UNIVERSITY OF POTSDAM

MASTERS THESIS

Abstractive Summarization of Scientific Documents Using Argument Structure

---

Author:
Jatin Karthik TRIPATHY

1st Supervisor:
Arne BINDER

2nd Supervisor:
Prof. Dr. Manfred STEDE

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in Cognitive Systems

May 8, 2024

Declaration of Authorship

I, Jatin Karthik TRIPATHY, declare that this thesis titled, "Abstractive Summarization of Scientific Documents Using Argument Structure" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 08.05.2024

UNIVERSITY OF POTSDAM

Abstract

Faculty of Human Sciences

Department of Linguistics

Master of Science in Cognitive Systems

Abstractive Summarization of Scientific Documents Using Argument Structure

by Jatin Karthik TRIPATHY

This study explores the integration of additional contextual information, particularly argument structure, to enhance abstractive summarization methodologies. Our research leverages argumentative discourse units (ADUs) within the Longformer-Encoder-Decoder (LED) architecture to provide nuanced contextual cues essential for producing coherent and contextually relevant summaries. Through meticulous exploration of ADUs and their diverse parsing methods as effective guidance signals, empirical findings highlight the superior performance of the fine-tuned LED model on ADUs, surpassing benchmarks set by the state-of-the-art LED baseline. Moreover, our investigations underscore the critical importance of alignment between the guidance signal and the source text, with experiments demonstrating superior performance outcomes when alignment is enhanced. Overall, this study contributes to advancing the state-of-the-art in abstractive summarization by 3 to 4 points in the ROUGE metric, enabling more accurate, informative, and contextually nuanced summarization outcomes.

UNIVERSITÄT POTSDAM

Zusammenfassung

Humanwissenschaftliche Fakultät

Department für Linguistik

Master of Science in Cognitive Systems

Abstractive Summarization of Scientific Documents Using Argument Structure

by Jatin Karthik TRIPATHY

Diese Studie untersucht die Integration zusätzlicher kontextueller Informationen, insbesondere der Argumentationsstruktur, zur Verbesserung abstraktiver Zusammenfassungsmethoden. Unsere Forschung nutzt argumentative Diskurseinheiten (ADUs) in der Longformer-Encoder-Decoder (LED) Architektur, um nuancierte kontextuelle Hinweise zu liefern, die für die Erstellung kohärenter und kontextuell relevanter Zusammenfassungen unerlässlich sind. Durch die sorgfältige Erforschung von ADUs und ihren verschiedenen Parsing-Methoden als effektive Orientierungssignale zeigen die empirischen Ergebnisse die überlegene Leistung des fine-tuned LED-Modells bei ADUs und übertreffen die Benchmarks, die von der State-of-the-Art-LED-Baseline gesetzt wurden. Darüber hinaus unterstreichen unsere Untersuchungen die kritische Bedeutung des Alignments zwischen dem Orientierungssignal und dem Ausgangstext, wobei die Experimente zeigen, dass bei verbessertem Alignment bessere Leistungsergebnisse erzielt werden. Insgesamt trägt diese Studie dazu bei, den Stand der Technik im Bereich der abstraktiven Zusammenfassung um 3 bis 4 Punkte in der ROUGE-Metrik zu verbessern, was genauere, aussagekräftigere und kontextuell differenziertere Zusammenfassungsergebnisse ermöglicht.

Contents

List of Figures

List of Tables

List of Abbreviations

ADU	**A**rgumentative **D**iscourse **U**nits
LED	**L**ongformer **E**ncoder **D**ecoder
NLP	**N**atural **L**anguage **P**rocessing
ROUGE	**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation

Chapter 1

Motivation

The primary focus of this research is to investigate the integration of argumentative structure information as a means to elevate the quality and effectiveness of abstractive summarization techniques. By incorporating context from argumentation theory, this study seeks to leverage the inherent structure of arguments within a text to enhance the coherence, comprehensiveness, and relevance of generated summaries using the ROUGE metric.

> Our hypothesis posits that integrating argumentative structure could improve the generated abstractive summaries, ensuring a logical flow and coherence that mirrors the underlying discourse of the source document.

Moreover, integrating argumentative structure is expected to mitigate the risk of critical points being overlooked or omitted in the generated summaries. By leveraging the inherent structure of arguments, the summarization model can prioritize and incorporate key arguments and supporting evidence.

In essence, this research aims to explore how the incorporation of argumentative structure information can serve as a guiding framework for abstractive summarization, enhancing both the structural coherence and content

fidelity of generated summaries. Through empirical investigation and experimentation, this study hopes to contribute to the advancement of abstractive summarization techniques, ultimately facilitating more accurate, informative, and contextually relevant text summarization solutions.

Chapter 2

Introduction

Abstractive summarization is a sophisticated facet of Natural Language Processing (NLP) that distils extensive textual data into concise and coherent summaries. Unlike extractive summarization methods, which merely select verbatim sentences or phrases from the source material, abstractive summarization delves deeper into context comprehension(Giarelis, Mastrokostas, and Karacapilidis, 2023; Zhang et al., 2022; Etemad, Abidi, and Chhabra, 2021).

This approach tries to emulate a more human-like comprehension by selecting key sentences, rephrasing, and restructuring content based on a nuanced understanding of the topic. Consequently, the generated summaries are not limited to the original wording, resulting in coherent, informative, and contextually appropriate outlines.

This thesis delves into the relationship between argument structure and abstractive summarization, exploring how leveraging Argumentative Discourse Units (ADUs) can enhance summarization quality. By incorporating ADUs extracted from the source text, we seek to enrich the summarization process with a deeper understanding of the underlying arguments, thereby facilitating the generation of more coherent and contextually relevant summaries.

The choice of scientific papers as the foundation for this research stems from their distinctive characteristics as extensive documents adhering to a standardized discourse structure. This standardized discourse structure helps with summarization since any approach does not need to consider any variation between the layouts of the documents. Moreover, the abstracts in the papers can be leveraged as a proxy for a gold-standard summary of the long-format document that is the scientific paper. Leveraging this established structure and the inherent proxy to summaries, this work aims to study long-format document summarization techniques.

This thesis hopes to bridge the gap between argumentative discourse analysis and abstractive summarization, offering insights into how integrating argument structure can enhance the quality and coherence of generated summaries. By leveraging the inherent structure of scientific papers and the nuanced relationships between arguments, we aspire to contribute to the advancement of abstractive summarization techniques, paving the way for more effective and contextually relevant text summarization solutions.

The layout of the thesis is structured into several sections, each addressing specific aspects of the research. In the Background section (Chapter 3), the foundational concepts of abstractive summarization, argument structure, Longformer, and GSum are discussed to provide the necessary context for the study. Following this, the Related Works section (Chapter 4) reviews existing literature on general abstractive summarization approaches and guiding abstractive summarization techniques.

Chapter 5 focuses on the Data used in the experiments, detailing the corpus and its characteristics. In Chapter 6, Methodology, the approach to using argument structure as guidance and modifications to the LED architecture are described. This chapter also addresses the alignment of the guidance signal with the text.

The Experimentation Setup (Chapter 7) outlines the specific configurations employed, including the choice of loss function, optimizer, metrics, experimental procedure, and operational setup. Chapter 8, Experimentation, presents the results of four experiments: Baseline LED, Fine-tuning on Argument Structure, Shared Embedding, and Shared Encoder. Each experiment builds upon the previous one to gain insights into effective strategies for abstractive summarization.

Chapter 9, Results and Discussion, delves into the findings of the experiments and discusses their implications. This chapter provides a comprehensive analysis of the performance of different configurations and their alignment with the research objectives. Finally, Chapter 10, Limitations and Future Work acknowledges the constraints of the study and outlines potential avenues for future research to address them.

Chapter 3

Background

This chapter will examine some core concepts and architectures this thesis builds upon. We will go over summarization, focusing more on abstraction summarization and argument structure, as these are the main theoretical aspects of this work. We also explain the Longformer architecture and its extension, the Longformer Encoder Decoder (LED) architecture, which we use and modify for abstractive summarization. Finally, we examine the GSum architecture, which helps us understand how adding more context can guide the generation of abstracts.

3.1 Abstractive Summarization

At its core, text summarization encompasses the task of condensing lengthy original texts or even collections of documents or paragraphs while preserving as much relevant information from within the source material as possible. There are two primary approaches to automatic text summarization: extractive and abstractive.

Extractive text summarization is accomplished by carefully selecting relevant phrases, paragraphs, or other textual units from the source material. These selected components are then put together to create a condensed version of the original text while preserving the semantics and coherence of the original

material.

The evaluation criteria used to select the relevant parts from the source text are significant in determining the effectiveness and quality of generated summaries. These criteria are benchmarks against which the summaries are evaluated regarding their coherence, relevance, and informativeness. These methods work towards generating concise summaries while providing important information that may be tailored to meet various informational requirements.

On the other hand, abstractive text summarization approaches the task in a slightly different direction. It aims to extract the core elements of the original information by detecting and capturing unique concepts, expressions, and ideas. Abstractive summarizing, thus, aims to provide summaries that compress information by generating novel phrases and expressions that may not be directly present in the original material.

There are numerous discernible aspects between abstractive summarization and its extractive equivalent. Extractive summarization focuses on selecting and rearranging existing text components, whereas abstractive summarization involves linguistic inventiveness and semantic comprehension. While extractive approaches tend to provide summaries that closely resemble the structure and phrasing of the original text, abstractive summarizing can generate summaries that show more abstraction and linguistic diversity.

Recent advancements in machine learning, specifically in deep learning architectures and natural language generation techniques, have made abstractive summarization a prominent focus of research and development (Giarelis, Mastrokostas, and Karacapilidis, 2023; Zhang et al., 2022; Etemad, Abidi, and Chhabra, 2021). However, both approaches have their advantages and challenges.

A point to note is the difference in linguistic complexity, as abstractive summarization requires a more thorough comprehension of language subtleties and context. Therefore, whereas extractive summaries may provide a simpler and more practical method, abstractive summaries can generate summaries that capture more profound insights and interpretations of the original material at the cost of being more complex.

3.2 Argument Structure

In linguistics and discourse analysis, argument structure pertains to the systematic organization and arrangement of elements within a speech or text to provide a logical and convincing argument (Walton, Reed, and Macagno, 2008).

> "Argument structure refers to the fundamental organization of ideas, the connections between claims, and rhetorical tactics to bolster a specific viewpoint or thesis." —Walton, Reed, and Macagno (2008)

Walton, Reed, and Macagno (2008) explain that the core elements of argument construction are premises, evidence, reasoning, and conclusions. Premises are fundamental assumptions that form the backbone of an argument, supporting later claims and assertions. Evidence comprises verifiable data, concrete illustrations, or authoritative viewpoints that validate the premises and enhance the argument's credibility. Reasoning connects premises and conclusions by explaining the logical procedures or inferential connections that relate the supporting evidence to the main argument. Conclusions contain the final claim or position of the argument, providing a concise summary of the primary point or critical message.

A compelling argument structure typically adheres to a discernible pattern, such as the traditional rhetorical framework comprising an introduction, body,

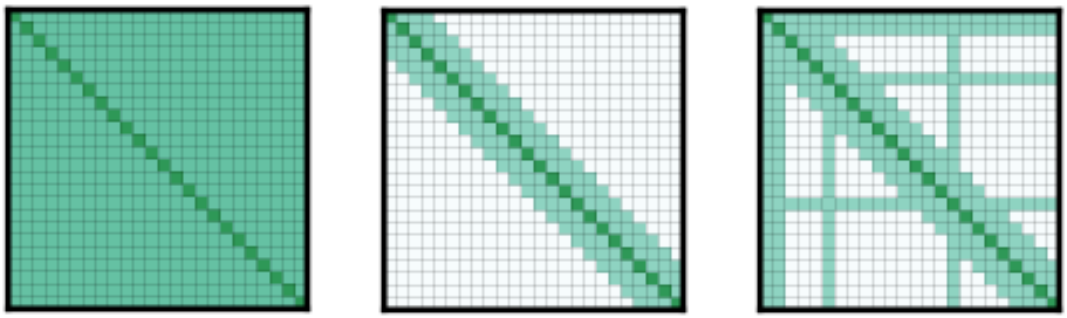
and conclusion. The introduction provides an initial framework by introducing the subject matter, providing the background, and stating the central thesis or argument. The main argument is further developed in the body of the text through a sequence of interrelated paragraphs or sections, each focusing on a particular component of the argument and presenting supporting evidence or logic. Paragraph transitions aid in the seamless progression of ideas and the maintenance of coherence, guaranteeing a logical advancement of the argument from one point to another. Finally, the conclusion strengthens the primary argument by succinctly summarizing crucial ideas.

Argumentative discourse units (ADUs) (Toulmin, 2003) are the essential elements or components of argumentation in a conversation. All text consists of these discrete ADUs, which contribute to its overall structure and coherence. ADUs can exhibit various dimensions and intricacy, encompassing everything from solitary assertions or premises to complete arguments or counterarguments.

ADUs are derived from the study of argumentation theory and discourse analysis. Work in this field aims to comprehend arguments' construction, presentation, and evaluation in communicative exchanges. By identifying and analyzing the components of argumentative speech, researchers can acquire valuable insights into the rhetorical methods, logical patterns, and persuasive approaches speakers or authors use to support their arguments.

3.3 Longformer

The Longformer (Beltagy, Peters, and Cohan, 2020) architecture extends the transformer architecture (Vaswani et al., 2017). Within a transformer model, the encoder-decoder structure (Devlin et al., 2019) effectively handles input sequences by selectively focusing on pertinent segments of the sequence at each layer. However, as the sequence length increases, the self-attention



(A) Full n^2 attention (B) Local attention (C) Global+Local attention

FIGURE 3.1: Types of attention
Figure sourced from Beltagy, Peters, and Cohan (2020)

mechanism's computational complexity increases exponentially, making it challenging to process long texts efficiently. The Longformer, however, cleverly resolves this problem by incorporating novel alterations to the conventional transformer architecture, allowing it to process lengthy sequences while preserving computing efficiency properly.

Its architecture is primarily characterized by its sparse attention mechanism, which efficiently handles lengthy sequences. Contrary to conventional transformers, the Longformer focuses on a specific group of tokens in the sequence rather than every token attending to every other token. Selective attention significantly decreases the computing load in processing lengthy sequences while maintaining the model's capacity to capture important contextual links.

The Longformer utilizes a local-global attention mechanism to capture local and global dependencies in extended sequences, shown in Figure 3.1. This technique partitions the input sequence into overlapping windows or segments and does attention calculations within each window. This sliding window technique guarantees that neighbouring windows exchange information, allowing the model to preserve contextual consistency across the page. Thus, Longformer can utilize the localized attention mechanism to efficiently collect intricate features inside each segment while considering wider

contextual relationships throughout the document.

To further enhance the capability of the Longformer, a combination of dilation and differing window sizes is employed. Various window sizes are applied across the layers, following the approach proposed by Sukhbaatar et al. (2019). Specifically, smaller window sizes are utilized in the lower layers, gradually increasing as the layers progress. This strategy enables the top layers to capture higher-level representations of the entire sequence while the lower layers focus on capturing local information, striking a balance between efficiency and performance.

Additionally, dilated sliding windows are implemented for the lower layers, with a small amount of increasing dilation applied to only two heads in the higher layers. This approach empowers the model to attend directly to distant tokens while preserving local context, thereby enhancing its ability to process long-range dependencies effectively.

Similar to previous transformer-based models, the Longformer employs positional encodings to represent the positioning information of tokens in the input sequence. Positional encodings facilitate the model in distinguishing tokens by their places in the sequence, enhancing its ability to capture sequential connections effectively. In addition, the Longformer employs specialized tokenization techniques that are optimized for effectively processing lengthy texts. These tokenization methods effectively represent textual inputs while minimizing computation costs, allowing the model to analyze lengthy texts effortlessly.

3.3.1 LED

The Longformer-Encoder-Decoder (LED) architecture presents a novel approach by integrating Transformer stacks in both the encoder and decoder,

focusing on optimizing attention mechanisms. Unlike conventional architectures, the encoder utilizes the local-global attention pattern inspired by the Longformer, which balances capturing long-range dependencies and maintaining computational efficiency. On the other hand, the LED decoder employs full self-attention to both all encoded tokens and previously decoded locations, enabling comprehensive information processing during the generation of summaries. This innovative design allows the model to effectively process large input sequences without compromising performance, making it well-suited for tasks such as text summarization.

A strategic initialization approach was adopted to address the computational demands of pre-training the LED model. LED parameters are initialized from BART (Lewis et al., 2020), a pre-trained model with a similar architecture, ensuring consistency in the number of layers and hidden sizes. This initialization strategy accelerates the training process and leverages the knowledge encoded within BART's parameters. As a result, the authors defined LED-base and LED-large architectures, each tailored to specific requirements, with 6 and 12 layers in both encoder and decoder stacks, respectively. This systematic approach to architecture design and parameter initialization sets the foundation for efficient and effective text summarization with the LED model.

3.4 GSum

A significant obstacle in abstractive summarization is the risk of generating summaries that are not accurate or faithful to the source text. These summaries may include factual mistakes, as well as imagined information that is not included in the original document. Any modifications from the original material might compromise the integrity and dependability of the summary, perhaps causing misconceptions or misinterpretations for the end-users.

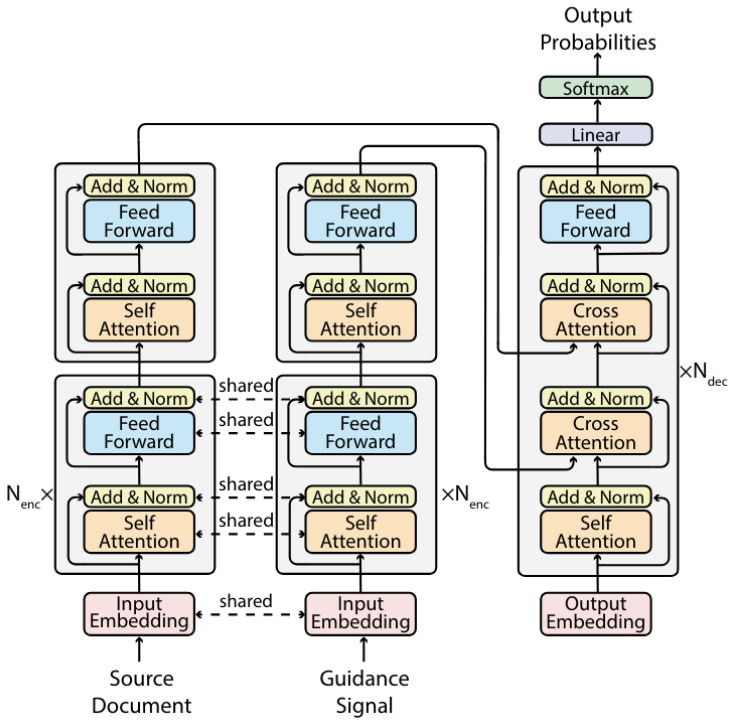


FIGURE 3.2: GSum Architecture
Figure sourced from Dou et al. (2021)

Dou et al. (2021) introduce a framework designed for guided neural summarization. Their work shows the synergistic nature of these guidance signals, revealing their complementary roles in enhancing the summarization process. Specifically, they observe that their models exhibit an increased capacity to generate novel words and produce summaries that adhere more closely to the original content. These findings underscore the versatility and effectiveness of their proposed approach in facilitating both improved summarization quality and increased user control over the process.

Their strategy involves employing distinct encoders — one specifically designed for the source material and another for handling various guide signals. The primary modification to the decoder is the addition of an extra cross-attention block, which allows it to focus efficiently on the output representations of the guiding encoder. The framework prioritizes these signals,

but its flexible structure enables the smooth incorporation of other forms of guidance signals, assuring versatility to accommodate various summarizing requirements and user preferences, shown in Figure 3.2.

The decoder first focuses on the guiding signals to create appropriate representations, which helps identify relevant characteristics in the source materials. After obtaining these guidance-aware representations, the decoder attends to the source material, utilizing the guidance signals to aid and contextualize the summarization process. The sequential attention technique guarantees that the created summaries align with the source text and use the information from the available guidance signals. Dou et al. (2021) show that their approach improves the overall quality and relevance of the summarization output.

Chapter 4

Related Works

This chapter will focus on other work that, while not completely similar to the goals of this thesis, still gives us an insight into how to approach our task. We primarily look into works of the two main aspects of this work —abstractive summarization and guiding abstractive summarization.

4.1 General Abstractive Summarization Approaches

Abstractive summarization techniques have garnered significant attention and research efforts owing to their potential to distil complex textual information into concise, coherent summaries. Prior studies have delved into various aspects of abstractive summarization, ranging from algorithmic advancements to linguistic insights.

Additionally, researchers have explored integrating external knowledge sources, such as ontologies and domain-specific databases, to enrich the summarization process. Understanding the evolution and nuances of these related works is instrumental in contextualizing our research endeavours and identifying avenues for innovation and refinement in the realm of abstractive summarization.

Cohan et al. (2018a)'s work was one of the first to tackle abstractive summarization for lengthy individual documents. They explore the use of discourse

attention by feeding attention based on section level as well as normal word-level attention. Although their method is relatively straightforward, they demonstrate the constructive influence of incorporating discourse structure on generating summaries from lengthy textual documents. While not exactly similar to the idea behind this thesis, this method of discourse structure can still be used to draw valuable insights.

ConvoSumm (Fabbri et al., 2021) is another pioneering benchmark featuring new conversation datasets developed via crowdsourcing. This benchmark, incorporating widely used datasets and advanced baselines, drove progress in abstractive summarization beyond news articles. Their study evaluates baseline models, integrating the text's argument structure to enhance viewpoint quantification in non-linear input. They also address challenges in modelling relevance and consistency, something this work also focuses on, in abstractive conversation summarization compared to traditional news summarization.

Roush and Balaji (2020) introduced DebateSum, a comprehensive argument mining and summarization dataset. Each entry in DebateSum comprises debate evidence, a word-level extractive summary of that evidence, and the corresponding argument. Their work provides insight in regards to using some form of extractive summary for the overall goal of abstraction summarization.

The work done by Zhu et al. (2021) is also interesting to consider due to their use of knowledge graphs for better summarization. They extract factual information from the article, present it in a knowledge graph, and integrate these graphs using graph neural networks. The resulting model excels in preserving facts during summarization, as demonstrated by evaluations conducted both automatically and by human assessors.

Jin, Wang, and Wan (2020)'s work is also similar enough to consider for the goals of this thesis. They investigate the integration of semantic dependency

graphs into abstractive summarization models. Their proposed model, SemSUM, utilizes information from original input texts and corresponding semantic dependency graphs to enhance the abstractive summarization process. The results indicate that their approach improves semantic relevance, leading to higher-quality summaries.

4.2 Guiding Abstractive Summarization

It can be challenging to predict or determine the specific elements of the original text that an abstractive system may prioritize or incorporate into the summary. If some form of control or guidance is absent, it impedes the capacity to customize the summary according to unique requirements or preferences.

A solution to the challenges mentioned above is to add extra context to the abstractive summarizer in the form of a guidance signal (Kikuchi et al., 2016; Li et al., 2018; Cao et al., 2018; Dou et al., 2021). For example, guiding signals can restrict generating a summary, guaranteeing that the resulting material stays faithful to the original text and reduces any deviations. Moreover, guiding signals can offer a method of control, enabling users to impact the content and structure of the generated summaries. Researchers seek to enhance the quality, accuracy, and practicality of produced summaries in various applications and scenarios by incorporating these guiding mechanisms into abstractive summarizing algorithms.

An approach introduced by Kikuchi et al. (2016) focuses on providing direction by setting the desired length of abstractive summaries. Researchers can guide the summarizing model to provide brief, informative summaries that match user preferences or application requirements by enforcing predetermined length limitations. Kikuchi et al. (2016) believe that this method ensures that the generated summaries remain concise while effectively capturing the crucial information from the source material.

Li et al. (2018) proposes a methodology that enhances models by using keywords to guide the process of summarization. The model acquires more context and guidance by including essential terms or phrases from the original material, reducing the likelihood of missing critical information in the final summaries. Li et al. (2018) say this method, which focuses on keywords, improves the model's ability to identify and prioritize important material, increasing the summaries' overall quality and relevancy.

On the other hand, Cao et al. (2018) provides a different approach that utilizes a retrieval and reference mechanism to direct abstractive summarization models. Their approach involves extracting relevant summaries from the training set and using them as references during summarizing. By extracting knowledge from preexisting summaries that capture comparable material or topics, the model acquires valuable reference points to guide its summary choices. Cao et al. (2018) show that their guiding mechanism, which focuses on reference, promotes coherence and consistency in the generated summaries. It ensures that the summaries correspond with known patterns and norms in the training data.

Although these strategies have demonstrated improvements in the quality and controllability of summarization, each focuses on unique types of assistance. Therefore, it is still unclear which strategy is preferable and whether they have complementing qualities.

Chapter 5

Data

The SciArg dataset (Lauscher, Glavaš, and Ponzetto, 2018) extends the DrInventor Argumentative Corpus with an annotation layer containing fine-grained argumentative components and relations.

This thesis uses the argumentative components within DrInventor Argumentative Corpus as our Argumentative Discourse Units (ADUs). These components serve as foundational elements for our analysis, facilitating the exploration of argumentative structures and their use as extra context for abstraction summarization. Given that there is no predefined split, we simply opt for a standard 70-30 split for the dataset.

Class	**Instances**
Data	4093
Own Claim	5445
Background Claim	2751

TABLE 5.1: Argumentative Component Classes

The DrInventor Argumentative Corpus (DrInventor) is an expanded iteration of the Dr Inventor corpus (Fisas, Saggion, and Ronzano, 2015; Fisas, Ronzano, and Saggion, 2016), featuring an annotation layer enriched with argumentative components and relations. Comprising 40 scientific publications from computer graphics, DrInventor encompasses approximately 12,000 labelled argumentative components, alongside annotations for other relevant

tasks.

The argumentative components within DrInventor are categorized into three classes: DATA, OWN CLAIM, and BACKGROUND CLAIM, as shown in Table 5.1. DATA pertains to factual information, while OWN CLAIM represents assertions made by the authors, and finally, BACKGROUND CLAIM denotes claims grounded in background knowledge, often referencing prior works by other authors.

Relation	**Instances**
Supports	5790
Contradicts	696
Semantically Same	44

TABLE 5.2: Argumentative Relations

Additionally, DrInventor delineates relation classes, including SUPPORTS, CONTRADICTS and SEMANTICALLY SAME. These relations capture the connections between argumentative components, explaining their interplay within the discourse.

The complexity of discourse structure in DrInventor, alongside instances of component fragmentation across non-contiguous parts of documents, presents significant challenges for argument mining. Over 1,000 instances exhibit split components, predominantly affecting claims but extending to data in fewer cases.

Furthermore, the unbalanced distribution among the three classes and the presence of split components exacerbates the difficulty of link prediction within the corpus. This challenge is made apparent by the low inter-annotator agreement reported in the original study, emphasizing the intricacies inherent in analyzing argumentative discourse within scientific publications.

However, for the goals of this thesis, these challenges do not cause any complications. Argument relations are not used in any form to provide extra context for abstraction summarization, and while the classes are very imbalanced, we care more about the actual argument component itself rather than what class it fits into. As stated above, we directly use the text of each argument component as our ADUs, thus mitigating most of the issues this dataset presents.

Chapter 6

Methodology

The methodology section describes the setups used to evaluate the performance of text summarization on the SciArg dataset.
We also explain how we use and modify the LED Large pretrained model [1], which has been trained on the arXiv dataset, to create summaries for scientific publications.

6.1 Using Argument Structure as Guidance

We introduce a few methods designed to provide varying levels of granularity and contextual information derived from the ADUs, thereby enriching the summarization process and enhancing the coherence and relevance of the generated summaries.
Three distinct methods for incorporating ADUs into the summarization process, namely SPAN TEXTS, ANNOTATED SPAN TEXTS, and ANNOTATED SOURCE TEXT.

SPAN TEXTS: In this approach, we extract the text content of each ADU and concatenate them with spaces to form continuous spans of text, Figure 6.1. This will let us see if the ADUs can be used like Zhao, Saleh, and Liu (2020)'s work, where extractive summarization is done first before attempting abstractive summarization. This method captures the essential content within

[1] https://huggingface.co/allenai/led-large-16384-arxiv

each ADU while maintaining a simplified representation without additional annotations or markup.
The SPAN TEXTS technique provides a simple form that effectively captures the core of each ADU without adding unnecessary complexity.

> ...Several recent major movie releases have demonstrated that the motion of clothing adds greatly to the appearance of a virtual character This effect is particularly compelling for scenes that include both real and synthetic actors those with Yoda and Anakin Skywalker in Episode II: Attack of the Clones the virtual clothing must move and be rendered so that it...

FIGURE 6.1: Example of Span Texts

ANNOTATED SPAN TEXTS: here, we build upon the SPAN TEXTS approach by introducing special tokens, `<ADU>` and `<\ADU>` to denote the beginning and end of each ADU span, Figure 6.2. These annotations serve to delineate the boundaries of individual ADUs within the concatenated text, providing a more structured representation while preserving the original content of each ADU.
The ANNOTATED SPAN TEXTS technique uses structured annotations to clearly define the borders of distinct argumentation Discourse Units (ADUs), which improves the model's capacity to recognize and include argumentation structure in the summaries.

> ...<ADU> Realistic human motion is an important part of media like video games and movies </ADU> <ADU> More lifelike characters make for more immersive environments and more believable special effects </ADU> <ADU> realistic animation of human motion is a challenging task </ADU> <ADU> people have proven to be adept at discerning the subtleties of human movement and identifying...

FIGURE 6.2: Example of Annotated Span Texts

ANNOTATED SOURCE TEXT: we augment the full source text with the same special tokens used in the ANNOTATED SPAN TEXTS method, Figure 6.3. By directly embedding these annotations within the source text, we hope to ensure that no information about the context of the ADUs is lost during the

summarization process.
The ANNOTATED SOURCE TEXT technique achieves the maximum level of integration by including ADU annotations directly into the source text.

> ...Where <ADU>visual fidelity is of the utmost importance</ADU>, with respect to <ADU>film quality animation</ADU>, <ADU>a combination of techniques including muscle simulation is used to achieve the realistic best in mesh deformation</ADU>. The attachment of mesh geometry to the underlying skeleton rig is called 'skinning' and <ADU>this can be understood as a function mapping of the skeleton parameters to a deformation field</ADU>. 1 One of the common skinning methods in...

FIGURE 6.3: Example of Annotated Source Texts

Each parsing approach has unique benefits and compromises regarding the level of detail and contextual information it provides to the summarization model. Once parsed, table 6.1 shows the differing text lengths.

	Type	**Min**	**Max**	**Mean**	**Standard Deviation**
Train	Full Text	4218	11454	7259	1950
	Abstract	92	444	171	69
	Span Texts	1356	5581	3158	908
	Annotated Span Texts	1665	6257	3717	1063
	Annotated Source Texts	4727	12313	7661	1843
Test	Full Text	4382	10518	6996	2179
	Abstract	88	220	155	46
	Span Texts	1365	4197	2569	910
	Annotated Span Texts	2292	6527	3969	1012
	Annotated Source Texts	4466	11271	7448	2330

TABLE 6.1: Comparison of different Guidance Signals along with the Full Text and Abstracts

6.2 Modifying LED

In this section of the methodology, we propose modifications to the LED architecture to effectively integrate guidance from the argument structure into

the text summarization process. This thesis primarily builds on the GSum architecture but makes a few modifications to accelerate training. We achieve this by leveraging pre-trained weights and avoiding training the entire architecture from scratch. This saves training time and improves the final model's performance by incorporating knowledge from the pre-trained weights.

Drawing inspiration from the GSum architecture, we introduce a separate LED encoder to process the guidance input derived from ADUs. This dual encoder setup allows for the simultaneous processing of both the source text and the guidance input, enabling the model to leverage contextual information from argument structure while generating summaries. We explore two distinct approaches for implementing this modification: SHARED EMBEDDING and SHARED ENCODER. Each offers unique advantages in enhancing summarization quality.

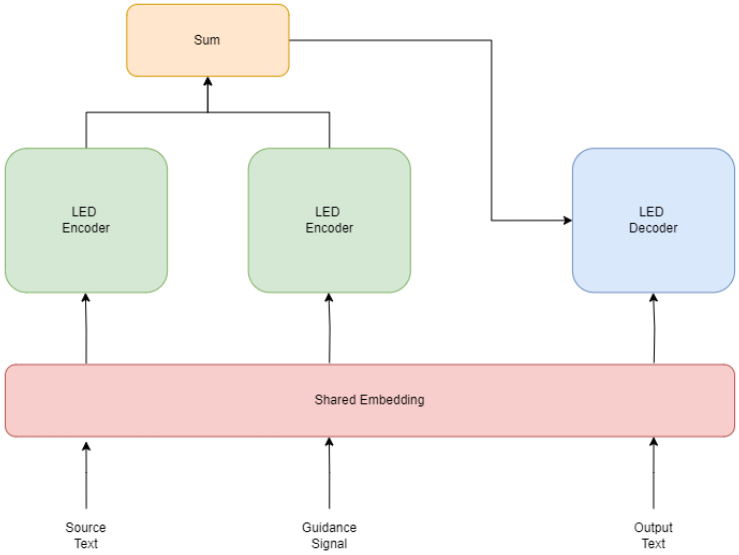


FIGURE 6.4: AbsArgSumm Shared Embedding Architecture

SHARED EMBEDDING: In this approach, we extend the existing LED architecture by sharing the embedding layer between the source text encoder and the guidance input encoder, Figure 6.4. This setup mirrors the standard LED architecture, where the encoder and decoder commonly share the embedding

layer. By sharing the embedding layer for both the source text and the guidance input, we enable the model to incorporate contextual information from the argument structure into the summarization process without introducing additional complexity.

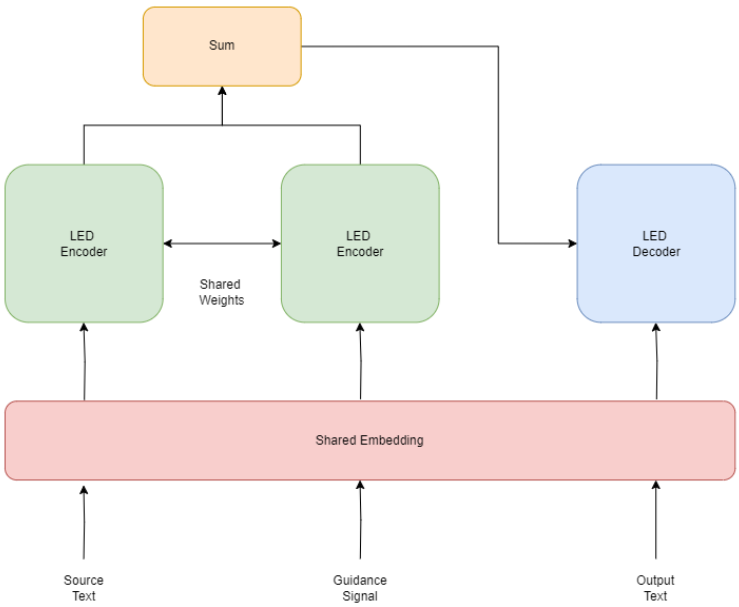


FIGURE 6.5: AbsArgSumm Shared Encoder Architecture

SHARED ENCODER: Building upon the SHARED EMBEDDING approach, we further extend the modification by sharing not only the embedding layer but also the encoder weights between the source text encoder and the guidance input encoder, Figure 6.5. This approach aligns more closely with the GSum architecture, where the encoder weights are shared across the source text and the guidance input. By sharing the encoder weights, we encourage the model to attend to both the source text and the guidance input simultaneously, facilitating a more comprehensive understanding of the underlying discourse and potentially leading to higher-quality generated summaries.

Since the traditional LED architecture does not allow extra input to be passed through, we modified it to allow leveraging guidance from the argument

structure. By introducing a separate LED encoder dedicated to processing the guidance input derived from ADUs, we aim to enrich the model's understanding of the underlying discourse and improve the coherence and relevance of the generated summaries.

The two proposed approaches offer distinct advantages in terms of implementation complexity and computational efficiency, with SHARED ENCODER expected to yield superior results by encouraging simultaneous attention to both the source text and the guidance input. We believe modifying the LED architecture to incorporate guidance from the argument structure represents a pivotal step in enhancing text summarization performance and is the main goal of this work.

6.3 Alignment of the Guidance Signal

An essential aspect of the modified architecture is its alignment mechanism, which governs the relationship between the guidance signal and the source input text. In both configurations of the modified architecture, a fundamental step involves combining the outputs from each encoder. This involves adding the information extracted from the source input text derived from the guidance signal. This addition process operates similarly to additive pooling, where the values of each input or feature dimension-wise sum up the representations obtained from different sources.

Chapter 7

Experimentation Setup

Our experiment setup is meticulously designed to ensure the reliability and stability of the proposed architecture. We define our loss function, opting for the widely-used cross-entropy loss. Complementing this, we employ AdamW as our optimizer for its adaptive learning rate capabilities and monitoring performance metrics, primarily ROUGE scores. We utilize high-performance computing resources for efficient model training and adhere to best practices in experimental design and documentation to maintain integrity and reproducibility.

7.1 Loss Function - Cross Entropy

The cross-entropy loss function (Mao, Mohri, and Zhong, 2023) is crucial for assessing the difference between predicted probability and actual class labels in classification tasks. By integrating cross-entropy loss into the optimization procedure, we aim to steer the model towards error minimization and accuracy enhancement by imposing penalties for departures from the ground truth. By utilizing the mathematical expression of cross-entropy loss, which entails calculating the negative logarithm of projected probabilities, we can adequately account for the uncertainty linked to token predictions and incentivize the model to produce more confident and precise classifications.

7.2 Optimizer - AdamW

AdamW (Loshchilov and Hutter, 2017) is an enhanced version of the Adam optimizer that integrates weight decay into its optimization process. This integration allows for better overfitting control and enhances the overall generalization performance. AdamW incorporates weight decay regularisation into Adam's parameter update process, which promotes the model to keep lower weights. This weight decay helps to reduce the likelihood of overfitting the training data. This regularisation method effectively penalizes high parameter values, encouraging a more streamlined optimization process and improving the model's capacity to generalize to unfamiliar data. In addition, AdamW preserves the benefits of the original Adam optimizer, including adjustable learning rates and momentum, which provide efficient and reliable convergence during training.

7.3 Metrics - ROUGE

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) is a commonly employed collection of exceptionally crafted criteria for evaluating the quality of summaries by comparing them to reference summaries or human-generated gold standards. ROUGE-N and ROUGE-L are measurements used to evaluate the efficacy and authenticity of summarization outputs. ROUGE-N analyses the overlap of n-grams between the generated summary and reference summaries, while ROUGE-L evaluates the longest common sub-sequence. ROUGE metrics are highly beneficial for evaluating produced summaries' informativeness, coherence, and relevance. They are essential for academics and practitioners in natural language processing.

We can evaluate how well our algorithms capture essential information and preserve semantic accuracy by calculating ROUGE ratings for our produced summaries compared to human reference summaries. In addition, ROUGE

metrics enable us to assess and evaluate the effectiveness of various summarising methodologies, aiding in identifying optimal procedures and parameters for attaining superior summary outcomes.

In our evaluation of text summarization models, we emphasize ROUGE-L over other ROUGE-N metrics due to its distinct capability to assess the overlap of the longest common subsequences between the generated and the reference summaries. While ROUGE-N metrics, such as ROUGE-1 and ROUGE-2, primarily focus on the overlap of n-grams at the token level, ROUGE-L considers the summaries' overall structure and coherence by highlighting the longest common subsequences. This emphasis makes ROUGE-L particularly suitable for evaluating the fluency and coherence of summaries, which are essential aspects of summarization quality.

By prioritizing ROUGE-L in our evaluation framework, we aim to underscore the importance of summarization models' ability to produce summaries that closely mirror the structure and content of the reference summaries, thereby offering a more comprehensive evaluation of their performance. Furthermore, ROUGE-L is less sensitive to minor variations in the generated summaries, enhancing its robustness for evaluating summarization systems across diverse domains and datasets. Overall, our focus on ROUGE-L enables us to gain a holistic insight into the summarization models' effectiveness in capturing essential information and coherence from the source documents.

7.4 Experimental Procedure

To assess the reliability of our model's performance, we carry out five runs, each yielding a distinct trained model, ensuring that they all use the same training data, hyperparameters, and initialization conditions. After completing the training process, we assess each trained model's performance on a

standardized test set using the selected evaluation metrics. To determine the overall stability of the model, we compute the mean scores achieved by the five trained models for each evaluation measure of interest. This method enables us to reduce the impact of random initialization and training variability, resulting in a more dependable assessment of the model's performance.

We can thoroughly evaluate the model's consistency and capacity by combining the outcomes from several iterations. This allows us to make reliable judgements about its effectiveness. This thorough testing process guarantees that our findings are not too dependent on the peculiarities of a single training session, hence improving the reliability and replicability of our study results.

7.5 Operational Setup

All experiments were conducted using an NVIDIA A100 GPU, allowing us to complete each experimental run within approximately 5 hours. This timeframe represents the experimental setup's computational efficiency and scalability, enabling swift iteration and analysis of results.

Chapter 8

Experimentation

In our experiments, we systematically explore various configurations to enhance abstractive summarization performance. We begin with a baseline experiment to establish a foundation, followed by fine-tuning on guidance signals to incorporate additional context. Building upon these findings, we then investigate a modified LED architecture with a shared embedding layer, aiming to improve alignment between text and guidance signals. Subsequently, we extend this approach by sharing both the embedding and encoder layers to refine alignment further and reduce model complexity. Each experiment builds upon the previous one, incrementally refining our understanding of effective strategies for abstractive summarization. We aim to elucidate key insights into the mechanisms underlying successful summarization techniques through systematic iteration and analysis.

8.1 Baseline - LED

The LED Large pretrained model is the fundamental basis for our initial experiments and is designed primarily to summarize lengthy scientific papers. Utilizing a substantial collection of arXiv articles for training, the LED Large model has exceptional performance on the Scientific articles dataset, making it an appropriate choice for assessing the quality of summarization on the SciArg dataset.

The reason for utilizing the LED Large pretrained model as the baseline is two-fold. Firstly, we aim to assess how efficiently the model's weights can be transferred to the SciArg dataset. Due to the resemblance between the SciArg dataset and the arXiv articles used for pretraining, we expect the model's performance to roughly match the official scoring reports on the Scientific Papers dataset (Cohan et al., 2018b). Furthermore, the baseline experiment serves as a standard against which the effectiveness of future fine-tuning and improvement experimentation may be evaluated. The presence of any noticeable enhancements compared to the baseline will showcase the effectiveness of the suggested approaches in improving the quality of summarization on the SciArg dataset.

Each scientific publication in the SciArg collection is considered an independent document for summarization in our experimental setting. The whole content of each article is used as the input text, which the LED Large pretrained model then processes to produce the matching abstract.

8.2 Fine-tuning on Argument Structure

The primary objective of this experiment is to assess the efficiency of altering the LED design for text summarization by integrating contextual information obtained from the argument structure. We use a straightforward strategy to fine-tune the pretrained LED model, employing each of the three parsing methods mentioned previously as the encoder's input. We then train the model to forecast the summary for each document in the dataset. This experiment is a starting point for evaluating the possible advantages of altering the LED structure and provides information for further alterations outlined in the next section.

This is intended to assess the effect of integrating argument structure on summarization performance. By including contextual information obtained

from ADUs, we aim to improve the model's capacity to comprehend the fundamental conversation, as the model acquires the ability to incorporate the contextual information given by the ADUs and provide short and coherent summaries using this enhanced input.

This will result in the generation of summaries that are more useful and pertinent. This initial experiment enables us to evaluate the potential value of further exploring the modification of the LED design to utilize argumentative structure better.

8.3 Shared Embedding

In this experiment, we check out how well the modified LED architecture with shared embedding performs. This modified version shares only the embedding layer, similar to the case of the traditional LED model. Our goal here is to see if this change affects how the model works, particularly how it understands and summarizes text.

We're a bit sceptical about how effective this setup will be because there might not be enough coordination between the text and guidance parts of the model. The LED architecture relies on a good balance between these parts to produce accurate summaries. So, we're running this experiment to see if we need to tweak the model further to make it work better.

Even though we're unsure how well this modified LED will work, this experiment is important for us. We're carefully looking at how the modified architecture performs in controlled conditions. Our goal is to understand its strengths and weaknesses better. By doing this, we hope to figure out if there are any improvements we can make to enhance how the model processes and summarizes text. This experiment is a key step in our research journey

to improve abstractive summarization methods.

8.4 Shared Encoder

Our final experiment focuses on the version of the LED model that diverges from the standard design by sharing both the embedding and encoder layers. Among the various experiments we've conducted, this particular modification offers the most promise for improved performance.

This belief is primarily based on two main reasons. Firstly, aligning the encoder layers between the text and guidance signals offers a compelling proposition. This alignment, similar to the approach in the GSum framework, is crucial for ensuring the model captures and synthesizes the key features of the input text, enhancing the quality and coherence of the generated summaries.

Secondly, reducing the model size by consolidating shared layers presents an interesting opportunity to enhance stability and efficiency. This downsizing effort aims to streamline computational resources and minimise redundancy, increasing the model's ability to handle complex linguistic structures and extensive textual inputs.

As we delve into this final experimental phase, our objectives are clear. We aim to dissect the mechanics of the modified LED architecture to evaluate its performance and uncover insights that will guide future refinements. We seek to unravel the intricate relationship between architectural choices and summarization effectiveness through experimentation and thorough analysis. Our ultimate goal is to pave the way for more advanced and impactful controlled abstractive summarization techniques.

Chapter 9

Results and Discussion

In this section, we present the findings of our empirical investigations and discuss the implications and significance of these results comprehensively. Through experimentation and analysis, we delve into the performance of our proposed methodologies, examining their efficacy in enhancing the quality and coherence of abstractive summarization. By presenting outcomes in detail and contextualizing them within the broader landscape of abstractive summarization research, we aim to explain the key insights and address research questions —will integrating argumentative structure improve the generated abstractive summaries.

9.1 Baseline Evaluation

Our primary objective in the initial experimentation phase is to evaluate the effectiveness of the state-of-the-art baseline model on the SciArg dataset. This experiment's focal point is to ascertain whether the pretrained model's weights can be successfully applied to the new dataset. Given the similarities between the Scientific Papers dataset (Cohan et al., 2018b) and the SciArg dataset (Lauscher, Glavaš, and Ponzetto, 2018), this evaluation tests the model's adaptability to a slightly different domain.

While the datasets exhibit similarities, any significant deviation in results would indicate that the model's transferability is ineffective. However, our analysis, shown in Table 9.1, reveals that the baseline model's performance on the SciArg dataset somewhat aligns with the results reported in the original work.

Dataset	**R1**	**R2**	**RL**
Scientific Papers	46.63	19.62	41.83
SciArg	44.09	15.26	23.77

TABLE 9.1: Comparison of Baseline Results

One possible reason for the misalignment of ROUGE-L scores could be the nuanced differences in the stylistic conventions and linguistic structures between the Scientific Papers dataset and the SciArg papers. While both datasets may cover similar subject matter, the variation in writing styles, terminology usage, and sentence structures could lead to discrepancies in how well the generated summaries align with the reference summaries.

Additionally, subtle differences in the annotation or formatting of the reference summaries within the SciArg dataset compared to the original Scientific Papers dataset may also contribute to the divergence in ROUGE-L scores. These factors highlight the importance of considering dataset-specific characteristics and potential noise caused due to domain shifts when evaluating summarization models, emphasizing the need for robustness and adaptability in model architectures and training methodologies.

Retaining the pretrained model as the foundation for further experimentation remains the most pragmatic choice, especially considering the alternative of pretraining a new model on the SciArg dataset. However, the SciArg dataset only consists of 40 samples, which is too limited for proper pretraining. Training a model on such a small dataset poses significant challenges, as the model may struggle to generalize well to new data and exhibit overfitting tendencies.

Moreover, the computational resources and time required for pretraining a model would be substantial, and there is no guarantee of achieving superior performance compared to the pretrained model. Starting from scratch could also introduce additional variability and uncertainty in the results, making it challenging to establish a reliable baseline for comparison. Therefore, despite the observed differences in ROUGE-L scores, using the pretrained model offers the most efficient approach by leveraging existing knowledge and resources to expedite the experimentation process while maintaining consistency with established practices in the field.

9.2 Fine-tuning

In the subsequent experiment, we explore the model's behaviour when given guidance signals extracted from the Argument Structure. Our primary focus here is solely on manipulating input data while keeping the model architecture unchanged. So, instead of passing the full text into the encoder, we input the parsed guidance signal derived from the argument structure.

While this approach deviates slightly from the primary objective of this thesis, which centres on integrating argument structure as guidance rather than sole input, it nonetheless presents an interesting idea for examination. By analyzing the model's performance under these altered conditions, we hope to gain invaluable insights into the potential efficacy of leveraging argument structure as a guiding mechanism.

Guidance Type	**R1**	**R2**	**RL**
Span Texts	43.71	15.69	23.51
Annotated Span Texts	45.18	15.98	24.59
Annotated Source Texts	45.07	15.68	24.52

TABLE 9.2: Fine-tuning using Guidance Signal

Our observations, shown in Table 9.2, yield an enhancement over the baseline model, particularly when employing either ANNOTATED SPAN TEXTS or ANNOTATED SOURCE TEXTS as guidance inputs. Both guidance types show an increase of approximately 1 point in ROUGE-L over the baseline approach, which uses the full text. On the other hand, the utilization of TEXT SPANS exhibits a decrease in performance relative to the baseline.

Due to this, we believe that it becomes evident that annotated guidance signals, such as ANNOTATED SPAN TEXTS and ANNOTATED SOURCE TEXTS, provide a richer context for the model to align more closely with the abstracts. Conversely, TEXT SPANS's inferior performance may stem from their inherent ambiguity. TEXT SPANS lack the explicit annotations necessary for the model to discern and effectively utilize the provided guidance since the model has no way of distinguishing between ADUs.

Another observation is the slight performance advantage of ANNOTATED SPAN TEXTS over ANNOTATED SOURCE TEXTS. This indicates the effectiveness of providing ADUs as guidance signals, reconfirming our hypothesis that ADUs offer sufficient context for the model to generate abstracts effectively. This clarifies the impact of different types of guidance signals and hints at potential avenues for refinement and optimization in future iterations of the model architecture.

9.3 Shared Embedding

Next, we explore the efficacy of different guidance types within the modified LED architecture featuring a shared embedding layer. This architecture, which more closely resembles the traditional LED model, is expected to offer stability and coherent abstract text generation.

We observe interesting outcomes after examining the results, shown in Table

9.3. While the modified architecture demonstrates the ability to produce coherent abstracts, the findings reveal a nuanced performance pattern. Using annotated SPAN TEXTS and ANNOTATED SOURCE TEXT, we observe slight improvements in ROUGE-1 scores compared to the baseline. However, other metrics exhibit inferior performance relative to the baseline.

Guidance Type	**R1**	**R2**	**RL**
Span Texts	43.15	15.31	23.29
Annotated Span Texts	44.97	15.38	24.45
Annotated Source Texts	44.26	15.36	23.55

TABLE 9.3: Fine-tuning Modified LED with Shared Embedding

Interestingly, when contrasted with fine-tuning on the standard LED architecture, we discern a consistent decrease of approximately 0.5 points across all metrics for all three guidance types.

This decrease in performance can be attributed to the limited alignment between guidance and source text within the modified architecture according to us. This lack of alignment complicates the model's processing, resulting in a decline in performance compared to the previous experiment, in which guidance signals are directly utilized as input. Consequently, the increased complexity introduced by the modified architecture impedes overall performance, highlighting the importance of effective alignment mechanisms for optimal model functioning.

9.4 Shared Encoder

To ensure improved alignment between guidance signals and source text, we revisit the results obtained from the previous experiment, which involved sharing only the embedding layer. To enhance alignment further, we opt to share the embedding layer and the weights of the encoder itself, drawing inspiration from the GSum architecture.

As expected, these adjustments to enhance alignment yield better results, shown in Table 9.4. Notably, all three types of guidance signals outperform the baseline approach, with Annotated Span Texts and Annotated Source Texts exhibiting a remarkable 2 to 3-point increase in both ROUGE-1 and ROUGE-L scores. We also observe notable improvements over the baseline and an approximate 1-point increase across all metrics when compared to fine-tuning the standard LED model directly on the guidance signals.

Guidance Type	**R1**	**R2**	**RL**
Span Texts	44.22	15.11	24.44
Annotated Span Texts	46.32	16.99	25.81
Annotated Source Texts	46.24	15.60	25.27

TABLE 9.4: Fine-tuning Modified LED with Shared Encoder

These findings showcase the efficacy of our approach in augmenting alignment between guidance signals and source text, thereby facilitating enhanced summarization performance. Moreover, the observed improvements highlight the potential of architectural modifications to yield tangible enhancements in model efficacy and underscore the importance of thoughtful design considerations in model development.

9.5 Alignment of the Guidance Signal with Full Text

One reason the simple addition of the two outputs from each encoder works so efficiently may be that each encoder successfully encodes the data into a form that no longer depends on the input itself. We believe that this representation is generic enough that the difference in input types, full text vs guidance signal, no longer matters so that they can be added together.

The modified architecture allows for the seamless integration of guidance information with the source text by adopting this additive pooling approach. This sum of contextual representation enables the model to leverage the extra information provided by the guidance signal, thereby improving the summarization process.

Furthermore, the additive pooling mechanism embodies a flexible approach to alignment, accommodating variations in guidance signal characteristics and source text complexities. This adaptability is a core feature of the modified architecture, enabling it to effectively handle diverse textual inputs and various types of guidance signals to form a better and more coherent abstract generation.

Chapter 10

Limitations and Future Work

Our study underscores promising results, yet it is essential to recognize and address certain limitations that affect future research. These limitations span different facets of the proposed architecture, each presenting distinct challenges and opportunities for improvement.

The first limitation concerns the architecture's reliance on the quality of extracted Argumentative Discourse Units (ADUs) from the source text. Addressing this limitation entails refining ADU extraction methods to ensure consistent and reliable guidance signal generation.
The architecture's effectiveness is contingent upon the accuracy and meaningfulness of these ADUs. Suboptimal ADU extraction may compromise the alignment between the guidance signal and the source text, hindering the architecture's performance. In this work, we bypass this issue by using human-annotated ADUs, but for the work done in this thesis to be more practical, automatic ADU extractions are imperative.

The second limitation concerns the guidance signal's encoding using a language encoder, which constrains the signal's format and representation. While our study demonstrates the feasibility of representing ADUs in plain text format, accommodating more complex guidance signals, such as Argument

Relations, has challenges. Encoding intricate concepts like Argument Relations in plain text format risks information loss or distortion during encoding. Overcoming this limitation requires exploring alternative encoding strategies like knowledge graphs (Ji et al., 2022) or specialized architectures like tree transformers (Wang, Lee, and Chen, 2019) capable of handling complex guidance signals effectively.

To mitigate these limitations and advance the state-of-the-art in abstractive summarization, future research could focus on refining ADU extraction methods, exploring alternative encoding strategies for complex guidance signals, and investigating hybrid approaches that combine linguistic insights with machine learning techniques. By addressing these challenges and exploring new research directions, we can develop more robust and adaptable summarization models capable of generating high-quality summaries across diverse domains and applications.

Chapter 11

Conclusion

In concluding this study, our primary target has been to enhance abstractive summarization methodologies through the strategic incorporation of additional contextual information, particularly in the form of argument structure. The selection of argumentative discourse units (ADUs) as our focal point stems from their inherent adaptability within plain text formats, affording us a robust foundation for integrating supplementary contextual cues. A comprehensive review of relevant literature has not only acquainted us with established abstractive summarization techniques tailored for lengthy documents but has also provided valuable insights into strategies for effectively integrating additional contextual information.

The adoption of the Longformer-Encoder-Decoder (LED) architecture as our core framework has been instrumental in realizing our research objectives. However, a significant departure from conventional LED architectures was necessary to accommodate the seamless integration of extra contextual cues through guidance signals. Consequently, the LED architecture underwent careful modification to ensure compatibility with our proposed approach. Moreover, insights gleaned from the GSum architecture have further enriched our methodology, enabling us to harness its potential for enhancing the coherence and logical flow of generated summaries.

An essential contribution of our research lies in the exploration of ADUs and

their diverse parsing methods to serve as effective guidance signals. By leveraging ADUs in various forms, we have strived to provide the summarization model with nuanced contextual cues essential for producing coherent and contextually relevant summaries. This exhaustive exploration has not only broadened our understanding of the potential applications of ADUs but has also shed light on their effectiveness in encapsulating pivotal information from the source text.

Empirical findings from our experimental results have yielded a wealth of intriguing insights into the efficacy of our proposed methodology. Foremost among these revelations is the remarkable performance of the fine-tuned LED model on ADUs, surpassing benchmarks set by the state-of-the-art LED baseline. This compelling result underscores the inherent capability of ADUs in encapsulating critical information from the source text, thereby facilitating the generation of more comprehensive and contextually relevant summaries.

Our empirical investigations have underscored the critical importance of alignment between the guidance signal and the source text. Instances of weak alignments, such as sharing only the embedding layer, have resulted in architectural instability and consequent performance degradation.

Modifying the LED architecture for better alignment, exemplified by the shared encoder architecture, has yielded superior performance outcomes. It outperforms both the baseline LED by 3 to 4 points on the ROUGE metric and direct fine-tuning on ADUs by 1 to 2 points on the ROUGE metric. This confirms the advantage of guidance signals for adding in crucial context that may have otherwise been lost in models reliant solely on the full text or direct fine-tuning on ADUs.

To conclude, this work shows the potential inherent in integrating argument structure information to enhance the efficiency and quality of abstractive

summarization techniques. By integrating ADUs as guidance signals and refining the alignment between guidance signals and source text, this research contributes to advancing the state-of-the-art in abstractive summarization, enabling more accurate, informative, and contextually nuanced summarization outcomes.

Bibliography

Beltagy, Iz, Matthew E Peters, and Arman Cohan (2020). "Longformer: The long-document transformer". In: *arXiv preprint arXiv:2004.05150*.

Cao, Ziqiang et al. (July 2018). "Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 152–161. DOI: 10.18653/v1/P18-1015. URL: https://aclanthology.org/P18-1015.

Cohan, Arman et al. (June 2018a). "A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 615–621. DOI: 10.18653/v1/N18-2097. URL: https://aclanthology.org/N18-2097.

– (2018b). "A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. DOI: 10.18653/v1/n18-2097. URL: http://dx.doi.org/10.18653/v1/n18-2097.

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed.

by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Dou, Zi-Yi et al. (June 2021). "GSum: A General Framework for Guided Neural Abstractive Summarization". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, pp. 4830–4842. DOI: 10.18653/v1/2021.naacl-main.384. URL: https://aclanthology.org/2021.naacl-main.384.

Etemad, Abdul Ghafoor, Ali Imam Abidi, and Megha Chhabra (2021). "A Review on Abstractive Text Summarization Using Deep Learning". In: *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pp. 1–6. DOI: 10.1109/ICRITO51393.2021.9596500.

Fabbri, Alexander et al. (Aug. 2021). "ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, pp. 6866–6880. DOI: 10.18653/v1/2021.acl-long.535. URL: https://aclanthology.org/2021.acl-long.535.

Fisas, Beatriz, Francesco Ronzano, and Horacio Saggion (May 2016). "A Multi-Layered Annotated Corpus of Scientific Papers". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari et al. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 3081–3088. URL: https://aclanthology.org/L16-1492.

Fisas, Beatriz, Horacio Saggion, and Francesco Ronzano (June 2015). "On the Discoursive Structure of Computer Graphics Research Papers". In: *Proceedings of the 9th Linguistic Annotation Workshop*. Ed. by Adam Meyers, Ines Rehbein, and Heike Zinsmeister. Denver, Colorado, USA: Association for

Computational Linguistics, pp. 42–51. DOI: 10.3115/v1/W15-1605. URL: https://aclanthology.org/W15-1605.

Giarelis, Nikolaos, Charalampos Mastrokostas, and Nikos Karacapilidis (2023). "Abstractive vs. Extractive Summarization: An Experimental Review". In: *Applied Sciences* 13.13. ISSN: 2076-3417. DOI: 10.3390/app13137620. URL: https://www.mdpi.com/2076-3417/13/13/7620.

Ji, Shaoxiong et al. (2022). "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.2, pp. 494–514. DOI: 10.1109/TNNLS.2021.3070843.

Jin, Hanqi, Tianming Wang, and Xiaojun Wan (2020). "SemSUM: Semantic Dependency Guided Neural Abstractive Summarization". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05, pp. 8026–8033. DOI: 10.1609/aaai.v34i05.6312. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6312.

Kikuchi, Yuta et al. (Nov. 2016). "Controlling Output Length in Neural Encoder-Decoders". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by Jian Su, Kevin Duh, and Xavier Carreras. Austin, Texas: Association for Computational Linguistics, pp. 1328–1338. DOI: 10.18653/v1/D16-1140. URL: https://aclanthology.org/D16-1140.

Lauscher, Anne, Goran Glavaš, and Simone Paolo Ponzetto (2018). "An argument-annotated corpus of scientific publications". In: Association for Computational Linguistics.

Lewis, Mike et al. (July 2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703.

Li, Chenliang et al. (June 2018). "Guiding Generation for Abstractive Text Summarization Based on Key Information Guide Network". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans,

Louisiana: Association for Computational Linguistics, pp. 55–60. DOI: 10.18653/v1/N18-2009. URL: https://aclanthology.org/N18-2009.

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: https://aclanthology.org/W04-1013.

Loshchilov, Ilya and Frank Hutter (2017). "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101*.

Mao, Anqi, Mehryar Mohri, and Yutao Zhong (2023). "Cross-entropy loss functions: theoretical analysis and applications". In: *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. Honolulu, Hawaii, USA: JMLR.org.

Roush, Allen and Arvind Balaji (Dec. 2020). "DebateSum: A large-scale argument mining and summarization dataset". In: *Proceedings of the 7th Workshop on Argument Mining*. Ed. by Elena Cabrio and Serena Villata. Online: Association for Computational Linguistics, pp. 1–7. URL: https://aclanthology.org/2020.argmining-1.1.

Sukhbaatar, Sainbayar et al. (July 2019). "Adaptive Attention Span in Transformers". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 331–335. DOI: 10.18653/v1/P19-1032. URL: https://aclanthology.org/P19-1032.

Toulmin, Stephen E. (2003). "Contents". In: *The Uses of Argument*. Cambridge University Press, v–vi.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Walton, Douglas, Christopher Reed, and Fabrizio Macagno (2008). *Argumentation schemes*. Cambridge University Press. Chap. 11.

Wang, Yau-Shian, Hung-yi Lee, and Yun-Nung Chen (2019). "Tree Transformer: Integrating Tree Structures into Self-Attention". In: *CoRR* abs/1909.06639. arXiv: 1909.06639. URL: http://arxiv.org/abs/1909.06639.

Zhang, Mengli et al. (2022). "A comprehensive survey of abstractive text summarization based on deep learning". In: *Computational intelligence and neuroscience* 2022.

Zhao, Yao, Mohammad Saleh, and Peter J. Liu (2020). *SEAL: Segment-wise Extractive-Abstractive Long-form Text Summarization*. arXiv: 2006.10213 [cs.CL].

Zhu, Chenguang et al. (June 2021). "Enhancing Factual Consistency of Abstractive Summarization". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, pp. 718–733. DOI: 10.18653/v1/2021.naacl-main.58. URL: https://aclanthology.org/2021.naacl-main.58.