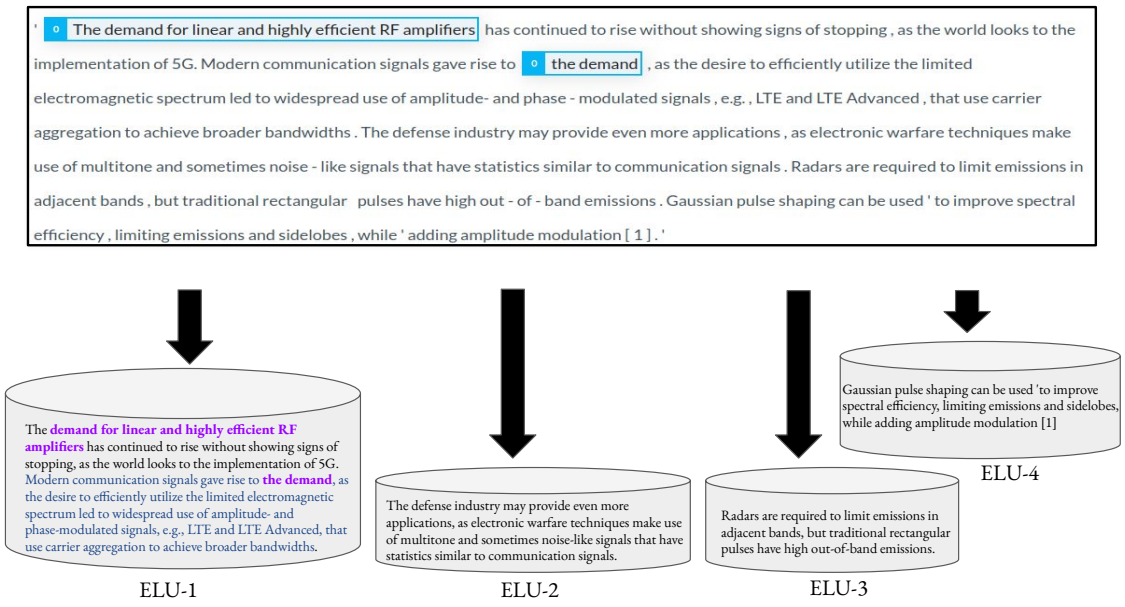

Figure 3.5: Extractive Language Unit (ELU) identification using coreference resolution and clusters initialization from core article. Segments of the article highlighted in blue show coreference dependency and contribute to ELU-1. Thus, sentences having common coreferents are kept together.

Once the extractive language units (ELUs) from the core article have instantiated clusters, the language units from the peripheral articles are placed into a cluster based on the cosine similarity between a language unit embedding of an ELU from the peripheral article and the language unit embeddings of the ELUs from the core article. A language unit embedding is constructed by concatenating the embeddings of the sentences using sentence-

[3] https://github.com/huggingface/neuralcoref

BERT [100] in the language unit and performing dimensionality reduction using t-SNE. The purpose of dimensionality reduction is to have a uniform dimension among language units even when they contain different numbers of sentences so that cosine similarity can be computed. We use 300 dimensions for representing an ELU for comparison with other ELUs. Equation 3.3 shows the technique to perform dimensionality reduction to represent an ELU.

$$ELU_{embd} = \text{t-sne}_{300}(\bigoplus_{i}^{N} sent - BERT(S_i)) \tag{3.3}$$

where $S_i \in Sentences\ in\ a\ Language\ Unit$

$N - Number\ of\ sentences\ in\ a\ language\ unit$

$\bigoplus - Concatenation\ Operator$

Multi-Sentence Compression The number of clusters formed in the centroid-based clustering stage is the same as the number of extractive language units (ELUs) in the core article. After clusters of Extractive Language Units (ELUs) are formed, we build word graphs [28] for each cluster. We use Python's NetworkX [4] to construct the word graph. Figure 3.6 shows a sample word graph constructed for a cluster consisting of the following ELUs:

ELU_1 = "Radars are required to limit emissions in adjacent bands, but traditional rectangular pulses have high out-of-band emissions."
ELU_2 = "Millimeter wave radars are popularly used in last-mile radar based defense systems."

We develop an algorithm for extracting paths based on topical coverage and relevance. A path is selected using an additional criterion that a candidate path should at least span two ELUs in the cluster. Next, we generate topically informative and relevant paths from

[4] https://networkx.github.io/

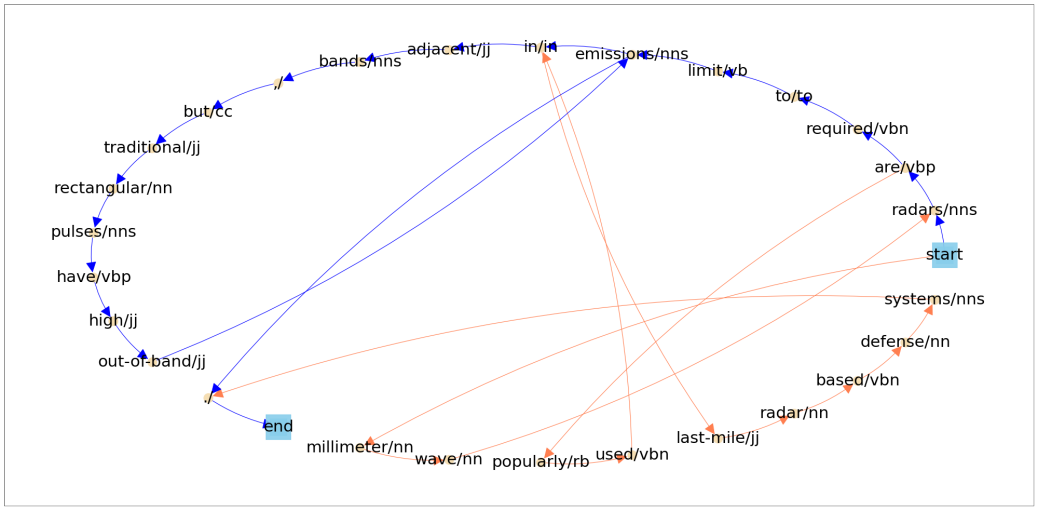

Figure 3.6: Word Graph for two ELUs using NetworkX. Tokens and PoS tags of the tokens are used for a node.

the word graph while maintaining the 100-word summary (MAX_LEN) limit. Topical coverage (Equation 3.4) measures how well a path covers the dominant topics discussed by the articles of the ELUs. Relevance (Equation 3.5) measures how relevant a path is to the ELUs. The cumulative score of a path (Equation 3.6) is determined by a weighted sum of topical coverage and relevance. We experimented with values of α in the range of 0 to 1.

Topical Coverage Formulation

$$Coverage(C_{path}, C_{topics}) = \frac{1}{|C_{path}|} \sum_{t_{C_{path}} \in C_{path}}^{|C_{path}|} \frac{1}{|C_{topics}|} \sum_{K_c \in C_{topics}}^{|C_{topics}|} maxcos(t_{C_{path}}, K_c) \tag{3.4}$$

where, C_{path} - Candidate path

C_{topics} - Cluster of topics

$t_{c_{path}}$ - Term t in Candidate path C_{path}

K_c - Topic K in cluster of topics C_{topics}

Topical coverage is measured with respect to the cluster of topics.

Topical coverage formulation given in Equation 3.4 is described below:

- Topical coverage is measured with respect to the cluster of the topics. A candidate path with coverage 1.0 indicates the path fully covers all the dominant topics.
- Consider word graph WG_1 built from a cluster C of some ELUs. Let this cluster C have ELUs in the range ELU_1, ELU_2, $\cdots$, ELU_n.
- We know that these ELUs are extracted from different articles in the range $Article_1$, $Article_2$, $\cdots$, $Article_n$.
- We identified that there are dominant topics $Topic_1$, $Topic_2$, $\cdots$, $Topic_m$ corresponding to each of these articles. We also assert that these topics are clustered into a cluster of topics using HAC.
- Consider a candidate path in the word graph that has words $word_1$, $word_2$, $\cdots$, $word_p$.
- Next, we iterate through each word in the candidate path and compute its maximum cosine similarity (as used in Equation 3.1) with each of the dominant topics in the set of dominant topics $Topic_1$, $Topic_2$, $\cdots$, $Topic_m$; thus, for each word, we generate $maxcos$ scores as many as the number of the dominant topics. We then determine the score of the word as the average of these $maxcos$ scores (inner summation in Equation 3.4).
- Finally, the topical coverage of the candidate path is computed as the average of the scores of the words (outer summation in Equation 3.4).

Path Relevance Formulation

$$Relevance(C_{path}, C_{ELU}) = \frac{\vec{v}(C_{path}) \cdot \vec{v}(C_{ELU})}{|\vec{v}(C_{path})| \cdot |\vec{v}(C_{ELU})|} \tag{3.5}$$

where, C_{path} - Candidate Path

C_{ELU} - Cluster of ELUs

$\vec{v}(C_{path})$ - Vectorial Representation of Candidate Path

$\vec{v}(C_{ELU})$ - Vectorial Representation of Cluster of ELUs

Path relevance is measured with respect to the ELUs.

Cumulative Score

$$\text{Score}_{\text{cumulative}}(C_{path}) = \alpha \cdot \text{Coverage}(C_{path}, C_{topics}) + (1 - \alpha) \cdot \text{Relevance}(C_{path}, C_{ELU}) \tag{3.6}$$

A path is selected from the word graph 1) if the path is longer than the average minimum length of a sentence in an FoS or DUC-2004 topic and smaller than the average maximum length of a sentence; 2) if the combined topical coverage and relevance for the path meets or exceeds a threshold τ of 0.5. If a path picked from the word graph is semantically similar to an already selected path by an order of threshold δ of 0.8 or more, we compare the combined topical coverage and relevance of the two paths and keep the one with a higher score and remove the other. The selection of 0.8 is based on empirical observations. Algorithm-1 outlines the path ranking and selection algorithm which is inspired by Maximal Marginal Relevance (MMR) [101].

Abstractive Phase

Figure 3.7 shows a pipeline of the abstractive phase of the proposed framework. A headline generation component is included for the DUC-2004 part since DUC-2004 news articles do not come with headlines while MAG-20 abstracts have titles which we use in the abstractive phase.

Algorithm 1: Path Ranking algorithm

procedure RankPaths *(τ, δ, MAX_LEN)*

Initialization

$paths_{selected} \leftarrow \emptyset$

$paths_{candidate} \leftarrow \forall p \in$ P | P: *the set of all paths in the word graph*

$langUnits \leftarrow \forall l \in$ L | L: *language units forming the word graph*

for $\forall c_{path} \in paths_{candidate}$ **do**

$\text{Score}_{\text{cumulative}}(c_{path}) = \alpha \cdot \text{Coverage}(c_{path}, C_{topics}) + (1 - \alpha) \cdot \text{Relevance}(c_{path}, C_{ELU})$

if $Score(c_{path}) \geq \tau$ **then**

if $c_{path} \notin langUnits$ **then**

$semSim = max_{s_{path} \in paths_{selected}} cosSim(c_{path},\ s_{path})$

$s_{path} = \underset{s \in paths_{selected}}{\text{argmax}} \ cosSim(c_{path}, s)$

if $semSim \geq \delta$ **then**

$path_{max} = \max_{path \in \{c_{path}, s_{path}\}} (Score(c_{path}), Score(s_{path}))$

$path_{min} = \min_{path \in \{c_{path}, s_{path}\}} (Score(c_{path}), Score(s_{path}))$

$Size = |\textbf{Tokenize}(paths_{selected})|$

if $Size \leq MAX\_LEN$ **then**

do ADD$(paths_{selected}, path_{max})$

do DELETE $path_{min}$

end

else

return $paths_{selected}$

end

end

else

do ADD$(paths_{selected}, c_{path})$

end

end

end

end

return $paths_{selected}$

end

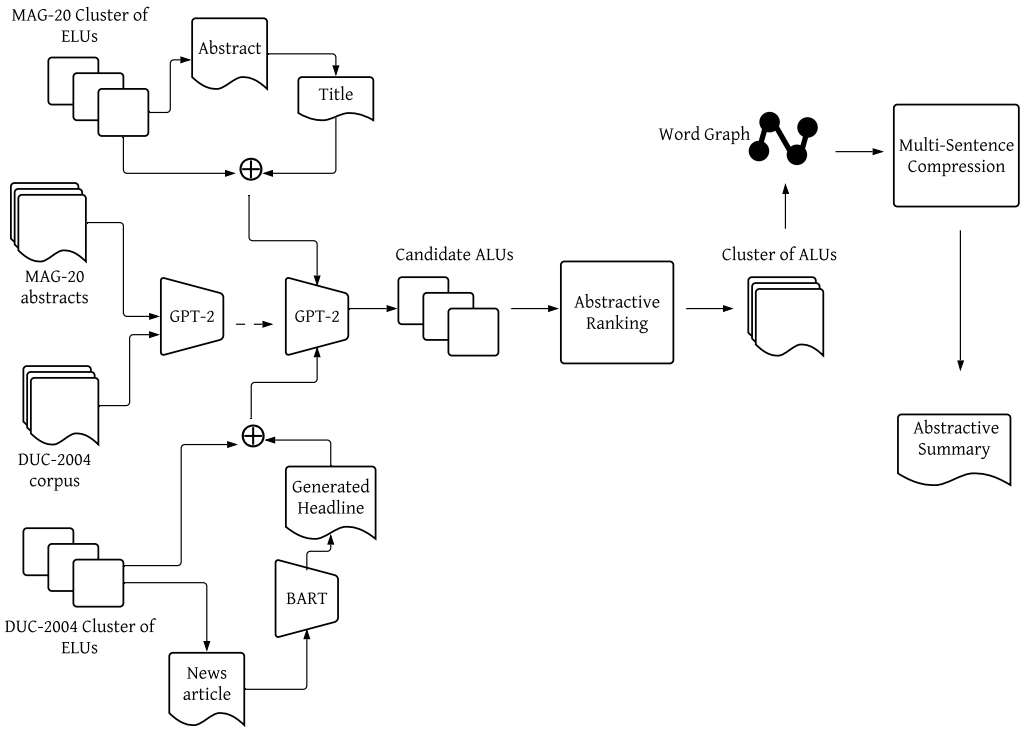

Figure 3.7: Abstractive Phase. Since each ELU is derived from an abstract, we use the abstract to query for the title of the paper from MAG. We use the combination of the title and ELU to generate candidate ALUs.

Abstractive Language Unit (ALU) Generation We start our abstractive phase with a pragmatic assumption that the title/headline of an article is an abstraction of the individual extractive language units (ELUs) within the same article. We propose a method to generate an ALU for an ELU using the ELU and title/headline as prompts for generating text. Combining bidirectional encodings of the title/headline with an ELU enables generating abstractive text. For ELUs consisting of two or more sentences, we encode each sentence using sentence-BERT [100] and then we concatenate these representations. Next, we perform dimensionality reduction using t-SNE to encode an ELU. For encoding a title/-headline, we use sentence-BERT without dimensionality reduction. We fine-tune a GPT-2 model (architecture shown in Figure 3.8) for an FoS (Figure 3.9) and use the fine-tuned GPT-2 model to generate ALUs given a concatenation of the bidirectional encodings of the

ELU and the title/headline of an article. We fine-tune a GPT-2 model [52] such that it has 124M parameters and generates 10 candidate ALUs.

GPT-2 is one of the class of Generative Pre-Trained Transformer (GPT - transformer decoder) models that is originally pretrained in a self-supervised manner with a language modeling loss for an autoregressive text generation task using millions of web pages (WebText dataset). Figure 3.8 shows the architecture of GPT-2 adapted from [50]. We use GPT-2 to specifically prompt generation of absractive version of an extractive language unit. As stated in [52] under §3.6 pp.6, we add the text "TL;DR" following an extractive language unit to generate abstractive language units.

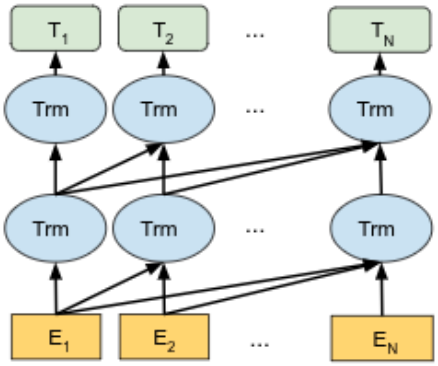

Figure 3.8: Architecture of GPT-2. E_t is the embedding of a token in a sequence at position t. T_t is the output token to be predicted.

The language modeling loss used to train GPT-2 follows Equation 3.7.

$$\mathcal{L}_{LM} = -\sum_{t} \log p(x_t|\mathbf{x}_{1:t-1}) \tag{3.7}$$

While fine-tuning, we set the temperature to 0.7, number of generated samples to 10, top_k random sampling to 2 to generate more abstractive ALUs and minimize redundancy [52]. We train the GPT-2 for 10 epochs with a batch size of 10 and attain a loss of 2.16. We select an ALU that maximizes semantic similarity and minimizes syntactic similarity with the ELU used for generation. We use the normalized sum of ROUGE-1(R_1) and ROUGE-2

(R_2) for syntactic similarity. We introduce an *abstractiveness score* for an ALU, as shown in Equation 3.8.

We use BART [66] for headline generation for each DUC-2004 article that is later used for ALU generation along with an ELU.

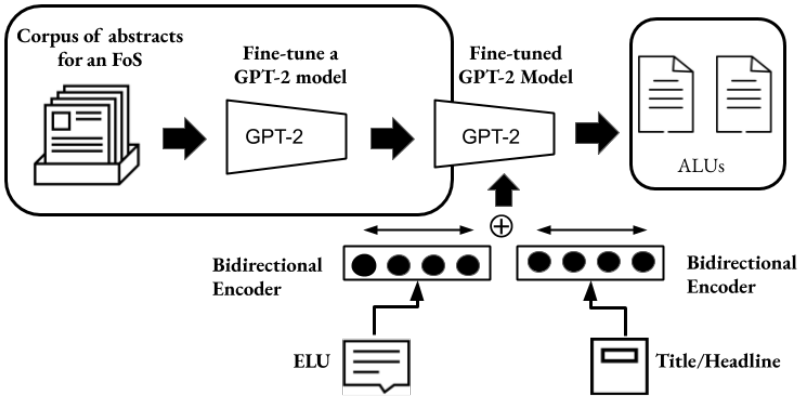

Figure 3.9: ALUs generation using GPT-2.

$$\text{Score}_{\text{abstractive}}(\text{ALU, ELU}) = cossim_{d\text{-BERT}}(\text{ALU, ELU}) - \frac{[R_1(\text{ALU, ELU}) + R_2(\text{ALU, ELU})]}{[R_1(\text{ALU, ALU}) + R_2(\text{ALU, ALU})]} \tag{3.8}$$

where

ALU - Abstractive Language Unit

ELU - Extractive Language Unit

$cossim_{d\text{-BERT}}$ - Cosine similarity on d-dimension BERT embeddings

We select an ALU that gives the highest *abstractiveness score* (Equation 3.8) from candidate ALUs since one of the things abstraction entails is higher semantic similarity, and lower lexical similarity (i.e., paraphrasing).

$$ALU_{selected} = \underset{ALU \in ALUs}{\text{argmax}} \; Score_{abstractive}(ALU, ELU) \tag{3.9}$$

where $ALUs$ - set of sample ALUs generated for an ELU by the fine-tuned GPT-2 model.

Table 3.2 shows a sample ELU and sample ALUs generated from which will be determined the ALU with the highest semantic similarity and lowest lexical similarity.

Multi-Sentence Compression After generating ALUs for a cluster, we build a word graph and run our MSC algorithm for each cluster as used in the extractive phase; i.e., the same ranking formulation and path selection algorithm are used for selecting informative, and topically relevant paths from a word graph built, this time from a cluster of ALUs. Figure 3.10 shows a cluster of ALUs and the generated fused paths that form the final abstractive summary.

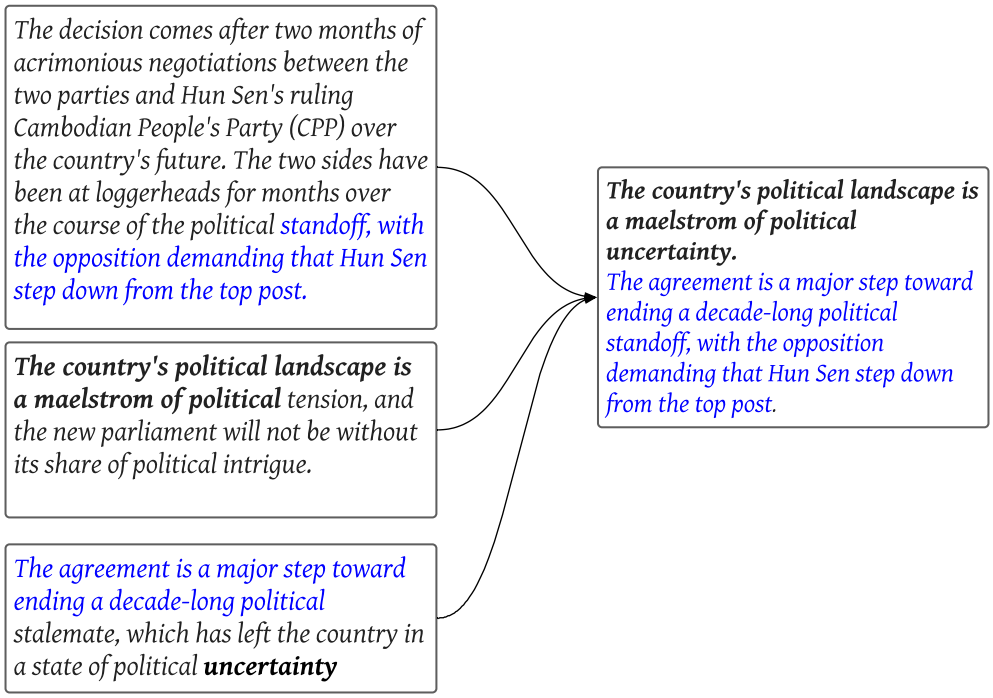

Figure 3.10: Fusing ALUs into final paths. This is for the DUC-2004 dataset.

Extractive LU	Source title	Abstractive LUs
Radars are required to limit emissions in adjacent bands, but traditional rectangular pulses have high out-of-band emissions.	The Various Angles of Outphasing PAs: Competitiveness of Outphasing in Efficient Linear PA Applications	1. The best approach to reduce out-of-band emissions is to use an out-of-band pulse to generate a pulse of high-frequency power. 2. The pulse of high-frequency power is proportional to the pulse width and pulse length. This is known as the Linear PA approach. 3. The high-frequency pulse is derived from the resonant frequency of the pulse. 4. The out-of-band pulse is a 2-dimensional wave that is not a 2-dimensional wave in the linear PA.
With the now widespread availability of software-defined radio technology for wireless networks, the distinction between jamming in the original electronic warfare sense and wireless cybersecurity attacks becomes hazy.	A communications jamming taxonomy	1. The problem is that jamming is an inherently insecure practice, because it requires that a computer or device be in close proximity to a transmitting signal and that the signal be encrypted and then broadcast over the Internet. 2. To be a successful jammer, one needs to be able to locate a specific target, but in order to do that, it is necessary to have some level of control over the target. 3. With the now widespread availability of software-defined radio technology for wireless networks, the distinction between jamming in the original electronic warfare sense and wireless cybersecurity attacks becomes hazy. A communications jamming taxonomy is presented that includes both. 4. A communications jamming taxonomy is useful to distinguish the types of hacking that can be performed by a single hacker, the attack that can be performed by multiple hackers working in parallel, and the attack that can be performed by a team of hackers working in parallel. 5. In the case of wireless jamming, the term ”jamming” has been used to refer to the intentional interruption of a wireless network in order to prevent or degrade its performance. The term ”jamming” can also refer to a deliberate and deliberate attempt to cause a network to malfunction or become inaccessible.
A taxonomy delineates these concepts in the rapidly expanding field of wireless security, classifying communication jammers' theoretical behaviors and characteristics.	A communications jamming taxonomy	1. A communications jamming taxonomy is a practical tool to classify wireless communications jammers, with the goal of making the use of jammers less dangerous and more efficient. 2. A communications jamming taxonomy is also useful for identifying the vulnerabilities of wireless systems and their prevention or mitigation strategies. 3. Operational communications jamming is a class of wireless communication jammers that uses a variety of methods to confuse, disrupt, or spoof a communications channel. 4. A communications jamming taxonomy for the wireless security field is developed, which is based on the fundamental research of the recent years. 5. The proposed taxonomy is based on the physical properties of wireless communication jammers, which are based on their overall physical design and configuration. The proposed taxonomy is not a definition of the behavior of a communications jammer, but rather a description of the characteristics of jammers.

Table 3.2: Abstractive Language Units generated using fine-tuned GPT-2.

3.3.3 Experiments and Results

Our proposed framework is compared with two baseline approaches which are also on the same task of unsupervised multi-document summarization. The baseline approaches were evaluated on news articles. We evaluate the baseline approaches and our approach on the benchmark dataset and our MAG-20 curated dataset. Tables 3.3 through 3.6 show comparison of our proposed approach and the baseline models on extractive, copy rate, and human evaluation metrics. Equations 3.10 - 3.15 show the formulation for computing the Recall Oriented Understudy for Gisting Evaluation (ROUGE). We use three variants of ROUGE, as used in the literature for quantifying the lexical overlap of a generated summary w.r.t a reference summary. The first two variants which are formalized as ROUGE-N are based on n-gram overlap and specifically are 1) ROUGE-1 (which measures the unigram overlap between a generated summary and the reference summary); and 2) ROUGE-2 (which measures the bigram overlap between a generated summary and a reference summary). The third ROUGE metric called ROUGE-L is a measure of the longest sub-sequence shared between a generated summary and a reference summary. ROUGE evaluation for MAG-20 is done against the source articles as the reference summaries since our task is unsupervised and we do not have human-written summaries, while for DUC-2004, evaluation is conducted against the human-written summaries as the reference summaries. Equation 3.16 formalizes copy rate which quantifies paraphrasing in terms of novel tokens generated. In this dissertation, all the ROUGE scores reported are the harmonic mean of ROUGE precision and ROUGE recall.

$$\text{ROUGE-N-recall} = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count(gram_n)} \tag{3.10}$$

$$\text{ROUGE-N-precision} = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S' \in CandidateSummaries} \sum_{gram_n \in S'} Count(gram_n)} \quad (3.11)$$

$$\text{ROUGE-N-F-1} = 2 \cdot \frac{\text{ROUGE-N-recall} \cdot \text{ROUGE-N-precision}}{\text{ROUGE-N-recall} + \text{ROUGE-N-precision}} \quad (3.12)$$

Similarly, ROUGE-L is formulated as follows.

$$\text{ROUGE-L-recall} = \frac{LCS(ReferenceSummary, CandidateSummary)}{|ReferenceSummary|} \quad (3.13)$$

$$\text{ROUGE-L-recall} = \frac{LCS(ReferenceSummary, CandidateSummary)}{|CandidateSummary|} \quad (3.14)$$

$$\text{ROUGE-L-F-1} = 2 \cdot \frac{\text{ROUGE-L-recall} \cdot \text{ROUGE-L-precision}}{\text{ROUGE-L-recall} + \text{ROUGE-L-precision}} \quad (3.15)$$

Model	**DUC-2004 Evaluation**			**MAG-20 Evaluation**		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
ILPSumm	39.24	11.99	9.34	43.37	16.72	11.26
ParaFuse	**40.07**	**12.04**	**11.28**	46.78	**18.93**	**12.47**
Ours (Proposed Method)	39.58	11.36	9.83	**47.43**	17.28	10.58

Table 3.3: MAG-20 and DUC-2004 Extractive Evaluation

$$\text{Copy Rate} = \frac{|Summary_{tokens} \cap Reference_{tokens}|}{|Summary_{tokens}|} \quad (3.16)$$

Task	Model	Copy Rate
DUC-2004	ILPSumm	0.99
	ParaFuse	0.76
	Our framework	**0.68**
MAG-20	ILPSumm	0.96
	ParaFuse	0.88
	Our framework	**0.72**

Table 3.4: Copy Rate Evaluation. Small copy rates mean more novel words are generated in the final abstractive summaries.

The generated abstractive summaries were evaluated by human evaluators using the following guidelines:

- Our co-author linguists independently reviewed the DUC-2004 and MAG-20 results generated using our approach, ILPSumm, and ParaFuse. Thus, three copies of the same results were shared with the human evaluators. Each of the abstractive summaries generated using the three approaches, for both DUC-2004 and MAG-20, is coupled with the source articles the summaries were synthesized from.

- For each abstractive summary, the linguists read the source articles in the order in which they were listed. While reading the source articles, they kept note of keywords – names of places, people, countries, events, or dates.

- When determining the rating for each criterion, they used the source articles to validate the summary. Then, they used their own compiled summaries to compare to the resulting abstractive summary. The closer the abstractive summary was to the details in their notes, the higher the Entailment.

- Grammar and Coherence did not influence each other in their rating, as grammar is its own separate criterion. Each human evaluator judged Coherence by sentence structure (subject, verb, predicate) and whether the sentences showed logical progression. Thus, they found it easy to differentiate Coherence and Grammar because they were looking past errant punctuation and focusing on the structure of the sentences and the

paragraph as a whole.

- When examining Conciseness, they looked for areas of the abstractive summary that were repeated. They also noted whether the following sentence carried the logical progression of the paragraph, backtracked, or added nothing.

- For Readability, just as with Coherence, our human evaluators did not take grammar or spacing into consideration. They looked for sentence fragments, word order, took note of instances of missing subjects or verbs that were essential to the meaning of the sentence or paragraph as a whole. If the omission or error impacted the overall meaning of the sentence/summary, a lower mark was assigned for Readability.

- When rating Grammar, our human evaluators gave the abstractive summary a lower rating for comma splices or extra spacing than if there were fragments or inappropriate punctuation that made it difficult to determine meaning.

Across 2 human evaluators, we achieve an inter-rater agreement Cohen Kappa score of 68%. In addition to the five human evaluation metrics, we also adopt copy rate [30] for evaluating abstractive summaries. Copy Rate is inversely proportional to the rate of novel word generation in an abstractive summarization task. As shown in Table 3.4, our framework achieves the lowest copy rate indicating that we are able to generate more novel words in the final summaries.

Human Evaluator	Model	Entailment	Coherence	Conciseness	Readability	Grammar
Evaluator-I	ILPSumm	0.60	0.26	0.22	0.20	0.20
	ParaFuse	0.62	0.47	0.55	0.46	0.53
	Ours	**0.66**	**0.52**	**0.63**	**0.50**	**0.60**
Evaluator-II	ILPSumm	0.50	0.38	0.34	0.34	0.40
	ParaFuse	0.64	0.51	0.50	0.45	0.51
	Ours	**0.66**	**0.54**	**0.55**	**0.48**	**0.57**

Table 3.5: DUC-2004 Human evaluation results

Human Evaluator	Model	Entailment	Coherence	Conciseness	Readability	Grammar
Evaluator-I	ILPSumm	**0.89**	0.63	0.71	0.53	0.38
	ParaFuse	0.82	0.64	**0.79**	0.61	0.56
	Ours	0.85	**0.70**	0.77	**0.65**	**0.59**
Evaluator-II	ILPSumm	**0.84**	0.71	0.70	0.65	0.47
	ParaFuse	0.83	**0.79**	0.76	0.68	0.60
	Ours	0.80	0.77	**0.81**	**0.70**	**0.67**

Table 3.6: MAG-20 Human evaluation results

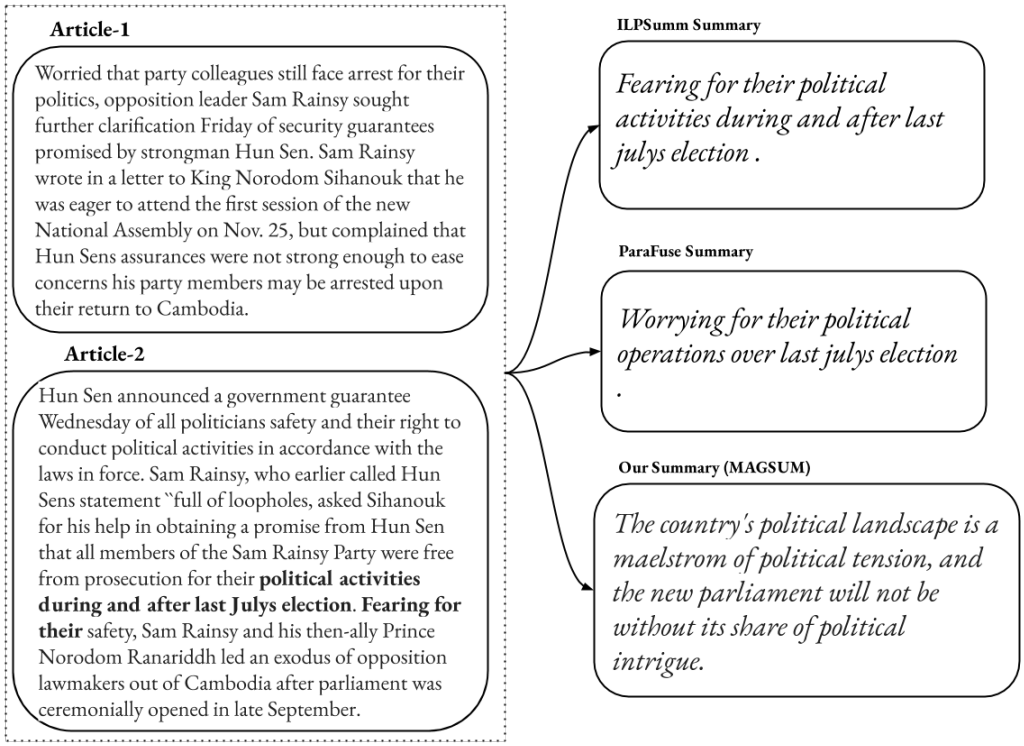

Figure 3.11: Comparison of abstractive summaries.

For DUC-2004, our proposed approach consistently performs better than ILPSumm or ParaFuse on the 5 human evaluation criteria. ILPSumm and ParaFuse show better results in entailment. In contrast, our approach generally performs comparably across the 5 criteria. Thus, we can clearly infer generating summaries that are entailed by source articles is easier than generating summaries that are coherent, concise, readable, and grammatical. This is because if summaries have words copied from the source articles, it is highly likely that they are entailed by the source articles. Since the baseline approaches (ILPSumm, and ParaFuse) have higher copy rate, they do well in entailment. However, with our approach, having a low copy rate and generating summaries that are entailed by the sources articles is difficult; yet, our proposed approach still has the best entailment score for task DUC-2004.

For MAG-20, our approach performs better than the baseline approaches in coherence, conciseness, readability, and grammar across two of our human evaluators, while marginally losing to the baselines according to one of our evaluators. As for entailment, ILPSumm performs the best which is attributed to the high copy rate by ILPSumm. Even though our approach generates significantly more novel words than ILPSumm or ParaFuse, we lose to the best entailment score by only 4%. Further, ILPSumm, ParaFuse, and our proposed approach perform generally better on MAG-20 than on DUC-2004. We surmise this is due to the headline generation task for DUC-2004, while we use author-provided titles for MAG-20.

3.4 Conclusion

We proposed an unsupervised multi-document abstractive summarization framework that, when given a set of documents from MAG, automatically clusters the documents and then generates summaries for each cluster. Our framework consists of extractive and abstractive phases. In the extractive phase, we use coreference resolution to extract groups of inter-dependent sentences from source articles and centroid-based clustering followed by

an enhanced multi-sentence compression algorithm to generate topically informative and relevant summaries. In the abstractive phase, we use text generation technique to generate abstractive language units that are synthesized into an abstractive summary. The number of summaries in our proposed method is adaptively determined based on the semantic analysis of the topics discussed in the documents. We introduce MAG-20, a dataset of topically-clustered groups of scientific articles across 20 Fields of Study and their abstractive summaries. Results show that our proposed approach performs better than state-of-the-art centroid-based summarization techniques on 5 human evaluation metrics and copy rate. In the future, we plan to use additional knowledge and metadata such as citation relationships among scientific articles for document summarization.

Generating Abstractive Summaries for a Scientific Article using Citation Contexts

"Under the right circumstances, groups are remarkably intelligent, and are often smarter than the smartest people in them."

—James Surowiecki, 1967 –

In the previous chapter, we discussed how multiple articles (scientific or news) can extractively or abstractively be summarized by utilizing their intrinsic semantic structure obtained from latent topical analysis. However, the content used for summarization was created by the original authors of the article. In this chapter, we investigate how expert-crowd-sourced knowledge in the form of citation contexts can be used in conjunction with the article to generate abstractive summaries that are comprehensive.

4.1 Why (Motivation)

A key source of knowledge curated by humans in the scientific literature is a summary (presented in the form of a citation context) of reference paper written by researchers while citing the reference paper. Citation contexts pointing to a scientific article can be leveraged to produce a comprehensive summary (overview) of the article that can include additional perspectives and information in the utility of the article. A hybrid summary of a scientific

article that is produced by integrating the abstract of the article (which the original authors produce) and the citation contexts (specified by other researchers) is crucial in bibliometric analysis of scientific literature.

4.2 What (Problem Statement)

In this research aim, we introduce an approach for citation-driven Abstractive Summarization of a scientific article using the abstract of an article and citation contexts (from papers citing the article) pointing to the article. The proposed approach consists of 1) a pipeline to retrieve an appropriate span of a citation context (that consists of citances [54] and sentences surrounding citances) from a citing article using the primary latent topic of a reference article; and 2) a model to fuse the citation contexts and the abstract of a scientific article to generate an abstractive summary of the article. We conducted evaluation of the generated citation-driven abstractive summaries against automated evaluation metrics (lexical and semantic metrics), and human evaluation metrics.

4.3 How (Approach)

4.3.1 Data Curation

For data collection, we focus on three fields of study from the Microsoft Academic Graph (MAG) [[10]: *Artificial Intelligence, Data Mining*, and *Machine Learning*. The reason for focusing on these fields of study is that these have papers with the highest number of citations per cited paper and we can acquire full texts of the citing papers from arXiv repo [1]. Further, we can access a sizeable number of citations for the three fields of study. Since MAG provides titles and abstracts of research papers, we use these to query arXiv for their

[1] https://arxiv.org/help/bulk_data

full texts. As we are interested in authoritative papers, we consider the top-ranked citing papers corresponding to each cited paper when querying arXiv. For this, we use the Rank [2] metadata from MAG to determine the highest ranked citing papers for each of the cited papers. The cited and citing papers used in this study are papers published in the years 2005 - 2020. To minimize noise, we focus on citances that appear in one of the following sections of a citing paper: 1) Introduction; 2) Related Work; or 3) Background. The reason for focusing on these sections is that authors of scientific papers normally summarize other papers in one of these sections while they focus on their own work in other sections. We report in Table 4.1 the statistics of MAGSumm-3000, a dataset we prepare for Research Aim-II. We use PDFMiner [3] and a LaTeX parser [4] to map each citance to its parent text and section in a full text citing article. Then, we extract citances that appear in Introduction, Related Work, or Background. We also note that a single citance may point to multiple cited papers, as reflected in Table 4.1 under column *# citation linkages*, and thus, the same citance is used for summarizing different reference papers if the citance has multiple citations.

Field of Study	# Reference Papers	# Citing Papers	# Unique Citances	# Citation Linkages
Artificial Intelligence	1000	11837	14765	17392
Data Mining	1000	9264	12583	13638
Machine Learning	1000	9736	12375	14739
TOTAL	3000	30837	39723	45769

Table 4.1: Dataset sizes for the three fields of study

We refer to the scientific article to be summarized as *the reference paper* (RP), and the article citing the RP as *the Citing Paper* (CP). Also, we call the primary topic of the combined abstract and introduction of the reference paper as *RP topic*. A *citance* [[54] is a sentence in a citing paper with explicit citation to the reference paper. Context sentences are sentences surrounding the citance. A *citation context* is a contiguous span of citance and zero or more context sentences in the citing paper.

[2] https://bit.ly/3bqSrPp
[3] https://pypi.org/project/pdfminer/
[4] https://github.com/alvinwan/TexSoup

Figure 4.1 shows the data curation pipeline for a reference paper in a field of study (FoS).

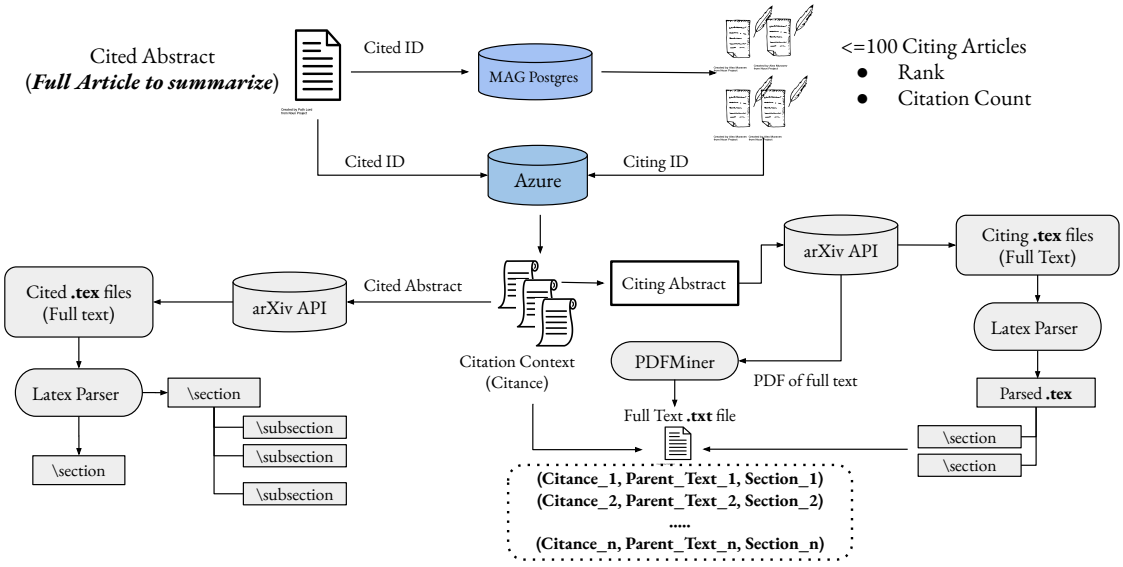

Figure 4.1: Data Preparation Pipeline. Querying for citation contexts corresponding to a reference (cited) paper. Data are stored in distributed system consisting of Postgres DB, and Microsoft Azure.

4.3.2 Proposed Framework

In this section, we discuss the components of the proposed framework.

In this research aim, we propose *TransFuse*, a framework built by amalgamating a Transformer-based encoder-decoder model and a sentence Fusion pipeline and apply it to abstractive summarization of reference papers using citation contexts. Our proposed framework, which consists of two stages of abstraction, is inspired by the limitations of encoder-decoder models which degenerate and produce bland and repetitive text, a phenomenon known as neural text degeneration. Figure 4.2 shows our proposed framework (TransFuse) along with TaCC retrieval module. The sub-sections below discuss the com-

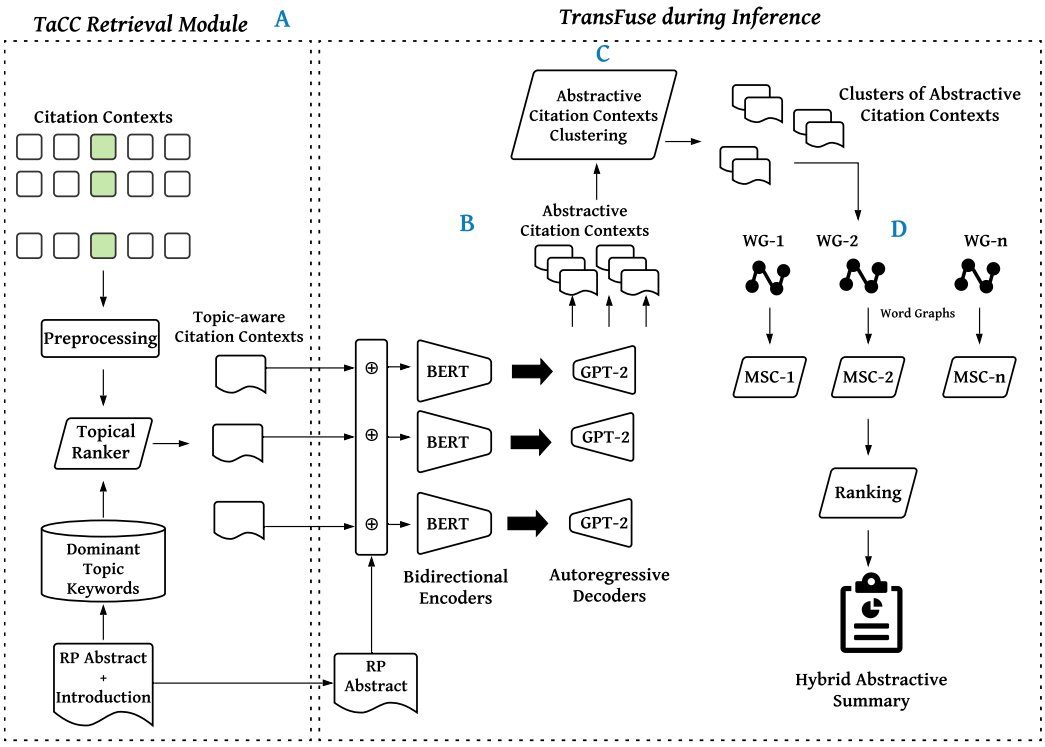

Figure 4.2: TaCC Retriever with TransFuse.

ponents (labeled A, B, C, D) in Figure 4.2.

Topic-aware Citation Context Retrieval In addition to citances, we consider context sentences surrounding citances based on how strongly a contiguous span of the context sentences and the citance reflects RP topic. For this, we perform topic-aware convolution over a sequence of explicit and implicit citing sentences in citing papers. Thus, for each field of study (FoS), we build a topic model using Latent Dirichlet Allocation (LDA) [[92] for a corpus composed of the abstract and introduction of all reference papers for the same FoS and produce the most dominant topic discussed by the RP abstract and introduction. Once we identify the most dominant topic and its constituent keywords for an RP abstract+introduction, we perform topic-aware citation context (TaCC) retrieval using a convolving kernel whose window size is within blocks of 0, ± 1, and ± 2 units (i.e., in terms of sentences spanned) from a citance. A moving kernel within a block of 0 unit

from the citance represents a kernel that spans only the citance as a citation context, while a moving kernel within a block of ± 2 units from the citance slides over three candidate citation contexts: 1) with the citance as the rightmost sentence preceded by two context sentences; 2) with the citance in the middle and a single context sentence on either side of the citance; and 3) with the citance as the leftmost sentence followed by two context sentences. The selection of a max block size of ± 2 from a citance is based on the arguments provided in [102]. For constructing the convolving kernel, we embed each keyword in the RP topic using SciBERT [94], and the sentences in a citation context to be convolved over using sentence-BERT [100]. We then perform ranking of citation contexts based on how strongly they reflect the RP topic and retrieve the citation context that scores the highest in cosine similarity across all the keywords as the TaCC for a citance. This operation is akin to extracting feature maps in convolutional neural networks [103]. The SciBERT model we use for embedding the RP topic keywords is the scivocab, uncased model. Similarly, the sentence-BERT model we use is the base model with the same hyperparameters as the SciBERT model. We show the steps to retrieve TaCC in algorithm 2. CC represents a Citation Context span whose topical rank is being computed. For a sequence of sentences that is centered around the citance with two units to the leftmost sentence and two units to the rightmost sentence, the index of the citance is at position pos=2. For the first iteration of a given kernel size w, the sliding kernel's $start$ index points at a sentence that is located w units to the left while its end index points at the location of the citance. Similarly, for the last iteration of kernel size w, $start$ index points at the citance while index end points at the context sentence w units to the right of the citance. In the algorithm, $\vec{v}(k)$ is a SciBERT embedding of a keyword in the RP topic, and $\vec{v}(CC_{[start,end]})$ is a sentence-BERT representation of a citation context.

Abstractive Citation Context Generation Once we identify the topic-aware citation context for each citance in a CP, we pass each TaCC into the first component of TransFuse

Algorithm 2: TaCC retrieval algorithm

Initialize $score_{max} \leftarrow 0.0$
Initialize $W \leftarrow 2$
Initialize $pos \leftarrow 2$
T - set of keywords in the RP topic
$\vec{v}(\cdot)$ - Embedded representation of a sequence or keyword
Function ***retrieve*** *TaCC*
  for $w = 0$ **to** W **do**
    for $offset = 0$ **to** w **do**
      $start \leftarrow pos - w + offset$
      $end \leftarrow pos + offset$
      $score_{CC_{[start,end]}} \leftarrow \frac{1}{|T|} \sum_{k \in T}^{|T|} cossim(\vec{v}(k), \vec{v}(CC_{[start,end]}))$
      if $score_{CC_{[start,end]}} > score_{max}$ **then**
        $score_{max} \leftarrow score_{CC_{[start,end]}}$
        $TaCC \leftarrow CC_{[start,end]}$
      end
    end
  end
end

where we perform abstraction of the topic-aware citation contexts. The rationale for independently abstracting the topic-aware citation contexts by generating multiple abstractive citation contexts is to minimize the generation of repetitive phrases which is one of the main problems with neural text generation models, especially with beam search decoding, since they are mainly optimized on Maximum Likelihood Estimation as pointed out in recent studies by [104] and [105]. Motivated by the work of [58], we experiment with generating Abstractive Citation Contexts (ACCs) using two input settings: 1) topic-aware citation contexts only; 2) topic-aware citation contexts and RP abstract. For the second (i.e., hybrid) configuration, we concatenate the RP abstract and all incoming TaCC, embed them, and feed their combination into an autoregressive decoder. We use special token **[SEP]** to separate RP abstract and TaCC in the hybrid setting. The use of **[SEP]** as a separator token is similar to that in [106]. We did not explore independently encoding each component of the input, combine the encodings and feed them into the autoregressive decoder. Our inspiration was based on [106] where [106] experimented with both approaches and

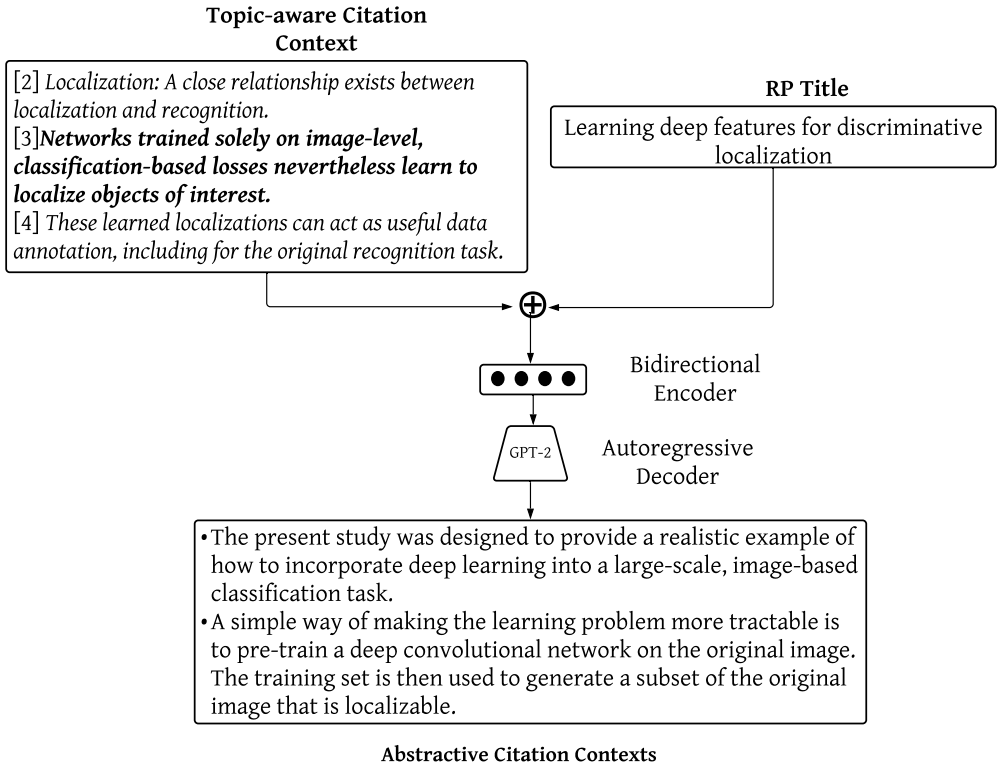

Figure 4.3: Sample Abstractive Citation Contexts.

found that the first approach gives better results. Since a topic-aware citation context could span a single sentence, or multiple sentences, we perform dimensionality reduction using t-SNE to 768 hidden units when TaCC is composed of a citance and context sentences in our experiment. The Transformer encoder-decoder part of TransFuse is fine-tuned on the arXiv dataset as the baseline models are fine-tuned on (details in section on the Domain and Task-specific fine-tuning). We use GPT-2 [52] with trigram blocking as our auto-regressive decoder. We set the hyperparameters of the decoder as: number of samples generated to 5, maximum length of a sample to 50, batch size to 5, and temperature to 0.7 to generate more coherent abstractive citation context. Abstractive citation context generation is performed in parallel across the Topic-aware Citation Contexts. Figure 4.3 shows sample abstractive citation contexts generated given a TaCC and an RP title.

Abstractive Citation Contexts Clustering The Transformer component of TransFuse generates multiple ACCs for each reference paper since each reference paper has multiple incoming citances, and hence multiple TaCCs. Further, for each TaCC, there are five corresponding Abstractive Citation Contexts (ACCs) generated. A certain number of these ACCs generated from different TaCCs are semantically similar or redundant. Therefore, we propose clustering ACCs to identify similar ACCs and separate them from others. We use K-means clustering where we dynamically determine the optimal number of clusters K by maximizing Calinski-Harabasz score. We use Calinski-Harabasz index because it is good at identifying dense and well separated clusters [5]. We find the same optimal number of clusters when we use Silhouette Coefficient [6]. We experimented with agglomerative clustering and obtained similar results.

Multi-Sentence Compression Each cluster of ACCs generated consists of ACCs with significant semantic overlap. Thus, we propose to project the sentences in each ACC onto a word graph and select paths from the word graph based on their thematic centrality and authoritative score of the ACCs the paths span. The intuition is that the path selected should be central to the theme of ACCs in a cluster and that a path that spans ACCs that are associated with higher authoritative sources [107] are ranked higher. We embed each ACC in a cluster using sentence-BERT followed by dimensionality reduction using t-SNE and then we compute the average of all ACC embeddings. We then traverse each word graph and pick paths that span at least two sentences in different ACCs and embed each path using sentence-BERT. Once each path is embedded using sentence-BERT, we compute its cosine similarity with the average of the ACC embeddings in a cluster and sort all the paths in non-ascending order of their cosine similarity score. Paths that are closest to the central theme of the ACCs in a cluster are ranked higher. The intuition is that we want to maximize information coverage or representativeness of paths per cluster. This is done for each cluster of

[5] https://bit.ly/3au3y7Q
[6] https://bit.ly/2MjMWIa

ACCs. Once candidate paths from each cluster are sorted based on their combined thematic centrality and authoritative score, we select paths into the final summary. We propose to have more informative sentences at the beginning and less informative sentences towards the end. Moreover, we assume each cluster is equally significant for the summary generation and thus, we sort paths based on their representativeness in their respective cluster to form the final summary. The purpose of the selection of paths from different clusters is to achieve better diversity in the final summary. There are as many word graphs as there are clusters of ACCs, and we want a certain number of paths selected from each word graph to meet our 250-word summary target. The 250-word summary target is determined by following the same guide as the CL-SciSumm shared tasks. To determine the average number of paths selected per word graph, we first estimate the average length of a path (which is essentially the same as the average length of an English sentence in number of words), the number of clusters of ACCs, and the expected number of words in the final summary. We assume the average number of words in an English sentence to be 20 as pointed out in [108]. Thus, given the 250-word summary target, the average number of words in an English sentence as 20, and the number of word graphs K, we compute the average number of paths per word graph as in Equation 4.1. The purpose is to have nearly equal number of paths from each cluster of ACCs, so one cluster does not dominate others.

$$nPaths_{wg} = \frac{250}{K * 20} \tag{4.1}$$

where K - number of clusters of ACCs

For the purpose of maximizing diversity, the path selection technique selects paths into the final summary so long as their semantic similarity with an already selected path is no more than 0.8 [29] on cosine similarity. If a path selected from the word graph is semantically similar by an order of 0.8 or more to an already selected path, and has a higher authoritative centrality score than the path in the summary, we replace the already

selected path with the new path, at the same location in the summary. This ensures that 1) more informative paths are at the beginning of the final summary; and 2) multiple runs of the path selection algorithm generate the same summary. Thus, the path selection algorithm optimizes between authoritative centrality of a path within a cluster and similarity with an already selected path. This approach is motivated by Maximal Marginal Relevance (MMR) as proposed in [101] and its variant is presented in chapter 3 algorithm 1. Further, a path is selected if its length is at least 8 words [24] and at most 25 words. For computing the authoritative centrality of a path, we first compute the normalized Ranks of the citing papers associated with the ACCs the path spans. Equation 4.2 computes the normalized Rank of a citing paper, which each ACC inherits from the CP. Since Ranks in MAG are sorted in ascending order with lower numbers associated with higher authoritativeness, we normalize the rank of each citing paper with respect to the citing papers whose ACCs are clustered together, and subtract from 1 as shown in Equation 4.2.

$$nRank(CP_i) = 1 - \frac{Rank_{CP_i}}{max\{Rank_{cp_r} : r = 1 \cdots N\}} \tag{4.2}$$

where N - # unique CPs associated with ACCs.

Each ACC associated with a CP is assigned the same normalized rank computed for the CP. Let the set of abstractive citation contexts $C = \{ACC_1, ACC_2, ACC_3, \cdots, ACC_M\}$ be abstractive citation contexts spanned by path P. We formulate the Rank of P as in Equation 4.3.

$$Rank(P) = \frac{1}{M} \sum_{i=1}^{M} nRank(ACC_i) \tag{4.3}$$

where M - number of ACCs spanned by path P

We compute Authoritative Centrality of a path from a word graph using Equation 4.4 which optimizes the average normalized Rank of a path, which captures its authoritative-

ness [107] and the proximity of the path to the mean of the embeddings of the ACCs in the cluster. The intuition for using the mean of the embeddings of entities in a cluster to decide a representative path based on proximity is similar to the approaches used in [39, 109] albeit for different tasks.

Let all the ACCs that are spanned by path P in a cluster be C. We compute Authoritative Centrality, AC of path P as in Equation 4.4.

$$AC(P) = \alpha \cdot Rank(P) + (1 - \alpha) \cdot cossim(Embed(P), Embed_{\text{avg}}(C)) \tag{4.4}$$

The first term represents the average Rank of the ACCs a path spans while the second term reflects the thematic centrality of the path. α varies from 0.1 to 0.9 in increments of 0.1 for each path under consideration and we pick the value of α that maximizes the authoritative centrality of the path.

Figure 4.4 shows sample word graph constructed given two ACCs in a cluster. Further, to preserve syntax, every token from the ACC is appended with its Part of Speech (PoS) tag when projected onto the word graph and two tokens share a node if their lowercase form and POS tag are the same.

ACC_1 = "The present study was designed to provide a realistic example of how to incorporate deep learning into a large-scale, image-based classification task."

ACC_2 = "A simple way of making the learning problem more tractable is to pre-train a deep convolutional network on the original image."

4.3.3 Experiments and Results

Domain- and Task-specific Fine Tuning Since our task is long document scientific article summarization, we fine tune all the pre-trained baseline models and the transformer

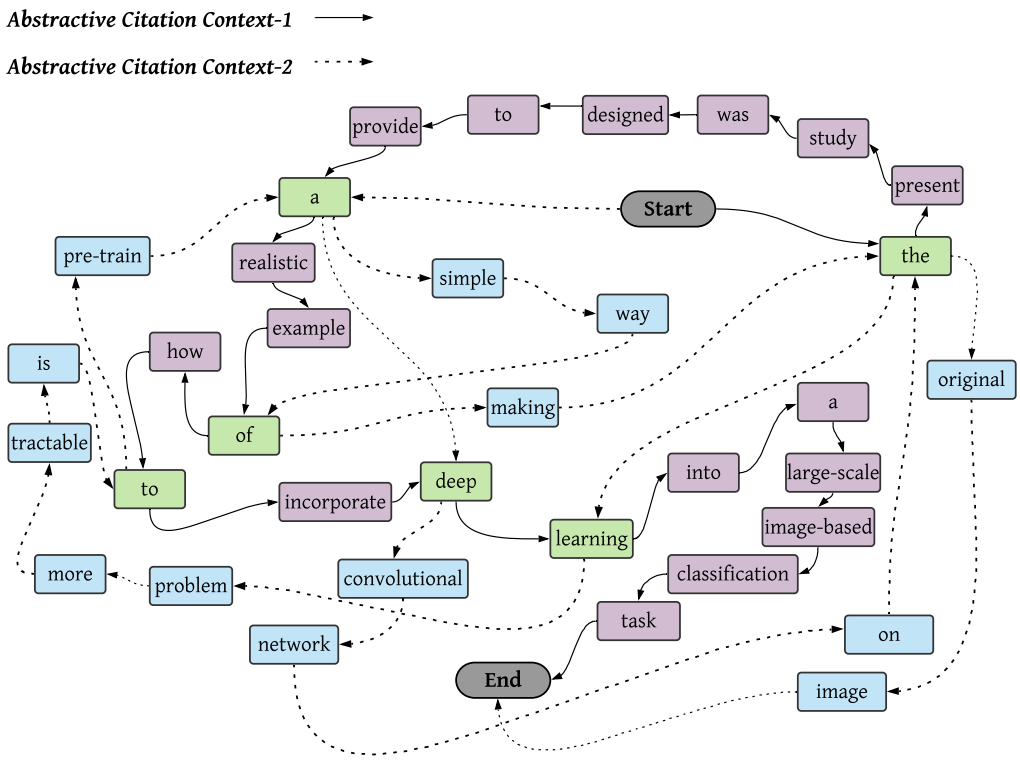

Figure 4.4: Nodes exclusively from ACC_1 are marked with a purple color and nodes exclusively from ACC_2 are marked with blue color. Common nodes between ACC_1 and ACC_2 are marked with green color.

part of TransFuse on the arXiv dataset [77] since this dataset is on scientific domain and the task is also abstractive summarization. We train each baseline model and the encoder-decoder component of TransFuse for 5 epochs with the following hyperparameter setting: learning rate of 5e-5, batch size of 16, optimization using Adam optimizer [110] and a drop out rate of 10%. For each model we achieve an average loss within the range 2.0 and 2.3 on the hold-out test set. The size of train/validation/test sets are such that there are 50,000 training instances, 6,000 validation instances and 6,000 testing instances. Since transformer-based models are designed to work with a maximum of 512 tokens of input sequence, for each baseline model, and for the transformer component of TransFuse, we split an input sequence into smaller chunks of 512 tokens each and encode each chunk followed by concatenation and then a linear transformation is applied to reduce the dimension back to 768 units. This approach of reducing long input sequence in transformer models

using chunking is explored in a recent study by [78]. Maximum decoded sequence length is set to 210 tokens as used in [77]. Similar to the way the baseline models are trained, the Transformer part of TransFuse is trained in such a way that the encoder part is initialized with BERT [50] where an input sequence is split into chunks of a maximum of 512 tokens, followed by separately encoding each chunk and then concatenating the encodings. The concatenated encoding is then passed to a feed forward layer with output dimension of 768. The decoder is initialized with GPT-2 [52]. During inference, a group of TaCCs (w/ or wo/ RP abstract) is passed to a trained baseline model in chunks of 512 tokens, while for TransFuse, each TaCC (w/ or wo/ RP abstract) is passed to the Transformer component to generate multiple ACCs as discussed in the Methods section. All baseline models and the transformer component of TransFuse are built and trained using PyTorch on NVIDIA Tesla T4 GPU.

Baseline Approaches

- *Text-to-Text Transfer Transformer (T5)* [64] is a unified framework that treats a text processing problem as a text-to-text problem. It is pre-trained on the "Colossal Clean Crawled Corpus" (C4) corpus, a 750GB corpus crawled from the Web. The pre-training is done using fill-in-the-blank style denoising objective and fine-tuning is performed for tasks including abstractive summarization.

- *BART* [66] is a denoising autoencoder consisting of a bidirectional encoder and a left-to-right autoregressive decoder (GPT) to generate an abstractive summary. The pre-training has two stages: 1) corruption of text with an arbitrary noising function; and 2) a sequence-to-sequence model trained to reconstruct the original document.

- *Pegasus* [67] is a transformer-based model that proposes a pre-training objective to mask a number of tokens and important sentences in an input document and learn to generate the important sentences from the context of the remaining sentences. It is trained based on Gap Sentence Generation and Masked Language Modeling.

For fair comparison with the SOTA, the transformer component of TransFuse is fine-tuned on the arXiv dataset as the baseline models are. We use the base variant of each baseline model and also of the building blocks in our framework. A similar approach was taken in a study by [111] while comparing different models for abstractive summarization.

Experiments and Results We design our experiments using the two input configurations. We use Hugging Face [7] implementation of models T5, BART, and Pegasus to fine-tune on arXiv dataset and generate summaries with maximum length of 250 words for the different inputs at test time. Tables 4.2 and 4.3 show our results on the benchmark SciSummNet-1000 dataset while Table 4.4 shows results on MAGSumm-3000 dataset. As can be seen, we do not evaluate novelty w.r.t ground truth summaries. Formally, we define Novelty as in Equation 4.5.

$$Novelty_{TaCC} = \frac{||\text{summary}_{ngrams} - \text{TaCC}_{ngrams}||}{||\text{summary}_{ngrams}||} * 100 \tag{4.5}$$

where $ngrams$ - set of unigrams, bigrams, and trigrams.

Due to the limitations of lexical based metrics to text generation tasks such as abstractive summarization, there has been a significant effort to quantify the semantic equivalence between a generated summary and a ground truth summary as extensively studied in a recent work by [2]. In our study, for computing semantic equivalence between a summary and a group of TaCCs, we experiment with a pretrained model for encoding the generated summaries, the ground truth summaries, and the citation contexts. We use SPECTER [106] which is based on a transformer model [36] and specifically pretrained on scientific documents by using citation networks as an additional feature. After the generated summary, a ground truth summary, and the group of citation contexts are embedded using

[7] https://huggingface.co/transformers/

SPECTER, we compute cosine similarity between the generated summary and the group of TaCCs and the generated summary and human-written summaries (for SciSummNet-1000). We also use an independent evaluation metric *Diversity*, which is less explored in document summarization. We posit that an abstractive summary should be diverse across its constituent sentences, at which beam search based decoding techniques [112, 105] do not perform well. For this, we leverage a variant of a self-BLEU metric as proposed in [113]. Specifically, we measure Diversity by computing the average pairwise cosine similarity among sentences in a summary, normalizing them and subtracting the result from 1. Tables 4.2, 4.3, and 4.4 show experimental results using different input configurations, and metrics. While ROUGE constitutes lexical evaluation, Novelty, and Semantic Equivalence represent abstractive metrics. Since our objective is not to maximize novelty of generated summaries with respect to ground truth summaries for SciSummNet-1000, we do not report novelty w.r.t the ground truth summaries in the experimental results.

Equation 4.6 formalizes Diversity.

$$\text{Diversity} = [1 - \frac{2}{|S| \cdot (|S| - 1)} \sum_{i=1}^{|S|-1} \sum_{j=i+1}^{|S|} \text{cossim}(\vec{v}(i), \vec{v}(j))] \cdot 100 \tag{4.6}$$

where S - set of sentences in a summary

$\vec{v}(p)$ - Sentence BERT embedding of sentence at location p

Model	ROUGE-1-F		ROUGE-2-F		ROUGE-L-F		Semantic Equivalence	
	CC	CC/w Abstract	CC	CC/w Abstract	CC	CC/w Abstract	CC	CC/w Abstract
T5	23.80	46.70	8.43	37.22	22.20	47.97	31.26	25.37
BART	**28.55**	**50.99**	**10.34**	**38.03**	**26.12**	**49.9**	35.64	26.59
Pegasus	13.38	32.40	4.21	22.33	12.42	31.71	24.76	21.73
TransFuse	26.50	48.06	7.66	35.61	21.61	48.32	**37.58**	**28.38**

Table 4.2: SciSummNet-1000 evaluation w.r.t. human summaries.

We report the ROUGE-F measure which is the harmonic mean of ROUGE precision and ROUGE recall. While ROUGE is based on n-gram overlap, Novelty metric measures

Model	ROUGE-1-F		ROUGE-2-F		ROUGE-L-F		Novelty (N-grams)		Semantic Equivalence	
	CC	CC/w Abstract	CC	CC/w Abstract	CC	CC/w Abstract	CC	CC/w Abstract	CC	CC/w Abstract
T5	32.05	22.71	21.56	9.1	31.61	21.5	28.85	43.41	34.21	31.85
BART	**39.38**	**29.40**	**25.98**	**14.37**	**38.53**	**27.86**	42.75	45.58	37.81	27.84
Pegasus	13.69	13.5	7.31	3.98	13.28	12.55	18.74	42.23	29.18	25.69
TransFuse	32.01	18.07	15.4	5.83	30.32	17.11	**44.78**	**47.79**	**38.62**	**32.74**

Table 4.3: SciSummNet-1000 evaluation w.r.t. citation contexts.

Model	ROUGE-1-F		ROUGE-2-F		ROUGE-L-F		Novelty (N-grams)		Semantic Equivalence	
	CC	CC/w Abstract	CC	CC/w Abstract	CC	CC/w Abstract	CC	CC/w Abstract	CC	CC/w Abstract
T5	27.99	18.04	**20.28**	6.42	27.68	16.83	30.28	56.89	23.91	18.35
BART	**29.93**	**21.57**	18.75	**8.97**	**29.11**	**20.12**	38.87	58.82	27.38	22.47
Pegasus	12.06	11.73	6.56	2.72	11.71	10.71	26.62	52.92	18.36	15.73
TransFuse	24.86	16.37	18.36	6.03	25.37	13.84	**42.27**	**59.21**	**31.68**	**25.37**

Table 4.4: MAGSumm-3000 evaluation wrt topic-aware citation contexts.

Model	Diversity	
	TaCC	TaCC/w Abstract
T5	0.29	0.33
BART	**0.62**	**0.42**
Pegasus	0.53	0.39
TransFuse (Ours)	0.58	**0.42**

Table 4.5: Diversity with different input configuration.

novel n-gram generation. The limitation of ROUGE metrics for evaluating abstractive summaries is highlighted in previous studies [35, 114, 40]. With regards to using a hybrid input configuration, we generally see lexical measures improve with the inclusion of RP abstract. In addition to the automatic evaluation, we also perform human evaluation of 20 randomly selected reference papers' summaries generated using the baseline models and our approach. The human evaluation was conducted by our three linguists who are native English speakers. The three linguists evaluated the summaries on two criteria: 1) Readability and 2) Coherence, each on a scale of 1 to 5. On Readability, the average scores are 2.31, 3.57, 2.68, and 2.61 for T5, BART, Pegasus, and TransFuse respectively. On coherence, the average scores are 2.17, 3.26, 2.21, and 2.11 for T5, BART, Pegasus, and TransFuse respectively. We compute inter-annotator agreement among the human evaluators and achieve a substantial kappa score [115] of 0.73.

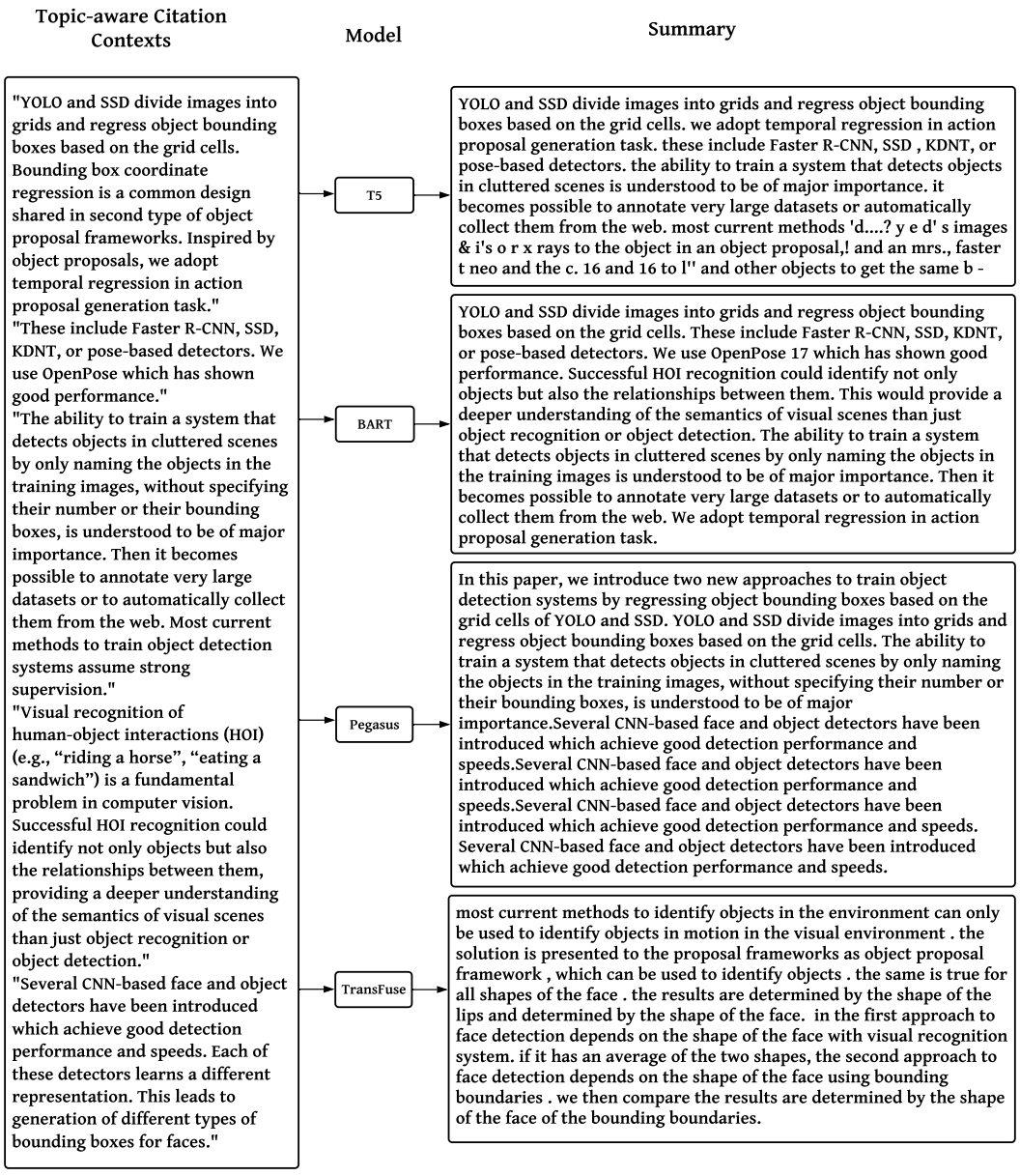

Figure 4.5: Comparison of abstractive summaries generated using the baseline models and TransFuse.

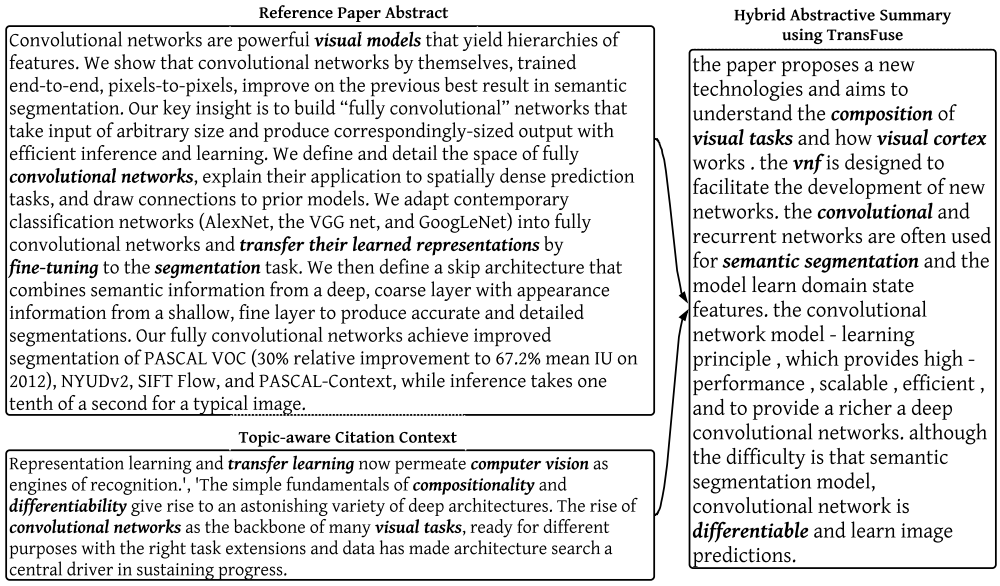

Figure 4.6: Sample Hybrid Abstractive Summary synthesized from the abstract of the reference paper and a topic-aware citation context.

4.4 Conclusion

In this chapter, we introduced *TransFuse*, a framework built by inserting a sentence Fusion pipeline on top of a Transformer-based model for abstractive summarization and applied to the task of scientific article summarization using citation contexts. Using experiments, we showed TransFuse outperforms in abstraction of generated summaries, and is comparable to strong abstractive approaches in ROUGE metrics and human evaluation. Figure 4.5 shows comparision of TransFuse with baseline models in terms of the qualities of the summaries generated while Figure 4.6 shows a sample hybird abstractive summary generated using TransFuse when the abstract of a reference paper and a topic-aware citation context are fed as input to the model during inference.

The key takeaway from this chapter, in terms of technical contribution, is that while the SOTA transformer encoder decoder models perform reasonably good enough on lexical metrics, one of their major limitations continues to be generating repetitive, non-sensical text (a result of what is known as neural text degeneration). TransFuse enables to generate

less repetitive, more novel words (i.e., paraphrasing source text), while preserving semantics. With respect to the problem the proposed model is applied to, abstractive merging of citation contexts (expert-crowd-sourced) and abstract of a scientific article (author-sourced) leads to generating an abstractive summary that is more comprehensive. On the other hand, we still believe there is a lot of room for improvement for this task in terms of coherence and readability of generated summaries.

Entity-driven Fact-aware Abstractive Summarization of Biomedical Literature

"We know very little, and yet it is astonishing that we know so much, and still more astonishing that so little knowledge can give us so much power."

—Bertrand Russell, 1872 – 1970

We explore how expert-curated knowledge bases can be used to guide abstractive summarization. While the previous chapter focused on what authoritative sources have to say about an article, which may or may not be factual, in this chapter, we leverage knowledge bases containing facts about named entities in an article to guide abstractive summarization and apply to the task of biomedical article summarization. The proposed approach is evaluated on entity-level factual accuracy and semantic equivalence metrics.

5.1 Why (Motivation)

There are over 2 million biomedical articles published and available on PubMed. These plethora of biomedical articles are related to one another in terms of their named entities and the relationships between the named entities (semantics). Named entities in biomedical literature can be leveraged to cluster multiple biomedical articles based on their entity-level relatedness and use the named entities as part of modeling abstractive summarization of

biomedical literature. The semantics of named entities can be made explicit and enhanced by using expert-curated domain-specific background knowledge bases. Concretely, named entities in biomedical literature can be used to retrieve relevant facts (i.e., facts related to the article the named entities appear in) from knowledge bases. Leveraging named entities and expert-curated facts in knowledge bases can improve abstractive summarization in terms of factual accuracy and semantic equivalence which are metrics not captured by widely used lexical metrics such as ROUGE, BLEU, and METEOR.

To this end, the World Health Organization (WHO) introduced the International Classification of Diseases (ICD), a catalog for the systematic study of human diseases. The ICD is organized into different chapters in such a way that diseases that share certain characteristics belong to a common ICD chapter. There have been different editions of ICD, the most recent one being the ICD-11. On the other hand, part of the class of biomedical literature being published by the biomedical community constitutes articles about human diseases such as the articles curated for the publicly available NCBI disease [116] and BC5CDR [117] disease corpora. While the NCBI-disease and BC5CDR corpora have extensively been used for NLP tasks including named entity recognition and relation extraction, there are no large scale equivalent dataset on diseases for biomedical articles summarization. With several biomedical articles about related or common diseases being published, it is important that biomedical articles with common disease mentions be grouped and a summary of the grouped articles be generated to help a biomedical researcher learn more about published works about related diseases. Consequently, we leverage the ICD-11 classification of diseases to query PubMed for abstracts (specifically, abstracts with disease mentions), cluster the abstracts into their corresponding ICD-11 chapter and guide the summary generation based on named entities about the diseases. Table 5.1 shows the ICD-11 Special Groups Catalog [1].

[1] https://icdcdn.who.int/icd11referenceguide/en/html/index.html#icd-chapter-structure

ICD Chapter	Chapter Title
1	Certain infectious or parasitic diseases
2	Neoplasms
3	Diseases of the blood or blood-forming organs
4	Diseases of the immune system
18	Pregnancy, childbirth, or the puerperium
19	Certain conditions originating in the perinatal period
20	Developmental anomalies
22	Injury, poisoning or certain other consequences of external cause

Table 5.1: ICD-11 special groups chapters and corresponding titles.

5.2 What (Problem Statement)

In this chapter, we devise a unified end-to-end model for Entity-driven Abstractive Summarization of biomedical articles where named entities and facts retrieved from biomedical knowledge bases are used for guiding abstractive summarization. The proposed approach is composed of two stages: 1) Entity-driven knowledge retriever, and 2) Knowledge-guided abstractive summarizer trained end-to-end. We evaluated the proposed approach against the state-of-the-art abstractive summarization models on semantic and lexical metrics.

5.3 How (Approach)

5.3.1 Data Curation

We use two datasets for this task: 1) 60,000 randomly selected article-to-summary pairs from benchmark PubMed dataset [77] which we refer to as PubMed-50k and use the 50,000 samples to train our models, and the remaining 10,000 for inference, and 2) ICD-11-Summ-1000 which we curate and use to test the trained baseline models and our model. Curation of ICD-11-Summ-1000 follows a data preparation pipeline that consists of: 1) ICD-11 disease lexicon (for each ICD-11 chapter) curation for querying PubMed for abstracts cor-

responding to each ICD-11 chapter; and 2) Entity-aware pseudo-document generation for a collection of semantically related PubMed abstracts collected using the keywords in the lexicon built for an ICD-11 chapter. Thus, for querying PubMed for abstracts for an ICD-11 chapter, we first query the biomedical knowledge bases for *"disease"* keywords using the names of each ICD-11 chapter and build a lexicon of diseases corresponding to each chapter. Figure 5.1 shows what a lexicon build-up for an ICD-11 chapter looks like. Once the disease related keywords are identified for an ICD-11 chapter, we use these keywords to query PubMed via the Bio Entrez parser [2] to capture the first 1000 abstracts (PMIDs) spanning a period of last 90 days from the moment we initiated the query. We do this for each of the eight *special groups* ICD-11 chapters.

[2] https://biopython.org/docs/1.75/api/Bio.Entrez.html

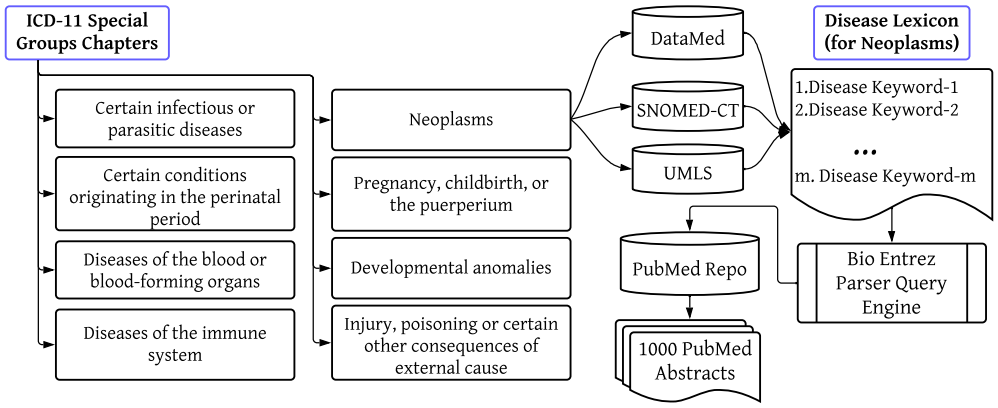

Figure 5.1: ICD-11 based lexicon construction and querying for abstracts from PubMed using Bio Entrez parser. For illustration purpose, we show the pipeline for ICD-11 chapter 2 (i.e., *Neoplasms*)

Figure 5.2 shows our ICD-11-Summ-1000 dataset preparation pipeline. Once we have obtained the 1000 PubMed abstracts for an ICD-11 chapter through the querying process, we conduct named entity recognition (NER) on each of the abstracts within a chapter using the SciSpacy NER model trained using the BC5CDR corpus [118]. Since we are interested in entity-level clustering of PubMed abstracts within an ICD-11 chapter, we first conduct clustering of the named entities using agglomerative clustering as used in [69]. We use BioBERT [119] for named entity representation followed by agglomerative clustering. Once the named entities pertaining to an ICD-11 chapter are clustered into different bins, our next task is to cluster the PubMed abstracts into a bin based on how the named entities within the abstracts are related to the entities within a cluster. We use cosine similarity between named entities identified in a PubMed abstract and entities characterizing a cluster to determine an entity-aware cluster the abstract belongs to. Next, for each cluster, we perform named entity-aware salient content selection to produce an extractive pseudo-document for each cluster. This paradigm of reducing a multi-document corpus (i.e., a cluster consisting of PubMed abstracts grouped based on entity-relatedness) into an extractive pseudo-document is explored for different tasks in other studies [120, 121, 122]. As part of the NER task, we use coreference resolution [99, 49] after named entities are extracted using SciSpacy to cluster the biomedical named entities and their coreferenced

mentions spanning the multiple abstracts within an ICD-11 chapter.

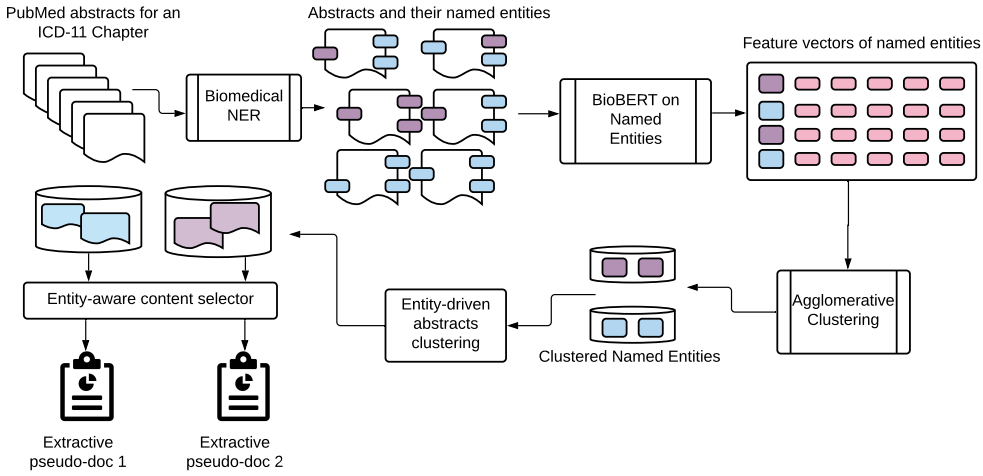

Figure 5.2: Entity-aware content selection to produce extractive pseudo-documents. The light blue and light lavender colored documents in the final bins represent abstracts whose named entities are semantically similar to one another.

During entity-aware content selection to produce an extractive pseudo-doc for a cluster of PubMed abstracts that are clustered based on named entity relatedness, we preserve the positioning of sentences within an abstract. We also use the following heuristics while constructing the extractive pseudo-doc: 1) a sentence shall have at least one named entity identified using SciSpacy-BC5CDR NER model; and 2) the selected sentences from an abstract are placed in the same order as they appear in the abstract. Further, we also take into account abstracts' relative importance scores where abstracts with higher document importance scores [48] have their sentences precede sentences from abstracts with less document importance scores while generating the extractive pseudo-document. Document importance D_{imp} of target abstract d_i is determined using pairwise cosine similarity between the BioBERT [119] embedding of d_i and other abstracts within the same cluster $\mathcal{C}$. Formally, document importance is defined as in Equation 5.1

$$D_{imp} = \frac{\sum_{d_i, d_j \in \mathcal{C}} cossim(d_i, d_j)}{|\mathcal{C}| - 1}, (i \neq j) \tag{5.1}$$

For all tasks throughout this chapter involving initializing of networks or for representation learning, we use BioBERT [119].

5.3.2 Proposed Framework

Our proposed approach is a two-stage framework consisting of 1) an entity-driven knowledge retriever, and 2) a knowledge-guided abstractive summarizer. In this section, we discuss both modules in detail.

Entity-driven Knowledge Retriever For each extractive pseudo-document generated in the data preparation stage for ICD-11 or input article for PubMed-50k designated by $\mathcal{D}$, we identify the named entities in the input document. The identified named entities are then used to retrieve facts from biomedical knowledge bases (UMLS, ICD-10, and SNOMED-CT). We use PyMedTermino [123] to work with the entire dump of UMLS [85] available at [3]. For m named entities (and their coreferenced mentions), we have a set of pairs of entities $\{(e_i, e_j) \mid 0 \leq i < j < m\}$ extracted from $\mathcal{D}$, where each pair (e_i, e_j) is used to query for c candidate facts $F_1, F_2, F_3, \ldots, F_{|c|}$ denoted collectively by $F_{\mathcal{D}}^{i,j}$ from the background knowledge bases $\mathcal{K}$ using full text search. The complete set of facts retrieved for all pairs of named entities in source document $\mathcal{D}$ is denoted by $\mathcal{F}_{\mathcal{D}}$.

The reason we use a pair of named entities to perform lexical query from $\mathcal{K}$ is to capture the relationship between a pair of named entities as it appears in a knowledge base to capture their semantics and assist in disambiguation. After the candidate facts $\mathcal{F}_{\mathcal{D}}$ are retrieved from the knowledge bases $\mathcal{K}$, we embed the candidate facts using BioBERT. Then, we perform efficient vector similarity search using Maximum Inner Product Search (MIPS) [124] implemented in the FAISS library [4] to query for the top-k facts among the candidate facts ($\mathcal{F}_{\mathcal{D}}$) using the input document $\mathcal{D}$ as the query. Formally, we define the similarity between fact $F_i \in \mathcal{F}_{\mathcal{D}}$ and document $\mathcal{D}$ as in Equation 5.2.

[3] https://bit.ly/3E0zrll
[4] https://github.com/facebookresearch/faiss

$$sim(F_i, \mathcal{D}) = \vec{\mathcal{V}}(F_i)^T \vec{\mathcal{V}}(\mathcal{D}) \tag{5.2}$$

where $\vec{\mathcal{V}}(F_i)$ - Vector representation of Fact F_i;
$\vec{\mathcal{V}}(\mathcal{D})$ - Vector representation of document $\mathcal{D}$

Thus, after the knowledge retrieval task, we have 1) the input document $\mathcal{D}$ which is obtained during the data preparation phase for ICD-11 and readily available for PubMed-50k; 2) the named entity chain (i.e., chain of named entities extracted from the pseudo-doc) $\mathcal{E}$ [125]; and 3) top-k facts $F_1, F_2, F_3, \ldots, F_{|K|}$ retrieved from the background knowledge bases collectively represented as $F_K \subseteq \mathcal{F}_\mathcal{D}$. We set the value of K to 3 following the study by [126]. We experiment with different values of K as detailed in the ablation studies section. The combination of these contextual signals will be used to guide the summarization model at training/inference time. The rationale for using maximum inner product search for knowledge retrieval is inspired by the works of [127, 128, 129, 130, 131], albeit they used it mainly for open domain question answering [132, 133]. [126] use a similar approach for exemplar retrieval in their RetrievalSum model which is based on contrastive learning [134] using a Siamese network [135] to learn representations for an input document and the exemplars and guide their summary generation. Our problem of retrieving the most relevant facts from the background KB, however, is framed as a dense passage retrieval problem. Named entities from the input document are extracted using the SciSpacy NER model trained on the BC5CDR corpus [118]. Table 5.2 shows the named entities based statistics of ICD-1000-Summ dataset we curate and Table 5.4 shows sample facts, as they appear and retrieved from UMLS KB for an input article with a given pair of named entities identified.

Knowledge-guided Abstractive Summarizer The backbone component of our knowledge-guided abstractive summarizer, which is a transformer encoder-decoder model, is based on

ICD Chapter	Chapter Title	Disease Related Named Entities Statistics					
		Max	Min	Avg	Std	Total Unique Count	Total Count
1	Certain infectious or parasitic diseases	49	0	7.0	5.8	2486	6448
2	Neoplasms	28	0	7.0	5.3	1729	6110
3	Diseases of the blood or blood-forming organs	40	0	8.0	5.9	2585	7467
4	Diseases of the immune system	40	0	7.4	5.3	2387	6929
18	Pregnancy, childbirth, or the puerperium	40	0	8.1	6.3	2774	7667
19	Certain conditions originating in the perinatal period	52	0	8.9	6.9	2774	8630
20	Developmental anomalies	40	0	9.5	6.9	3749	9248
22	Injury, poisoning or certain other consequences of external cause	38	0	7.9	6.5	3045	7544

Table 5.2: Statistics of Disease-related Named Entities for ICD-Summ-1000 Dataset. Note that these statistics are for the total 1000 raw abstracts queried using the lexicons built. That is why some abstracts do not have named entities (based on SciSpacy NER) (as reflected by Min = 0 in each ICD-11 Chapter. The extractive psuedo-docs, however, are guaranteed to have at least three entities. Keywords are not necessarily the same as named entities

ICD Chapter	Chapter Title	Disease Related Named Entities Statistics					
		Max	Min	Avg	Std	Total Unique Count	Total Count
1	Certain infectious or parasitic diseases	357	3	94.8	135.0	332	569
2	Neoplasms	378	4	136.5	169.0	315	546
3	Diseases of the blood or blood-forming organs	400	3	75.5	138.9	387	680
4	Diseases of the immune system	199	3	43.4	71.5	218	347
18	Pregnancy, childbirth, or the puerperium	171	5	55.7	65.4	251	390
19	Certain conditions originating in the perinatal period	130	3	45.0	53.9	172	270
20	Developmental anomalies	132	3	48.0	65.1	214	288
22	Injury, poisoning or certain other consequences of external cause	168	3	44.2	54.8	240	354

Table 5.3: Statistics of Disease-related Named Entities for Extractive Pseudo-doc Dataset.

Named Entity Pair	Entity-driven Facts from UMLS KB
(iron, anemia)	*Iron* deficiency *anemia* secondary to inadequate dietary *iron* intake. *Iron* deficiency *anemia* in mother complicating childbirth.
(dementia, depression)	Primary degenerative *dementia* of the Alzheimer type, presenile onset, with *depression*. Arteriosclerotic *dementia* with *depression*.
(diabetes, hypertension)	*Hypertension* in chronic kidney disease due to type 1 *diabetes* mellitus. *Hypertension* concurrent and due to end stage renal disease on dialysis due to type 2 *diabetes* mellitus.

Table 5.4: Pairs of named entities and sample facts mined from UMLS for each pair. The maximum number of facts extracted is discussed in the experiments section.

the work by [136]. Figure 5.3 shows the proposed end-to-end model architecture. We use this architectural setup for all the models we experiment with. We designate a model augmented with one of the knowledge signals as model-EFAS. We train the models on the 50k samples obtained from PubMed abstractive scientific summarization dataset [77] using different combinations of signals (with and without named entities and facts). The top-k facts retrieved by the biomedical knowledge retriever, corresponding to each pair of named entities in an input extractive pseudo-doc or input article, are separated by a special token **[SEP]**. The input article is passed as one input document prepended with **[CLS]** and appended with **[SEP]** token. The named entity chain is passed as one segment prepended with **[CLS]** and appended with **[SEP]** token. There have been different approaches to combining different signals such as concatenating the different pieces to prime the generation component such as the one proposed in Fusion-in Decoder [130] and [131]. The top-k retrieved facts are initialized using BioBERT and the concatenated encoding is then passed through a sequence of transformer layers to be projected onto a 768- dimension vector to later be attended to by the autoregressive decoder. Similarly, the named entity chain is initialized with BioBERT and passed through a sequence of transformer encoders. Each transformer encoder layer is composed of self-attention and feed forward sub-layers. At training time, a batch of input-output pairings is passed to the encoder and decoder respectively in the form $\langle x, y \rangle$. The encoder undergoes the following transformations to the input sequence x which in this formalism is used to represent the first hidden layer h^0 of the stacked sequence of l transformer encoder layers.

$$\tilde{h}_x^l := LayerNorm(h_x^{l-1} + MHAtt(h_x^{l-1}))$$
$$h_x^l := LayerNorm(\tilde{h}_x^l + FFN(\tilde{h}_x^l))$$

The decoder component, which is trained using teacher forcing [65] at training time, consists of two cross-attention sub-layers to attend to: 1) the input source article; and 2)

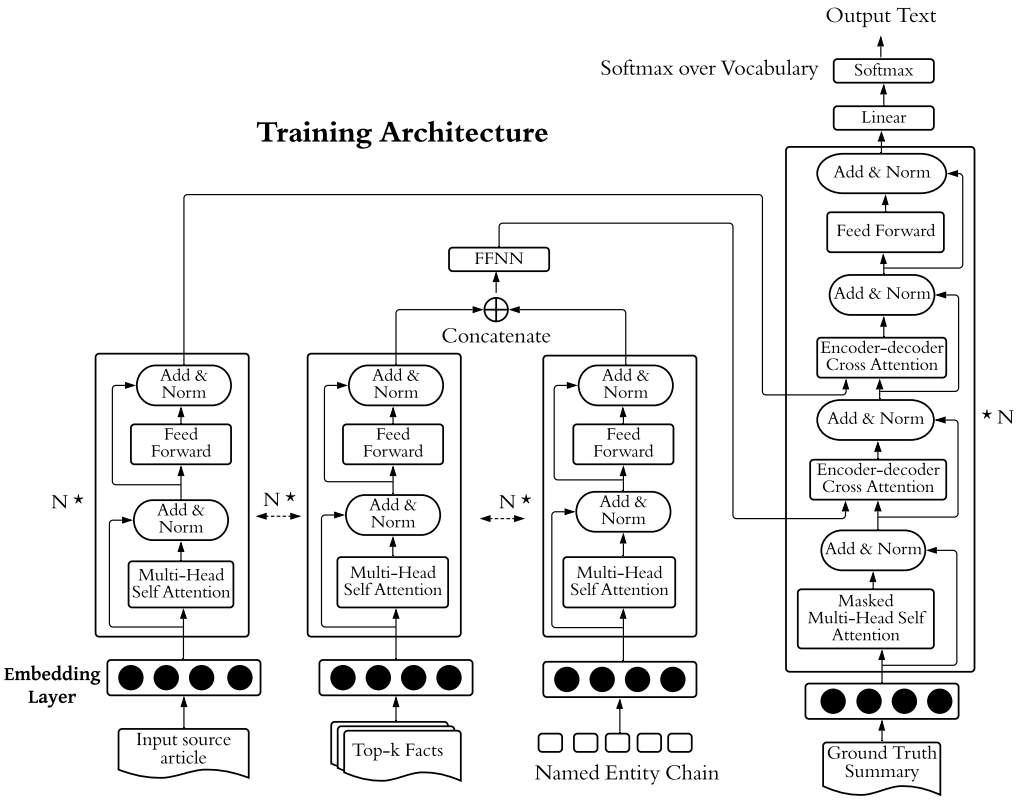

Figure 5.3: The Proposed Framework. The encoder networks have their parameters shared. The two cross attention sub-layers in the decoder attend to the input source article, and a linear transformed projection of encodings of facts, and the chain of named entities. This architecture best represents the three traditional transformer models. For BigBird, and LED, the full self attention layer gets replaced with sparse attention.

the affine transformed concatenation of facts and named entity chain's encodings. The following formulations show the transformations in the decoder component where the ground truth output sequence y is passed to the sequence of transformer decoder layers and is used to initialize the first hidden layer h^0 of the decoder network. Note that we have a Masked Multi-head Self-Attention in the decoder network denoted by $MMHAtt$.

$$\tilde{h}_y^{l-2} := LayerNorm(h_y^{l-3} + MMHAtt(h_y^{l-3}))$$
$$\tilde{h}_y^{l-1} := LayerNorm(\tilde{h}_y^{l-2} + CrossAtt(\tilde{h}_y^{l-2}, F_K, \mathcal{E})$$
$$\tilde{h}_y^{l} := LayerNorm(\tilde{h}_y^{l-1} + CrossAtt(\tilde{h}_y^{l-1}, x))$$
$$h_y^{l} := LayerNorm(\tilde{h}_y^{l} + FFN(\tilde{h}_y^{l}))$$

5.3.3 Experiments and Results

Model Training All models are trained with a cross-entropy loss using backpropagation formally defined in Equation 5.3:

$$\mathcal{L}_\theta = -\frac{1}{n}\sum_{k=1}^{n}\mathcal{P}(t_k|t_{<k}, X, \mathcal{E}, F_K; \theta) \tag{5.3}$$

Where X - the input sequence to be summarized;

$\mathcal{E}$ - the named entities chain in the input sequence X;

F_K - top-k facts extracted from biomedical KB;

θ - model parameters.

We train each model using cross entropy loss to generate the ground truth summaries for the PubMed-50k dataset. We use the following hyperparameters setting: number of epochs is 5, fixed learning rate is set to 5e-5 with Adam optimizer [110], batch size to 8, beam size to 5 with a length penalty [137] α between the range of 0.6 and 1 [41], at inference time. To deal with long-document summarization using the traditional transformer encoder-decoder models, we split the source article into chunks of a maximum of 512 tokens and independently encode each chunk, after which we concatenate and project back to 768 dimension using a linear layer. The approach of splitting the long input sequence into smaller chunks of 512 tokens and then embedding independently is motivated by the recent work to [78]. For Longformer-Encoder-Decoder (LED) [138] and BigBird [139], however, we set the maximum length of the input sequence to 8192 tokens since they can deal with long input sequences without having to truncate; maximum output sequence length is set to 210 tokens following the experiments by [77]. To mitigate redundancy in the generated summaries, we enable trigram blocking [140] during inference. For each backbone model, we use its base variant with 12 encoder and 12 decoder layers. The train/validation/test sizes for PubMed-50k are 50,000/5,000/5,000 and each model is trained using early stopping. A checkpoint of the model that performs the best (in terms of validation loss) on the

validation set across different epochs is saved to the file system. All models are built and trained using PyTorch on NVIDIA Tesla T4 GPU. We perform model training experiments with different input guidance settings as shown in Table Table 5.5:

Training Setting	Training Configuration
TC-I	Input document only
TC-II	Input document + named entities chain
TC-III	Input document + named entities chain + knowledge facts

Table 5.5: Training Configuration.

For our base summarization model, we experiment with five transformer-based encoder-decoder models and show that our entity-driven knowledge-aware approach enables us to achieve the best performance in entity-level factual consistency, N-gram novelty, and semantic equivalence while performing comparably on the commonly used ROUGE metrics. At inference time, we experiment with two settings (w/o named entities, and w/ named entities).

Experiments While all models are trained using the PubMed-50k corpus, they are evaluated using a hold-out test set from the original PubMed dataset as well as the ICD-11-Summ-1000 corpus we curate. The experimental results are shown in Table 5.6 through Table 5.13. Results of evaluation w.r.t source articles reported are average results for both the ICD-11-Summ-1000 and PubMed corpora since the ICD-11 pseudo-extractive documents do not have a ground truth summary. For lexical (ROUGE) evaluation, we report ROUGE F1 scores [141]. Similarly, for evaluation conducted w.r.t source articles, the results reported are average across the PubMed and ICD-11-Summ-1000 corpora. Entity-level factual accuracy [142] is measured in terms of precision, and recall w.r.t ground truth summary (for PubMed), and w.r.t source articles (for both PubMed and ICD-11-Summ-1000). Entity-level precision and recall w.r.t ground truth summaries are denoted with precision-target and recall-target; similarly, entity-level precision, and recall w.r.t the source

article are designated with precision-source, and recall-source, respectively. The F1 score is the harmonic mean of the precision and recall for either case.

Equations 5.4 - 5.9 formalize recall-target, precision-target, precision-source, recall-source as well as their harmonic sums.

$$\text{Recall}_{target} = \frac{\sum_{S \in TargetSummaries} \sum_{entities \in S} Count_{match}(entities)}{\sum_{S \in TargetSummaries} \sum_{entities \in S} Count(entities)} \tag{5.4}$$

$$\text{Precision}_{target} = \frac{\sum_{S \in TargetSummaries} \sum_{entities \in S} Count_{match}(entities)}{\sum_{S' \in CandidateSummaries} \sum_{entities \in S'} Count(entities)} \tag{5.5}$$

$$\text{F-1}_{target} = 2 \cdot \frac{\text{Recall}_{target} \cdot \text{Precision}_{target}}{\text{Recall}_{target} + \text{Precision}_{target}} \tag{5.6}$$

$$\text{Recall}_{source} = \frac{\sum_{S \in SourceArticles} \sum_{entities \in S} Count_{match}(entities)}{\sum_{S \in SourceArticles} \sum_{entities \in S} Count(entities)} \tag{5.7}$$

$$\text{Precision}_{source} = \frac{\sum_{S \in SourceArticles} \sum_{entities \in S} Count_{match}(entities)}{\sum_{S' \in CandidateSummaries} \sum_{entities \in S'} Count(entities)} \tag{5.8}$$

$$\text{F-1}_{source} = 2 \cdot \frac{\text{Recall}_{source} \cdot \text{Precision}_{source}}{\text{Recall}_{source} + \text{Precision}_{source}} \tag{5.9}$$

For measuring semantic equivalence between generated summaries and ground truth summaries, we leverage BERTScore as proposed by [2]; specifically, we use BioBERT for representing each token in a generated summary and in the ground truth summary after which we perform pairwise cosine similarity as proposed in [2]. All experimental results are reported in percentages. The average full text length of input source articles in PubMed-

50k is 3,224 words and the average abstract length is 218 words, while for ICD-11-Summ-1000, the average length of an extractive pseudo-doc (i.e., input source article) is 4816 words.

Backbone Model	Model Variant (K=3)	Training Config	R-1	R-2	R-L
T5	T5 Vanilla (Baseline)	TC-I	**31.333**	**12.821**	**29.018**
	T5 w/ named entities (Ours)	TC-II	29.915	11.352	27.667
	T5 w/ named entities /w facts - EFAS (Ours)	TC-III	28.643	11.286	26.591
BART	BART Vanilla (Baseline)	TC-I	**34.214**	**13.830**	**31.545**
	BART w/ named entities (Ours)	TC-II	32.377	11.733	29.910
	BART w/ named entities /w facts - EFAS (Ours)	TC-III	31.283	10.528	28.174
Pegasus	Pegasus Vanilla (Baseline)	TC-I	28.851	11.274	26.859
	Pegasus w/ named entities (Ours)	TC-II	30.365	11.483	28.003
	Pegasus w/ named entities /w facts - EFAS (Ours)	TC-III	**30.872**	**12.031**	**28.263**
BigBird	BigBird Vanilla (Baseline)	TC-I	**35.426**	**13.801**	**32.537**
	BigBird w/ named entities (Ours)	TC-II	33.491	12.362	30.184
	BigBird w/ named entities /w facts - EFAS (Ours)	TC-III	31.936	13.162	28.730
LED	LED Vanilla (Baseline)	TC-I	**36.218**	**14.173**	**32.862**
	LED w/ named entities (Ours)	TC-II	33.734	13.825	30.614
	LED w/ named entities /w facts - EFAS (Ours)	TC-III	33.283	13.582	29.038

Table 5.6: Lexical (ROUGE) Evaluation w.r.t Ground Truth Summary (*vanilla input @ inference time*). The input in this experimental setting is the *raw input article to be summarized (i.e., w/o named entity chain)*. It can be seen that ROUGE scores are generally higher with the vanilla setting except for Pegasus.

Backbone Model	Model Variant (K=3)	Training Config	Entity-level Factual Consistency		
			Precision-target	Recall-target	F1 score-target
T5	T5 Vanilla (Baseline)	TC-I	27.008	**21.175**	**23.738**
	T5 w/ named entities (Ours)	TC-II	**27.564**	19.246	22.666
	T5 w/ named entities /w facts - EFAS (Ours)	TC-III	27.329	19.136	22.510
BART	BART Vanilla (Baseline)	TC-I	**28.315**	**20.404**	**23.718**
	BART w/ named entities (Ours)	TC-II	27.949	19.105	22.695
	BART w/ named entities /w facts - EFAS (Ours)	TC-III	27.241	18.792	22.241
Pegasus	Pegasus Vanilla (Baseline)	TC-I	17.911	20.212	18.992
	Pegasus w/ named entities (Ours)	TC-II	22.950	21.335	22.113
	Pegasus w/ named entities /w facts - EFAS (Ours)	TC-III	**23.572**	**22.956**	**23.260**
BigBird	BigBird Vanilla (Baseline)	TC-I	16.523	19.384	17.840
	BigBird w/ named entities (Ours)	TC-II	23.273	21.831	22.529
	BigBird w/ named entities /w facts - EFAS (Ours)	TC-III	**25.317**	**23.839**	**24.556**
LED	LED Vanilla (Baseline)	TC-I	17.830	20.173	18.929
	LED w/ named entities (Ours)	TC-II	24.528	22.573	23.510
	LED w/ named entities /w facts - EFAS (Ours)	TC-III	**26.827**	**25.322**	**26.053**

Table 5.7: Entity-level Factual Consistency Evaluation w.r.t Ground Truth Summary (*vanilla input @ inference time*). The input in this experimental setting is the *raw input article to be summarized (i.e., w/o named entity chain)*. We see that entity-level factual consistency metrics improve for Pegasus, BigBird, and LED as we inject intrinsic and extrinsic semantic signals during training. On the other hand, since we are using vanilla input during inference for this experimental setting, we also see the vanilla-trained versions of T5, and BART perform well when tested with vanilla input.

Backbone Model	Model Variant (K=3)	Training Config	Entity-level Factual Consistency		
			Precision-source	Recall-source	F1 score-source
T5	T5 Vanilla (Baseline)	TC-I	**55.076**	**7.976**	**13.934**
	T5 w/ named entities (Ours)	TC-II	54.015	7.232	12.756
	T5 w/ named entities /w facts - EFAS (Ours)	TC-III	53.284	6.275	11.228
BART	BART Vanilla (Baseline)	TC-I	58.592	**5.623**	**10.261**
	BART w/ named entities (Ours)	TC-II	60.422	5.361	9.848
	BART w/ named entities /w facts - EFAS (Ours)	TC-III	**61.593**	4.739	8.801
Pegasus	Pegasus Vanilla (Baseline)	TC-I	33.821	7.401	12.144
	Pegasus w/ named entities (Ours)	TC-II	46.757	7.743	13.286
	Pegasus w/ named entities /w facts - EFAS (Ours)	TC-III	**48.387**	**8.263**	**14.116**
BigBird	BigBird Vanilla (Baseline)	TC-I	34.288	9.261	14.583
	BigBird w/ named entities (Ours)	TC-II	48.283	8.625	14.636
	BigBird w/ named entities /w facts - EFAS (Ours)	TC-III	**48.572**	**9.583**	**16.008**
LED	LED Vanilla (Baseline)	TC-I	59.361	6.731	12.091
	LED w/ named entities (Ours)	TC-II	62.479	6.382	11.581
	LED w/ named entities /w facts - EFAS (Ours)	TC-III	**63.731**	**6.821**	**12.323**

Table 5.8: Entity-level Factual Consistency *w.r.t source article*. The input in this experimental setting is the raw input article to be summarized @ inference time (i.e., w/o named entity chain). From this table, we see that injecting named entity chain and facts during training generally enables the transformer models to hallucinate less as evidenced by the precision-source scores.

Backbone Model	Model Variant (K=3)	Training Config	R-1	R-2	R-L
T5	T5 Vanilla (Baseline)	TC-I	29.837	11.386	27.493
	T5 w/ named entities (Ours)	TC-II	**32.183**	**13.725**	**29.398**
	T5 w/ named entities /w facts - EFAS (Ours)	TC-III	29.372	9.682	28.275
BART	BART Vanilla (Baseline)	TC-I	34.762	12.592	29.387
	BART w/ named entities (Ours)	TC-II	**35.281**	**12.938**	**31.276**
	BART w/ named entities /w facts - EFAS (Ours)	TC-III	33.731	11.923	30.285
Pegasus	Pegasus Vanilla (Baseline)	TC-I	26.592	10.052	24.386
	Pegasus w/ named entities (Ours)	TC-II	32.562	**13.864**	30.174
	Pegasus w/ named entities /w facts - EFAS (Ours)	TC-III	**33.824**	13.841	**30.639**
BigBird	BigBird Vanilla (Baseline)	TC-I	28.174	11.371	25.692
	BigBird w/ named entities (Ours)	TC-II	32.281	**14.263**	**31.863**
	BigBird w/ named entities /w facts - EFAS (Ours)	TC-III	**34.728**	13.264	31.752
LED	LED Vanilla (Baseline)	TC-I	34.265	10.826	26.173
	LED w/ named entities (Ours)	TC-II	**36.840**	13.773	**32.156**
	LED w/ named entities /w facts - EFAS (Ours)	TC-III	34.927	**14.003**	30.851

Table 5.9: Lexical (ROUGE) Evaluation w.r.t Ground Truth Summary (*input article + named entity chain @ inference time*); i.e., the input in this experimental setting is the *raw input article to be summarized with the named entities (i.e., w/ named entity chain)*. Here, we mostly see that ROUGE scores (evaluated with named entities included during inference) are higher with the inclusion of named entities during training. This is expected as named entities used during training are similarly used during inference.

Backbone Model	Model Variant (K=3)	Training Config	Entity-level Factual Consistency		
			Precision-target	Recall-target	F1 score-target
T5	T5 Vanilla (Baseline)	TC-I	26.194	19.759	22.526
	T5 w/ named entities (Ours)	TC-II	**29.826**	**22.952**	**25.941**
	T5 w/ named entities /w facts - EFAS (Ours)	TC-III	28.582	20.738	24.036
BART	BART Vanilla (Baseline)	TC-I	26.581	18.381	21.733
	BART w/ named entities (Ours)	TC-II	**27.949**	**19.105**	**22.696**
	BART w/ named entities /w facts - EFAS (Ours)	TC-III	27.241	18.792	22.241
Pegasus	Pegasus Vanilla (Baseline)	TC-I	15.386	19.382	17.154
	Pegasus w/ named entities (Ours)	TC-II	24.638	23.529	24.071
	Pegasus w/ named entities /w facts - EFAS (Ours)	TC-III	**25.498**	**24.374**	**24.923**
BigBird	BigBird Vanilla (Baseline)	TC-I	15.942	19.873	17.692
	BigBird w/ named entities (Ours)	TC-II	26.315	**24.728**	**25.497**
	BigBird w/ named entities /w facts - EFAS (Ours)	TC-III	**26.638**	24.163	25.340
LED	LED Vanilla (Baseline)	TC-I	17.284	20.692	18.835
	LED w/ named entities (Ours)	TC-II	**28.173**	25.866	**26.970**
	LED w/ named entities /w facts - EFAS (Ours)	TC-III	26.116	**26.830**	26.468

Table 5.10: Entity-level Factual Consistency w.r.t Ground Truth Summary. The input in this experimental setting is the *raw input article to be summarized with the named entities (i.e., w/ named entity chain)* @ inference time. We see that precision-target and recall-target of models improve when they are trained with the inclusion of the additional semantic signals.

Backbone Model	Model Variant (K=3)	Training Config	Entity-level Factual Consistency		
			Precision-source	Recall-source	F1 score-source
T5	T5 Vanilla (Baseline)	TC-I	52.183	5.792	10.427
	T5 w/ named entities (Ours)	TC-II	**56.803**	**10.816**	**18.172**
	T5 w/ named entities /w facts - EFAS (Ours)	TC-III	55.728	8.629	14.944
BART	BART Vanilla (Baseline)	TC-I	56.611	5.031	9.241
	BART w/ named entities (Ours)	TC-II	**62.385**	**7.284**	**13.045**
	BART w/ named entities /w facts - EFAS (Ours)	TC-III	61.938	6.382	11.572
Pegasus	Pegasus Vanilla (Baseline)	TC-I	31.492	6.792	11.174
	Pegasus w/ named entities (Ours)	TC-II	48.389	8.396	14.309
	Pegasus w/ named entities /w facts - EFAS (Ours)	TC-III	**48.964**	**9.491**	**15.900**
BigBird	BigBird Vanilla (Baseline)	TC-I	31.882	8.177	13.016
	BigBird w/ named entities (Ours)	TC-II	48.733	9.267	15.573
	BigBird w/ named entities /w facts - EFAS (Ours)	TC-III	**50.373**	**11.274**	**18.424**
LED	LED Vanilla (Baseline)	TC-I	58.316	6.472	11.651
	LED w/ named entities (Ours)	TC-II	63.722	**8.537**	**15.057**
	LED w/ named entities /w facts - EFAS (Ours)	TC-III	**65.180**	8.374	14.841

Table 5.11: Entity-level Factual Consistency *w.r.t input source article* (*input article + named entity chain @ inference time*); i.e., the input in this experimental setting is the *raw input article to be summarized with the named entities (i.e., w/ named entity chain)*. In this experimental setting, we see that precision-source and recall-source consistently improve when a model is trained and tested with the inclusion of semantic signals, which means the models is less prone to hallucinating irrelevant entities while generating summaries.

Backbone Model	Model Variant (K=3)	Training Config	N-gram Novelty	
			w/o named entities	w/ named entities
T5	T5 Vanilla (Baseline)	TC-I	52.930	49.699
	T5 w/ named entities (Ours)	TC-II	50.079	50.967
	T5 w/ named entities /w facts - EFAS (Ours)	TC-III	**53.817**	**52.841**
BART	BART Vanilla (Baseline)	TC-I	54.816	54.997
	BART w/ named entities (Ours)	TC-II	54.959	57.811
	BART w/ named entities /w facts - EFAS (Ours)	TC-III	**57.360**	**61.370**
Pegasus	Pegasus Vanilla (Baseline)	TC-I	51.260	50.035
	Pegasus w/ named entities (Ours)	TC-II	52.558	51.269
	Pegasus w/ named entities /w facts - EFAS (Ours)	TC-III	**54.621**	**52.702**
BigBird	BigBird Vanilla (Baseline)	TC-I	49.783	51.374
	BigBird w/ named entities (Ours)	TC-II	52.729	**54.836**
	BigBird w/ named entities /w facts - EFAS (Ours)	TC-III	**53.661**	53.827
LED	LED Vanilla (Baseline)	TC-I	53.732	53.288
	LED w/ named entities (Ours)	TC-II	55.826	58.637
	LED w/ named entities /w facts - EFAS (Ours)	TC-III	**59.283**	**61.482**

Table 5.12: N-gram Novelty w.r.t source articles w/o and w/ named entity chain during inference. As can be seen, the models' capability of paraphrasing a source article improves when we include semantic signals during training and inference. Particularly, training the models with both intrinsic and extrinsic semantic signals and using the intrinsic signals during inference enables us to achieve high N-gram novelty (paraphrasing).

Backbone Model	Model Variant (K=3)	Training Config	BioBERTScore	
			w/o named entities	w/ named entities
T5	T5 Vanilla (Baseline)	TC-I	52.269	51.682
	T5 w/ named entities (Ours)	TC-II	51.868	52.739
	T5 w/ named entities /w facts - EFAS (Ours)	TC-III	**53.162**	**54.164**
BART	BART Vanilla (Baseline)	TC-I	51.799	50.283
	BART w/ named entities (Ours)	TC-II	51.783	**53.618**
	BART w/ named entities /w facts - EFAS (Ours)	TC-III	**52.072**	51.472
Pegasus	Pegasus Vanilla (Baseline)	TC-I	53.168	51.381
	Pegasus w/ named entities (Ours)	TC-II	53.401	**55.761**
	Pegasus w/ named entities /w facts - EFAS (Ours)	TC-III	**54.382**	55.263
BigBird	BigBird Vanilla (Baseline)	TC-I	55.271	53.620
	BigBird w/ named entities (Ours)	TC-II	**56.813**	54.271
	BigBird w/ named entities /w facts - EFAS (Ours)	TC-III	56.372	**55.088**
LED	LED Vanilla (Baseline)	TC-I	53.732	52.427
	LED w/ named entities (Ours)	TC-II	**54.163**	55.791
	LED w/ named entities /w facts - EFAS (Ours)	TC-III	53.814	**57.284**

Table 5.13: Semantic Equivalence (BioBERTScore [2]] w.r.t ground truth summaries w/o and w/ named entity chain during inference. Since we are using BioBERT for representation learning, we refer to the metric as BioBERTScore, a variant of BERTScore. As can be seen, we obtained the best semantic equivalence scores when the models are trained with the inclusion of the semantic signals during training and the semantic signals included during inference.

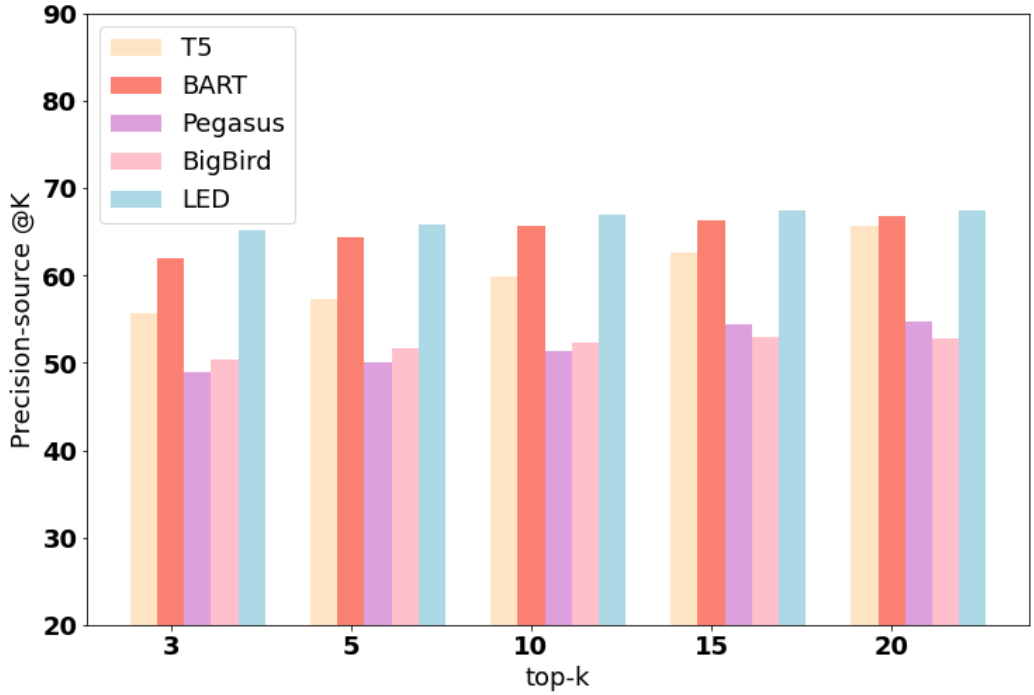

Figure 5.4: Precision-source for different values of K.

Ablation Studies To assess the impact of facts mined on the quality of summaries generated, we conduct an ablation study where we experiment with different values of K in top-k for the backbone models. Figure 5.4 and Figure 5.5 shows results of ablation to assess precision-source, and recall-target. Since we want to minimize entity hallucination which is measured in terms of precision-source and want to maximize the number of entities in the ground truth summary that are retrieved in the generated summary as measured by recall-target, we report the impact of different values of K for these two metrics. As shown in the two plots, precision-source and recall-target consistently improve as we retrieve more relevant facts from the biomedical knowledge bases and train our models.

Discussion of Results From the results reported in the previous section, we generally see entity-level factual consistency (particularly, precision-source, and recall-target) improve when a model is trained with named entities and/or facts included as an additional signal in the training with the same objective of generating the ground truth summary using

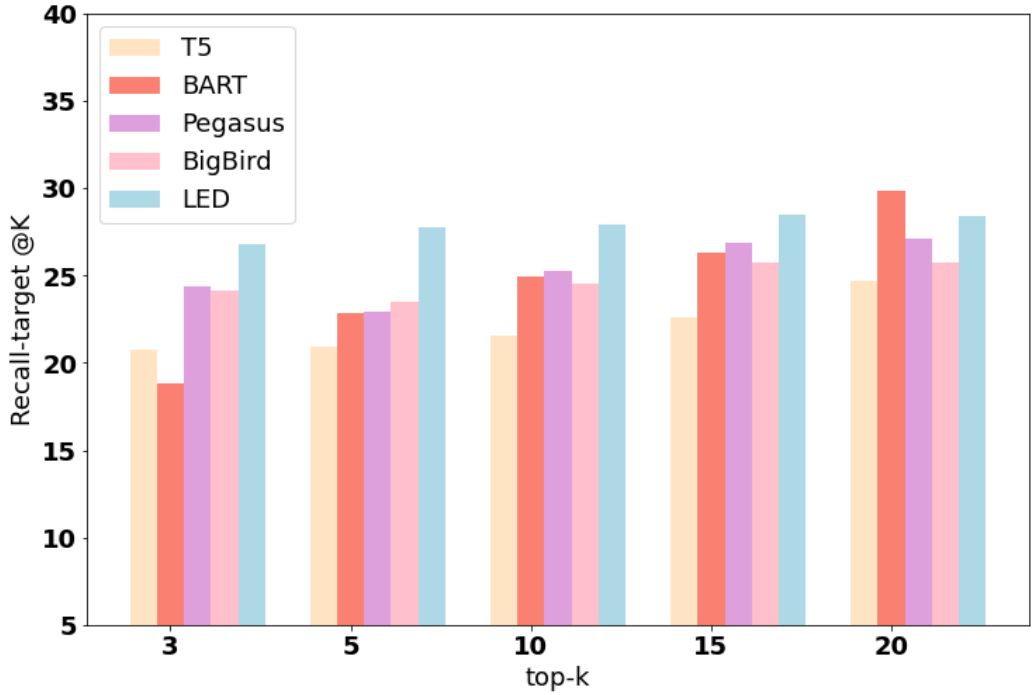

Figure 5.5: Recall-target for different values of K.

cross-entropy loss. The addition of more facts further improves entity-level factual consistency as shown in Figure 5.4 and Figure 5.5. Further, we notice N-gram novelty improves with our proposed framework for the five backbone models. Semantic equivalence generally improves when named entities and/or facts are included during training for all models. Thus, the corresponding entries for the various models and training configurations show improvement in semantic based scores. The ROUGE scores, however, drop slightly from when there is no additional context at training or inference time. The drop in ROUGE is a result of augmenting the models with facts from background knowledge bases which in turn leads to higher N-gram novelty. Thus, the proposed framework enables us to achieve better abstractive scores in terms of entity-level factual consistency, paraphrasing and semantic equivalence. With regards to evaluation across the ICD-11 chapters, while there is a slight variation among the chapters, we did not notice a significant difference in evaluation metrics (semantic equivalence, entity-level factual consistency, and N-gram novelty).

PubMed Article Abstract/Ground Truth Summary: cigarette smoke is considered a major risk factor for vascular diseases . there are many toxic compounds in cigarette smoke , including acrolein and other ,-unsaturated aldehydes , which are regarded as mediators of inflammation and vascular dysfunction . furthermore , recent studies have revealed that acrolein , an ,-unsaturated aldehyde in cigarette smoke , induces inflammatory mediator expression , which is known to be related to vascular diseases . in this study , we investigated whether korean red ginseng (krg) water extract suppressed acrolein - induced cyclooxygenase (cox)-2 expression in human umbilical vein endothelial cells (huvecs) . acrolein - induced cox-2 expression was accompanied by increased levels of phosphorylated p38 in huvecs and krg inhibited cox-2 expression in huvecs . these results suggest that krg suppresses acrolein - induced cox-2 expression via inhibition of the p38 mitogen - activated protein kinase signaling pathway . in addition , krg exhibited an inhibitory effect on acrolein - induced apoptosis , as demonstrated by annexin v propidium iodide staining and terminal deoxynucleotidyl transferase - mediated dutp nick end - labeling assay . consistent with these results , krg may exert a vasculoprotective effect through inhibition of cox-2 expression in acrolein - stimulated human endothelial cells .

BART-vanilla summary: backgroundive. the objective of our study was to investigate the vascular protective effect of korean red ginseng (krg) in human umbilical vein endothelial cells (huvecs). methods. we examined the effects of krg on the expression of cox-2 and apoptosis in acrolein - stimulated peripheral blood vessels (pgs). results. in addition, in vitro evidence suggests that krg may play an important role in the pathogenesis of vascular diseases such as hypertension, stroke, and and/or decreased vasodilation. conclusions. the results suggest that the effect of bg on endothelial damage is not due to bg alone.

BART-EFAS summary: the aim of this study was to investigate the effect of cyclooxygenase (cox) on the expression of cox-2 in human umbilical vein endothelial cells (huvecs). [subjects and methods] forty - four huvec cells were cultured in acrolein - stimulated human cells (ash 2.5%, total crude saponin 70 mg / g, and total ginsenosides 20 mg/ g. [results] the results of our study suggest that krg acts as a cytoprotective agent of inflammation and atherosclerosis. [conclusion] the present study suggests that the anti - inflammatory effect of krg powder may be due to the inhibition of the signaling pathway and may be related to the development of vascular diseases.

Figure 5.6: Sample Generated Summaries using BART vanilla and our proposed variants of BART.

5.4 Conclusion

In this chapter, we proposed a framework to integrate named entities in a source article and facts extracted from biomedical knowledge bases pertaining to the named entities using transformer-based encoder-decoder models and applied to the task of abstractive summarization of biomedical literature. Through extensive experiments, we showed the proposed approach improves the reliability and coverage of generated summaries in terms of entity-level factual consistency and semantic equivalence w.r.t ground truth summaries while generating novel words w.r.t source articles.

Improving the Factual Accuracy and Interpretability of Abstractive Clinical Text Summarization

"I don't know what's the matter with people: they don't learn by understanding; they learn by some other way - by rote, or something. Their knowledge is so fragile!."

—Richard Feynman, 1918 – 1988

In chapter 5, we looked at how abstractive summarization can be improved by utilizing intrinsic semantics in the form of named entities and extrinsic semantics in the form of facts retrieved from external knowledge bases. However, the named entities and associated facts obtained were used as additional contextual input signals with the objective still being generating a summary as semantically close as possible to a ground truth summary. In this chapter, we go one step further, and jointly optimize three objectives: ground truth summary given an input sequence, named entities in ground truth summary given named entities from the input sequence, and entity-relevant facts in ground truth summary given entity-relevant facts in an input sequence. Thus, while the previous chapters dealt with single-objective optimization, in this chapter, we propose multi-objective optimization and apply it to the task of abstractive summarization of clinical text. Further, in this chapter, we investigate how semantically sound summaries can be explained in terms of entity-level

factual accuracy. In other words, we study the relationship between semantic equivalence and entity-level factual accuracy to assess if improvement in entity-level factual accuracy could lead to improved semantic equivalence.

6.1 Why (Motivation)

Recent advances in sequence to sequence models [31] have led to progress in abstractive summarization of news articles, scientific articles, and social media data. However, these models have not been well investigated in the healthcare domain where automated clinical summary generation [143] for a set of findings in clinical notes can be helpful to clinicians for timely and effective clinical decision making. One of the clinical practices entails the task of recording *findings* of diagnosis, treatment or procedures followed by manually summarizing the findings into a form called *impressions*. Inspired by recent efforts in modeling findings-to-impression as summarization [144, 145, 84] , we propose to automate this process of writing an impression from findings to assist clinicians with their practice, making the clinical workflow more efficient. As part of the task of abstractive clinical text summarization, two of the critical aspects of informative summary generation are 1) preserving semantics; and 2) discovering portions of an input clinical note that have led to semantically informative summaries, motivated by a modeling paradigm known as interpretability [146].

6.2 What (Problem Statement)

In the previous chapter, we introduced an approach for leveraging named entities and associated facts mined from medical knowledge bases to model abstractive summarization of biomedical articles. In this chapter, we propose an end-to-end training framework using multi-objective optimization for the task of abstractive summarization of clinical text, pre-

sented as *findings* and assess the entity-level factual accuracy of generated summaries and the interpretability of the quality of semantics in terms of factual accuracy.

Concretely, we propose an abstractive clinical text summarization framework based on multi-objective optimization where we jointly optimize three cost functions in our proposed architecture during training: *generative loss, entity loss*, and *knowledge loss*. We evaluate the proposed architecture on three different datasets. We experiment with three transformer encoder-decoder architectures and demonstrate that optimizing different loss functions leads to improved performance in terms of entity-level factual accuracy and semantic equivalence. We also evaluate how entity-level factual accuracy relates to semantically sound summaries in a fundamental attempt to explainability of abstractive clinical text summarization.

6.3 How (Approach)

6.3.1 Data Collection

We collect clinical notes of 1200 patients with heart failure (HF) from the University of Illinois Hospital & Health Sciences System (UI Health) for our study. Among the clinical notes collected for the 1200 patients, there are a total of 15183 de-identified procedure notes spanning a period of over 4 years (5/2016 - 8/2020). Out of the total 15183 notes, we filtered the ones with no *Findings* or *Impressions* since our research aim is to generate an impression from a set of findings. The findings play the role of input text to be summarized and the impression serves as the ground truth summary. After pre-processing the data, we have 6182 notes consisting of *findings-to-impression* pairings along with other metadata. In addition to our Heart Failure data, we evaluate the proposed approach for Research Aim-4 on two benchmark datasets. The benchmark datasets are 1) radiology reports from the Indiana Network for Patient Care [147]; and 2) 50000 randomly selected chest x-ray

reports from the MIMIC-III-CXR dataset [148] originally curated by Beth Israel Deaconess Medical Center. Figure 6.1 illustrates what a typical clinical note (a record) for a patient with HF in our cohort looks like.

6.3.2 Proposed Framework

The sections below discuss the components in the proposed framework.

Clinical Knowledge Retriever Since our goal is to use named entities and entity-aware facts (from knowledge bases) for modeling abstractive summarization, our first task is to conduct named entity recognition on the *findings* and *impression* of the clinical notes. For this, we use an off-the-shelf Stanza package from Stanford for clinical named entity recognition (NER) [9]. Specifically, the Stanza model we use is the one trained on the i2b2 clinical text dataset. The knowledge bases to query for facts using the named entities are composed of UMLS, SNOMED-CT, and ICD-10. Figure 6.2 shows named entities and entity-aware facts for a given set of *findings* and *impression*, apiece, from the heart failure data. The fact retrieval module follows the same Maximum Inner Product Search using FAISS approach implemented in chapter 5.

For each named entity identified from a set of findings/impression, we perform full-text lexical query of the KBs and return the top-k facts where we set the value of K to 5 [126].

Model Training using Multi-Objective Optimization We experiment with three state-of-the-art transformer encoder-decoder models pretrained using different self-supervised objectives. We propose to train these models using a loss function that optimizes summary generation, named entity chain generation, and fact generation where our task is not only to auto-regressively generate the target summary, but also to generate the named entities in the

Procedure_name:
[PERSONALNAME] Abd and Pelv w/o [PERSONALNAME] cont

Indication:
64-year-old female with history of incarcerated hernia, concern for small bowel obstruction

---

Technique:
Multidetector multiplanar noncontrast [PERSONALNAME] images through the abdomen and pelvis were obtained.

Comparison:
[PERSONALNAME] examination of the abdomen and pelvis

Findings:
Lack of intravenous contrast limits exam interpretation. LUNG BASES: There is a moderate right pleural effusion. There is dependent atelectasis in the lung bases. The heart is slightly increased in size compared to prior examination. There is atherosclerotic calcification of the coronary arteries. LIVER: The liver demonstrates cirrhotic morphology with nodular surface contours. GALLBLADDER AND BILIARY SYSTEM: There are no calcified gallstones. SPLEEN: The spleen is borderline enlarged measuring up to 13 cm in length. PANCREAS: Evaluation of the pancreas is suboptimal in the absence of [PERSONALNAME] contrast. ADRENAL GLANDS: There is a low attenuating lesion in the left adrenal gland measuring approximately 2.2 cm (series 2 image 22) that appears slightly enlarged since February 27, 2018 when it measured approximately 1.9 cm. This is favored to represent an adenoma. KIDNEYS: In the inferior pole cortex of the left kidney there is a 1.2 cm simple cyst, more conspicuous than the prior study. STOMACH: The stomach is mildly distended with air and debris. BOWEL: Postsurgical changes with bowel sutures are again seen in the right lower quadrant. There is mild small bowel dilatation adjacent to the suture line, probably within normal limits postsurgical. There is a focally dilated loop of small bowel in the left mid abdomen measuring up to 4.2 cm (series 2 image 30) with passage of oral contrast distally, suspicious for partial small bowel obstruction. There is passage of oral contrast to the level of the terminal ileum. There is amorphous soft tissue in the mid abdomen (series 2 images 53, 54) which likely represents unopacified small bowel loops rather than mass. There is scattered stool in the colon. PERITONEUM AND RETROPERITONEUM: There is mild to moderate volume ascites. There is no intraperitoneal free air. The abdominal aorta is normal in course and caliber with atherosclerotic calcifications throughout its abdominal course extending into the common iliac arteries.There is no mesenteric or retroperitoneal lymphadenopathy. PELVIS: The urinary bladder is well distended and unremarkable. There is no pelvic lymphadenopathy. Multiple phleboliths are again seen. BONES: There are mild degenerative changes of the spine with a diffuse disc bulges at L4-L5 and L5-S1. Sclerosis of L4-L5 appears unchanged since prior examination. SOFT TISSUES: There is anasarca in the subcutaneous soft tissues. A midline laparotomy scar is again seen.

Impression:
1. Findings suspicious for a proximal, partial small bowel obstruction.
2. Moderate right pleural effusion.
3. Cirrhotic liver morphology.
4. Moderate volume ascites.
5. Postsurgical changes of small bowel resection in the right lower quadrant.
6. Slight increase in size of left adrenal nodule favored to represent an adenoma. These images were reviewed and interpreted with attending radiologist Dr. [PERSONALNAME] before dictation of this final report by resident Dr. [PERSONALNAME].

Figure 6.1: Example de-identified clinical record for the heart failure data collected through the Center for Clinical and Translational Science, University of Illinois, Chicago.

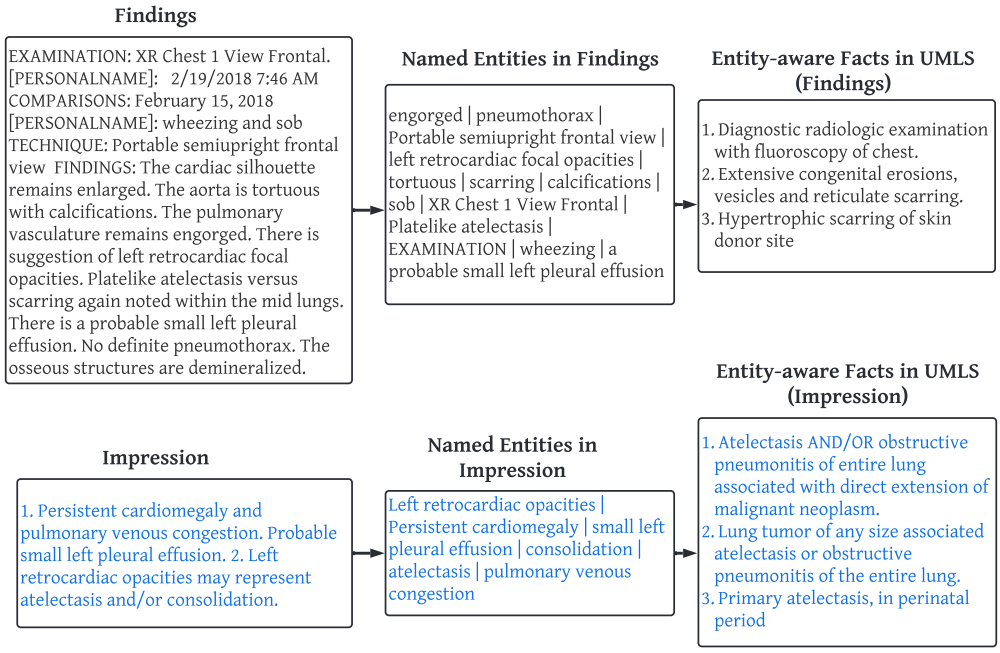

Figure 6.2: Named entities and entity-aware facts for *findings* and *impression*.

impression and to generate the facts associated with the named entities in the impression. Figure 6.3 shows the proposed end-to-end architecture where three networks, with shared parameters, are jointly trained using the loss functions stated in Equation 6.1.

We optimize the total aggregate loss function during the training phase for the proposed model in use. We use Bayesian optimization [149] to search for the best combination of *generative* and *regularization* hyperparameters. The generative hyperparameter is denoted in the formulation using λ_{gen} while the knowledge and entity-based regularization hyperparameters are denoted using λ_k, $\lambda_{\mathcal{E}}$. Each of the hyperparameters takes on values in the range of [0.1, 0.9] with increments of 0.3 and we evaluate the validation loss in each epoch during training to save the model checkpoint with the least validation loss. We experiment with three optimization configurations: i) with generative loss alone; ii) with generative loss and entity chain loss (Dual Multi-Objective Optimization - Dual MOO); and iii) with generative loss, knowledge loss, and entity chain loss (Triple MOO).

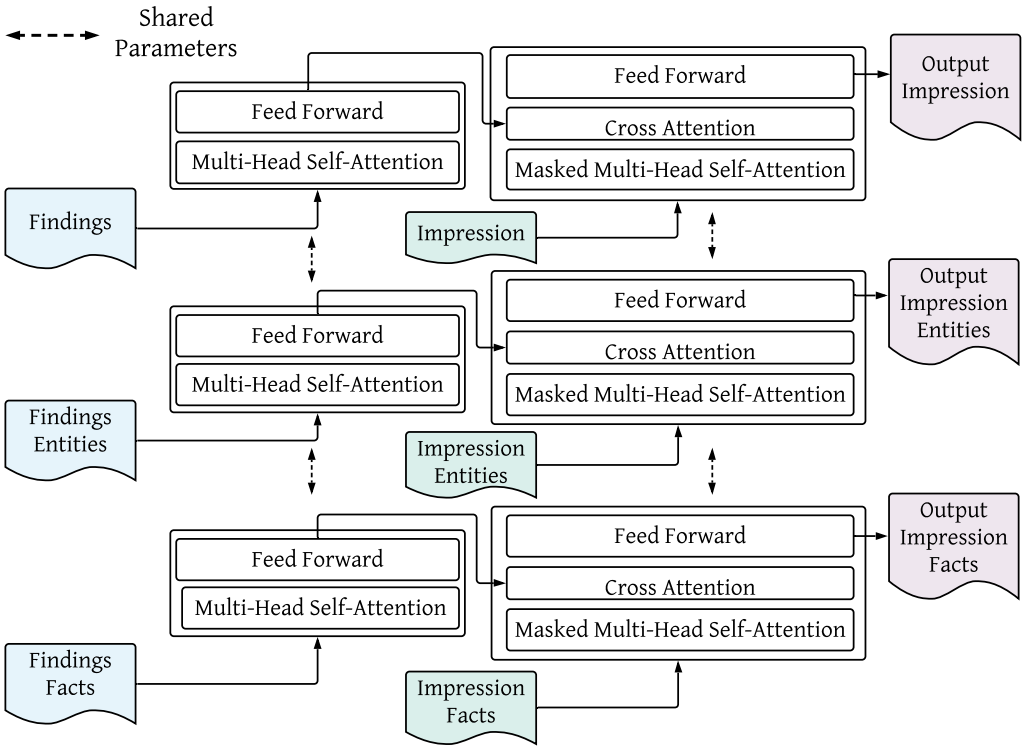

Figure 6.3: The proposed training architecture.

$$\mathcal{L}_{total} = \lambda_{gen} \cdot \mathcal{L}_{gen} + \lambda_k \cdot \mathcal{L}_k + \lambda_{\mathcal{E}} \cdot \mathcal{L}_{\mathcal{E}} \tag{6.1}$$

Each of the loss functions is based on cross-entropy criterion (Equation 6.2).

$$\mathcal{L}_\theta = -\frac{1}{n} \sum_{k=1}^{n} \mathcal{P}(t_k | t_{<k}, \chi; \theta) \tag{6.2}$$

Where χ - the input sequence (i.e., finding, or named entity chain in a finding, or a sequence of facts retrieved from the knowledge bases associated with named entities in a finding). The proposed models are trained with the objective of minimizing the aggregate loss function defined in Equation 6.1. All models are built and trained using PyTorch on Google Cloud NVIDIA Tesla T4 GPU.

6.3.3 Experiments and Results

Table 6.1 shows the statistics of the datasets and Table 6.2 shows the results of evaluation against the impressions (ground truth summary). Our experimental results show that jointly optimizing the task of traditional language modeling with task-specific objectives such as preserving entity-aware factual accuracy improves performance of a model. Specifically, we demonstrate this by leveraging three pre-trained abstractive summarization models and fine-tuning on our datasets using multi-objective optimization. It can be seen from Table 6.2 that Precision-target, and Recall-target increase with our training objective as compared to the language modeling training objective used with the baseline models. As extensively discussed in the literature [150, 145], we also argue that lexical measures (i.e., ROUGE) do not fully quantify the factual accuracy of a generated summary while a metric that measures entity-level overlap between a ground truth summary (impression) and a model-generated summary better reflects the extent to which semantics are preserved in abstractive summarization since named entities constitute significant semantics in a clinical text. As investigated in the preceding research aims, one quantitative measure of abstractive summarization is the rate of novel word generation (i.e., paraphrasing). We report in Table 6.3 the N-gram novelty of the generated summaries (corresponding to different base models and different variants of training configuration) measured w.r.t. *findings*. It can be seen from Table 6.3 that training with Multi-Objective Optimization leads a model to generate more novel words than a vanilla setting for the three backbone models. Further, we also measure the semantic equivalence between 1) *findings* and generated summary; and 2) *impression* and generated summary using the approach we pursued in chapter 5 (i.e., BERTScore). Unlike BioBERTScore, in this chapter, we use Clinical BERT [151] to encode the tokens in findings, impression or the generated summary. Tables 6.4 and 6.5 show semantic equivalence measured w.r.t *findings* and *impression* respectively.

A key limitation of our proposed approach is that it is computationally more expensive and takes longer to train than with customary single task objective training. Another

limitation we observed is that the proposed model training approach can be sensitive to hyperparameter initialization.

Dataset	Train	Validation	Test	Avg # tokens per Findings	Avg # tokens per Impression
Heart Failure (HF)	4000	1091	1091	142	48
IU X-Ray	2200	593	593	33	12
MIMIC-CXR	40000	5000	5000	52	18

Table 6.1: Statistics of the experimental datasets.

Model	R-1	R-2	R-L	Entity-level Factual Accuracy		
				Precision-target	Recall-target	F1 score-target
T5 Vanilla (Baseline)	**35.113**	**19.503**	**34.921**	25.150	42.577	31.621
T5 w/ named entities (dual MOO) - Ours	32.628	18.361	33.827	**29.672**	46.581	36.252
T5 w/ named entities /w facts (triple MOO) - Ours	28.761	17.382	30.599	29.327	**48.148**	**36.451**
BART Vanilla (Baseline)	**22.951**	**16.283**	**22.657**	18.321	29.679	22.656
BART w/ named entities (dual MOO) - Ours	19.827	13.693	19.792	20.629	33.839	25.632
BART w/ named entities /w facts (triple MOO) - Ours	15.721	12.173	16.582	**23.182**	**34.159**	**27.620**
Pegasus Vanilla (Baseline)	**28.193**	**11.387**	**28.079**	21.739	28.593	24.699
Pegasus w/ named entities (dual MOO) - Ours	27.370	9.728	25.372	22.058	**29.781**	**25.344**
Pegasus w/ named entities /w facts (triple MOO) - Ours	24.263	7.836	22.174	**25.661**	25.349	25.504

Table 6.2: Experimental results. Dual MOO refers to dual multi-objective optimization where only the generative loss and entity chain loss are jointly optimized during training. Triple MOO refers to modeling where the three loss functions are jointly optimized. Due to space constraints, we report average scores across the three datasets.

Backbone Model	Training Configuration (K=5)	N-gram Novelty	
		w/o named entities	w/ named entities
T5	T5 Vanilla (Baseline)	5.753	7.319
	T5 w/ Dual MOO (Ours)	15.431	12.824
	T5 w/ Triple MOO (Ours)	**37.165**	**36.817**
BART	BART Vanilla (Baseline)	24.313	23.182
	BART w/ Dual MOO (Ours)	26.724	25.182
	BART w/ Triple MOO (Ours)	**41.278**	**41.255**
Pegasus	Pegasus Vanilla (Baseline)	17.382	16.118
	Pegasus w/ Dual MOO (Ours)	22.735	22.347
	Pegasus w/ Triple MOO (Ours)	**29.813**	**28.601**

Table 6.3: Generated summaries N-gram Novelty w.r.t. *findings* w/o and w/ named entity chain during inference.

Findings: Density in the left upper lung on PA XXXX XXXX represents superimposed bony and vascular structures. There is calcification of the first rib costicartilage junction which XXXX contributes to this appearance. The lungs otherwise appear clear. The heart and pulmonary XXXX appear normal. In the pleural spaces are clear. The mediastinal contour is normal. There are degenerative changes of thoracic spine. There is an electronic cardiac device overlying the left chest wall with intact distal leads in the right heart.

Impression: 1. Irregular density in the left upper lung on PA XXXX, XXXX artifact related to superimposed vascular bony structures. Chest fluoroscopy or XXXX would confirm this 2. Otherwise, no acute cardiopulmonary disease.

BART-vanilla summary: 1. No acute radiographic cardiopulmonary process. 2. Hiatal leads in the left upper lung which XXXX represents the first rib costicartilage junction XXXX contributes to this appearance. 3. No evidence of pulmonary edema. No definite pleural effusion or pneumothorax. No typical appearance of pulmonary XXXX..

BART-Dual-MOO summary: Cardiopulmonary heart disease. The heart and pulmonary heart disease are normal. The pulmonary heart is normal. There is no heart disease in the right heart. The left heart is healthy. The right heart is stable. The lungs are normal and the pulmonary heart and lungs are stable. There are no heart and lung diseases. The lung is stable and stable.

BART-Triple-MOO summary: superimposed bony and vascular structures in the right heart. The heart and pulmonary heart are normal. The cardiopulmonary heart is normal. There is an electronic cardiac device overlying the left chest wall with intact distal leads. The right heart is healthy. The left heart is stable. The pulmonary heart is fine. The lungs are normal and the heart is good. The lung is healthy and stable. There are no complications in the left heart.

Figure 6.4: Sample summaries generated using vanilla, and knowledge-augmented optimization objectives.

Backbone Model	Training Configuration (K=5)	Clinical BERT Score	
		w/o named entities	w/ named entities
T5	T5 Vanilla (Baseline)	49.481	50.824
	T5 w/ Dual MOO (Ours)	48.628	49.271
	T5 w/ Triple MOO (Ours)	**52.602**	**53.183**
BART	BART Vanilla (Baseline)	53.793	52.337
	BART w/ Dual MOO (Ours)	51.694	52.825
	BART w/ Triple MOO (Ours)	**59.792**	**62.278**
Pegasus	Pegasus Vanilla (Baseline)	47.581	49.744
	Pegasus w/ Dual MOO (Ours)	48.291	47.763
	Pegasus w/ Triple MOO (Ours)	**51.382**	**51.803**

Table 6.4: Generated summaries' Semantic Equivalence w.r.t *findings* w/o and w/ named entity chain during inference.

Backbone Model	Training Configuration (K=5)	Clinical BERT Score	
		w/o named entities	w/ named entities
T5	T5 Vanilla (Baseline)	48.173	46.379
	T5 w/ Dual MOO (Ours)	47.862	48.921
	T5 w/ Triple MOO (Ours)	**51.364**	**53.061**
BART	BART Vanilla (Baseline)	50.372	51.741
	BART w/ Dual MOO (Ours)	**51.702**	**54.273**
	BART w/ Triple MOO (Ours)	50.391	52.286
Pegasus	Pegasus Vanilla (Baseline)	46.992	47.630
	Pegasus w/ Dual MOO (Ours)	47.251	48.379
	Pegasus w/ Triple MOO (Ours)	**50.772**	**49.402**

Table 6.5: Generated summaries Semantic Equivalence w.r.t *impression* w/o and w/ named entity chain during inference.

Discussion of Results From the experimental results reported, we can infer that optimizing three cost functions leads to improved performance in terms of semantic equivalence and paraphrasing w.r.t source clinical notes (findings) across all model variants. With regards to entity-level factual accuracy w.r.t the ground truth summaries (impressions), we see improved performance is observed with double optimization or triple optimization. One limitation of the proposed cost minimization strategy is each of the loss functions is based on cross entropy and it is possible to extend the approach using other cost functions such as KL divergence, and euclidean distance to name a few.

To better understand and explain the models we experimented with, we select the top-5 semantically sound summaries generated using BART w/Triple MOO variant (w/-named entity chain during inference) and analyzed how entity-level factual accuracy and semantic equivalence relate to each other. Table 6.6 shows the semantic equivalence scores w.r.t *findings* and entity-level factual accuracy for BART based models for the top-5 (based

on semantic equivalence for summaries generated using BART w/ Triple MOO) clinical notes. Table 6.7 shows semantic equivalence w.r.t *impressions* vs entity-level factual accuracy. From both tables, we generally see that for the top semantically sound summaries, their entity-level factual accuracy is also high and that as the entity-level factual accuracy goes down, semantic equivalence also shows a trend of going down. While not fully conclusive, from this observation, we can infer that entity-level factual accuracy and semantic equivalence are positively correlated.

Heart Failure Clinical Note Record ID	BART model variant	Semantic Equivalence w.r.t *Findings*	Entity-level Factual Accuracy (*F1-target*)
2176	BART vanilla	58.31	29.63
	BART w/ Dual MOO	63.27	32.85
	BART w/ Triple MOO	**68.43**	**34.69**
1305	BART vanilla	59.74	30.82
	BART w/ Dual MOO	64.39	31.72
	BART w/ Triple MOO	**67.25**	**33.81**
1938	BART vanilla	59.27	28.13
	BART w/ Dual MOO	65.91	31.29
	BART w/ Triple MOO	**65.48**	**34.16**
2406	BART vanilla	57.33	26.38
	BART w/ Dual MOO	63.81	29.41
	BART w/ Triple MOO	**65.13**	**32.05**
926	BART vanilla	60.82	25.73
	BART w/ Dual MOO	61.03	27.61
	BART w/ Triple MOO	**63.94**	**30.15**

Table 6.6: Correlation between semantic equivalence w.r.t findings and entity-level factual accuracy. Note that the boldfaced numbers are to show the non-ascending order of semantic equivalence for BART w/ Triple MOO and is not meant to compare with other model variant.

Thus, the proposed multi-objective optimization of cost functions enables for better explainability of summaries whose semantics are preserved since the informativeness of summaries can be better explained by the context in which named entities appear in *findings* and/or *impressions*. In other words, named entities enable us to capture features that contribute positively to the summary generation as certain tokens contribute higher than other tokens in regular deep learning classification tasks [152]. Since preserving semantics is the heart of abstractive summarization, we selected the top semantically sound (as measured by Clinical BERT Score) representative summaries generated using the vanilla, dual MOO, and tripe MOO model variants and show that tokens that are identified as named entities contribute positively to good summary generation. This is the result of the fact

Heart Failure Clinical Note Record ID	BART model variant	Semantic Equivalence w.r.t *impression*	Entity-level Factual Accuracy (*F1-target*)
269	BART vanilla	57.83	34.83
	BART w/ Dual MOO	59.29	35.92
	BART w/ Triple MOO	**61.37**	**37.21**
2738	BART vanilla	56.11	35.66
	BART w/ Dual MOO	59.72	36.86
	BART w/ Triple MOO	**61.03**	**36.27**
79	BART vanilla	58.25	33.17
	BART w/ Dual MOO	60.31	34.02
	BART w/ Triple MOO	**60.26**	**35.97**
683	BART vanilla	58.07	31.2
	BART w/ Dual MOO	57.31	33.15
	BART w/ Triple MOO	**58.39**	**33.78**
1847	BART vanilla	57.93	31.84
	BART w/ Dual MOO	58.61	32.62
	BART w/ Triple MOO	**58.22**	**30.19**

Table 6.7: Correlation between semantic equivalence w.r.t impression and entity-level factual accuracy. Note that the boldfaced numbers, as in Table 6.6, show the non-ascending order of semantic equivalence for BART w/ Triple MOO and is not meant to compare with other model variants.

that the summary generation model puts more weights on named entities than other tokens in the *findings* and/or *impressions*. This attempt of explaining the black box generative models is inspired by the works of [152, 146, 153].

6.4 Conclusion

In this chapter, we proposed a framework based on a transformer encoder-decoder network and transfer learning for clinical text summarization using knowledge-aware multi-objective optimization. We experimentally demonstrated that jointly optimizing generative loss, knowledge loss, and entity-based loss functions significantly improves the quality of generated summaries in terms of entity-level factual accuracy which is critical but less explored in the healthcare domain.

In addition to improving the factual accuracy and semantic equivalence of summaries in abstractive clinical text summarization, we have also made an attempt to go one step further and assess the role named entities play for the task of semantically sound summary generation. Particularly, we empirically analyzed how the semantics of summaries generated using the generative models in this chapter could be interpreted w.r.t named entities. We believe our work lays the foundation for interpretability [153, 152] of generative models

in the context of clinical text summarization.

Summary of Contributions

In this dissertation, we introduced a semantics-driven abstractive document summarization paradigm by exploring document semantics at various levels of granularity and investigated across four domains/tasks. The key contributions of the dissertation can be summarized as follows:

- A framework for utilizing intrinsic semantics of documents (scientific articles, biomedical literature, and clinical notes) for guiding abstractive summarization models in unsupervised and supervised settings for single document as well as multi-document summarization tasks.
- An approach for retrieving extrinsic semantics of documents from domain-specific knowledge bases in the form of related facts and amalgamating the intrinsic semantics (named entity chain) and the extrinsic semantics (facts) into abstractive summarization models during training and inference phases.
- A model that augments supervised sequence to sequence models with unsupervised graph based technique for addressing the classic neural text degeneration problem in neural decoding algorithms.
- A supervised abstractive summarization technique based on knowledge-aware multi-objective optimization driven by intrinsic and extrinsic semantics of documents.
- Datasets, lexicons, codes, and trained model checkpoints to be publicly released.

Future Research Directions

While we explored how intrinsic semantics and extrinsic semantics (derived from knowledge bases) can be used for the task of abstractive document summarization for different domains and tasks, a possible extension of this work is to integrate symbolic knowledge (e.g., logical rules governing a given domain) into the deep learning frameworks used. Further, whereas we focused on a single modality (text), in the future, we plan to investigate the application of multi-task learning (explored in chapter 6) along with logic-based rules derived from expert-curated knowledge bases for learning from multi-modal data in the domain of healthcare for multi-modal abstractive summarization. To achieve this, we aim to investigate multi-modal neuro-symbolic modeling in conjunction with the Mixture-of-Experts (MoE) architecture [154, 155, 156].

Further, even though different evaluation metrics have been investigated and used in this study for evaluating the quality of summaries generated, these metrics are still limited in terms of whether semantics and factual accuracy are respected in summarization. The factual accuracy metrics quantify entity-level precision and recall with respect to source articles or ground truth summaries. A plausible extension to these metrics is to evaluate how relationships between a pair of named entities are preserved during summary generation. Similarly, while we employ semantic matching (equivalence) metric using domain- and task-specific transformer encoder-decoder models, a potential avenue of deep semantic-based metrics can be explored that reflect whether pragmatics, in addition to semantics are

respected as well.

Bibliography

[1] ChengXiang Zhai and Sean Massung. *Text data management and analysis: a practical introduction to information retrieval and text mining*. Morgan & Claypool, 2016.

[2] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[3] Tim J Berners-Lee. The world-wide web. *Computer networks and ISDN systems*, 25(4-5):454–459, 1992.

[4] Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. *Advances in automatic text summarization*, pages 111–121, 1999.

[5] William John Hutchins. The concept of â€˜aboutnessâ€™ in subject indexing. In *Aslib proceedings*. MCB UP Ltd, 1978.

[6] John M Giorgi and Gary D Bader. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, 36(1):280–286, 2020.

[7] Mourad Gridach. Character-level neural network for biomedical named entity recognition. *Journal of biomedical informatics*, 70:85–91, 2017.

[8] Hyejin Cho and Hyunju Lee. Biomedical named entity recognition using deep neural networks with contextual information. *BMC bioinformatics*, 20(1):1–11, 2019.

[9] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*, 2020.

[10] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246, 2015.

[11] Paul Over and James Yen. An introduction to duc-2004. *National Institute of Standards and Technology*, 2004.

[12] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

[13] Phyllis B Baxendale. Machine-made index for technical literature—an experiment. *IBM Journal of research and development*, 2(4):354–361, 1958.

[14] Harold P Edmundson. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285, 1969.

[15] Udo Hahn and Ulrich Reimer. Computing text constituency: An algorithmic approach to the generation of text graphs. In *Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 343–368, 1984.

[16] Lisa F Rau, Paul S Jacobs, and Uri Zernik. Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management*, 25(4):419–428, 1989.

[17] Kathleen McKeown and Dragomir R Radev. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82, 1995.

[18] Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic text structuring and summarization. *Information processing & management*, 33(2):193–207, 1997.

[19] Dragomir R Radev. Generating natural language summaries from multiple on-line sources. 1997.

[20] Mandar Mitra, Amit Singhal, and Chris Buckley. Automatic text summarization by paragraph extraction. In *Intelligent Scalable Text Summarization*, 1997.

[21] Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. Summarization evaluation methods: Experiments and analysis. In *AAAI symposium on intelligent summarization*, pages 51–59. Palo Alto, CA, 1998.

[22] Regina Barzilay, Kathleen McKeown, and Michael Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 550–557, 1999.

[23] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[24] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.

[25] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[26] Regina Barzilay and Kathleen R McKeown. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328, 2005.

[27] Katja Filippova. Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*, pages 322–330, 2010.

[28] Florian Boudin and Emmanuel Morin. Keyphrase extraction for n-best reranking in multi-sentence compression. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2013.

[29] Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. Multi-document abstractive summarization using ilp based multi-sentence compression. *arXiv preprint arXiv:1609.07034*, 2016.

[30] Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, 2018.

[31] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[32] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[33] Sumit Chopra, Michael Auli, and Alexander M Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, 2016.

[34] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

[35] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[37] Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K Reddy. Neural abstractive text summarization with sequence-to-sequence models. *arXiv preprint arXiv:1812.02303*, 2018.

[38] Chandra Khatri, Gyanit Singh, and Nish Parikh. Abstractive and extractive text summarization using document context vector and recurrent neural networks. *arXiv preprint arXiv:1807.08000*, 2018.

[39] Eric Chu and Peter Liu. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232, 2019.

[40] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

[41] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019.

[42] Yang Liu and Mirella Lapata. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, 2019.

[43] Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*, 2019.

[44] Kundan Krishna and Balaji Vasan Srinivasan. Generating topic-oriented summaries using neural attention. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1697–1705, 2018.

[45] Kexin Liao, Logan Lebanoff, and Fei Liu. Abstract meaning representation for multi-document summarization. *arXiv preprint arXiv:1806.05655*, 2018.

[46] Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A Smith. Toward abstractive summarization using semantic representations. *arXiv preprint arXiv:1805.10399*, 2018.

[47] Shibhansh Dohare, Vivek Gupta, and Harish Karnick. Unsupervised semantic abstractive summarization. In *Proceedings of ACL 2018, Student Research Workshop*, pages 74–83, 2018.

[48] Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. Multi-document abstractive summarization using ilp based multi-sentence compression. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[49] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*, 2018.

[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[51] Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, 2015.

[52] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[53] Samaneh Karimi, Luis Moraes, Avisha Das, Azadeh Shakery, and Rakesh Verma. Citance-based retrieval and summarization using ir and machine learning. *Scientometrics*, 116(2):1331–1366, 2018.

[54] Preslav I Nakov, Ariel S Schwartz, and Marti Hearst. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR*, volume 4, pages 81–88. Citeseer, 2004.

[55] Chrysoula Zerva, Minh-Quoc Nghiem, Nhung TH Nguyen, Sophia Ananiadou, et al. Cited text span identification for scientific summarisation using pre-trained encoders. *Scientometrics*, pages 1–29, 2020.

[56] Amjad Abu-Jbara and Dragomir Radev. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 500–509, 2011.

[57] Arman Cohan and Nazli Goharian. Scientific article summarization using citation-context and article's discourse structure. *arXiv preprint arXiv:1704.06619*, 2017.

[58] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7386–7393, 2019.

[59] Tarek Saier and Michael Färber. Bibliometric-enhanced arxiv: A data set for paper-based and citation-based tasks. In *BIR@ ECIR*, pages 14–26, 2019.

[60] Arman Cohan and Nazli Goharian. Contextualizing citations for scientific summarization using word embeddings and domain knowledge. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1133–1136, 2017.

[61] Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*, 2014.

[62] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*, 2019.

[63] Wei-Cheng Chang, Felix X Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. Pre-training tasks for embedding-based large-scale retrieval. *arXiv preprint arXiv:2002.03932*, 2020.

[64] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[65] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

[66] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[67] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.

[68] Yu Yan, Weizhen Qi, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*, 2020.

[69] Hao Zhou, Weidong Ren, Gongshen Liu, Bo Su, and Wei Lu. Entity-aware abstractive multi-document summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 351–362, 2021.

[70] Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. An entity-driven framework for abstractive summarization. *arXiv preprint arXiv:1909.02059*, 2019.

[71] Frederik Schulze and Mariana Neves. Entity-supported summarization of biomedical abstracts. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 40–49, 2016.

[72] Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. Enhancing factual consistency of abstractive summarization. *arXiv preprint arXiv:2003.08612*, 2020.

[73] Beliz Gunel, Chenguang Zhu, Michael Zeng, and Xuedong Huang. Mind the facts: Knowledge-boosted coherent abstractive text summarization. *arXiv preprint arXiv:2006.15435*, 2020.

[74] Gaur Manas, Vamsi Aribandi, Ugur Kursuncu, Amanuel Alambo, Valerie L Shalin, Krishnaprasad Thirunarayan, Jonathan Beich, Meera Narasimhan, Amit Sheth, et al. Knowledge-infused abstractive summarization of clinical diagnostic interviews: Framework development study. *JMIR Mental Health*, 8(5):e20865, 2021.

[75] Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

[76] Muhammad Afzal, Fakhare Alam, Khalid Mahmood Malik, and Ghaus M Malik. Clinical context–aware biomedical text summarization using deep neural network: Model development and validation. *Journal of medical Internet research*, 22(10):e19810, 2020.

[77] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*, 2018.

[78] Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486*, 2021.

[79] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[80] Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*, 2020.

[81] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701, 2015.

[82] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[83] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. Tldr: Extreme summarization of scientific documents. *arXiv preprint arXiv:2004.15011*, 2020.

[84] Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W Filice. Ontology-aware clinical abstractive summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1013–1016, 2019.

[85] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[86] Curtis P Langlotz. Radlex: a new method for indexing online educational materials, 2006.

[87] Sajad Sotudeh, Nazli Goharian, and Ross W Filice. Attend to medical ontologies: Content selection for clinical abstractive summarization. *arXiv preprint arXiv:2005.00163*, 2020.

[88] Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. A novel system for extractive clinical note summarization using ehr data. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, 2019.

[89] Wei-Hung Weng, Yu-An Chung, and Schrasing Tong. Clinical text summarization with syntax-based negation and semantic concept identification. *arXiv preprint arXiv:2003.00353*, 2020.

[90] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[91] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[92] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[93] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961. Association for Computational Linguistics, 2012.

[94] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[95] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[96] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

[97] Hao Zheng and Mirella Lapata. Sentence centrality revisited for unsupervised summarization. *arXiv preprint arXiv:1906.03508*, 2019.

[98] Mengyun Cao and Hai Zhuge. Grouping sentences as a better language unit for extractive text summarization. *Future Generation Computer Systems*, 2020.

[99] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*, 2017.

[100] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[101] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998.

[102] Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 596–606, 2013.

[103] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[104] Xiang Lin, Simeng Han, and Shafiq Joty. Straight to the gradient: Learning to use novel tokens for neural text generation. *arXiv preprint arXiv:2106.07207*, 2021.

[105] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[106] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, 2020.

[107] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[108] Bam Bahadur Kadayat and Evelyn Eika. Impact of sentence length on the readability of web for screen reader users. In *International Conference on Human-Computer Interaction*, pages 261–271. Springer, 2020.

[109] Dimo Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

[110] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[111] Travis R Goodwin, Max E Savery, and Dina Demner-Fushman. Flight of the pegasus? comparing transformers on few-shot and zero-shot multi-document abstractive summarization. In *Proceedings of COLING. International Conference on Computational Linguistics*, volume 2020, page 5640. NIH Public Access, 2020.

[112] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*, 2017.

[113] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100, 2018.

[114] Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. Improving abstraction in text summarization. *arXiv preprint arXiv:1808.07913*, 2018.

[115] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[116] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, 2014.

[117] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

[118] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*, 2019.

[119] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[120] Logan Lebanoff, Kaiqiang Song, and Fei Liu. Adapting the neural encoder-decoder framework from single to multi-document summarization. *arXiv preprint arXiv:1808.06218*, 2018.

[121] Jianmin Zhang, Jiwei Tan, and Xiaojun Wan. Towards a neural network approach to abstractive multi-document summarization. *arXiv preprint arXiv:1804.09010*, 2018.

[122] Yang Gao, Wei Zhao, and Steffen Eger. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. *arXiv preprint arXiv:2005.03724*, 2020.

[123] Lamy Jean-Baptiste. Using medical terminologies with pymedtermino and umls. In *Ontologies with Python*, pages 207–239. Springer, 2021.

[124] Anshumali Shrivastava and Ping Li. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). *arXiv preprint arXiv:1405.5869*, 2014.

[125] Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simoes, Vitaly Nikolaev, and Ryan McDonald. Planning with learned entity prompts for abstractive summarization. *arXiv preprint arXiv:2104.07606*, 2021.

[126] Chenxin An, Ming Zhong, Zhichao Geng, Jianqiang Yang, and Xipeng Qiu. Retrievalsum: A retrieval enhanced framework for abstractive summarization. *arXiv preprint arXiv:2109.07943*, 2021.

[127] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

[128] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020.

[129] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*, 2020.

[130] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020.

[131] Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34, 2021.

[132] John M Prager. Open-domain question-answering. *Found. Trends Inf. Retr.*, 1(2):91–231, 2006.

[133] Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*, 2020.

[134] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[135] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[136] Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. Gsum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*, 2020.

[137] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[138] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[139] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. In *NeurIPS*, 2020.

[140] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.

[141] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157, 2003.

[142] Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. Entity-level factual consistency of abstractive text summarization. *arXiv preprint arXiv:2102.09130*, 2021.

[143] Rimma Pivovarov and Noémie Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, 2015.

[144] Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. Learning to summarize radiology findings. *arXiv preprint arXiv:1809.04698*, 2018.

[145] Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D Manning, and Curtis P Langlotz. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. *arXiv preprint arXiv:1911.02541*, 2019.

[146] Wang Haonan, Gao Yang, Bai Yu, Mirella Lapata, and Huang Heyan. Exploring explainable selection to control abstractive summarization. *arXiv preprint arXiv:2004.11779*, 2020.

[147] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[148] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[149] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

[150] Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 166–175, 2019.

[151] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[152] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[153] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

[154] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[155] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[156] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1930–1939, 2018.