Summarizing User-generated Discourse

DISSERTATION

Accepted by the
Faculty of Mathematics and Computer Science at Leipzig University
for the attainment of the academic degree of

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

in the field of
Computer Science

by Mr. Shahbaz Syed (M.Sc.)
born on 24. June 1990 in Hyderabad, India.

Acceptance of the dissertation was recommended by:

1. Prof. Dr. Martin Potthast, Leipzig University
2. Prof. Dr. Elena Cabrio, Université Côte d'Azur, France

The academic degree is awarded upon successful defense on 30.05.2024 with the overall grade of magna cum laude.

Winning gives birth to hostility. Losing, one lies down in pain. The calmed lie down with ease, having set winning and losing aside.

– Siddhartha Gautama, The Buddha
Dhammapada, verse 201

Abstract

Automatic text summarization is a long-standing task with its origins in summarizing scholarly documents by generating their abstracts. While older approaches mainly focused on generating extractive summaries, recent approaches using neural architectures have helped the task advance towards generating more abstractive, human-like summaries.

Yet, the majority of the research in automatic text summarization has focused on summarizing professionally-written news articles due to easier availability of large-scale datasets with ground truth summaries in this domain. Moreover, the inverted pyramid writing style enforced in news articles places crucial information in the top sentences, essentially summarizing it. This allows for a more reliable identification of ground truth for constructing datasets. In contrast, user-generated discourse, such as social media forums or debate portals, has acquired comparably little attention, despite its evident importance. Possible reasons include the challenges posed by the informal nature of user-generated discourse, which often lacks a rigid structure, such as news articles, and the difficulty of obtaining high-quality ground truth summaries for this text register.

This thesis aims to address this existing gap by delivering the following novel contributions in the form of datasets, methodologies, and evaluation strategies for automatically summarizing user-generated discourse: (1) three new datasets for the registers of social media posts and argumentative texts containing author-provided ground truth summaries as well as crowdsourced summaries for argumentative texts by adapting theoretical definitions of high-quality summaries; (2) methodologies for creating informative as well as indicative summaries for long discussions of controversial topics; (3) user-centric evaluation processes that emphasize the purpose and provenance of the summary for qualitative assessment of the summarization models; and (4) tools for facilitating the development and evaluation of summarization models that leverage visual analytics and interactive interfaces to enable a fine-grained inspection of the automatically generated summaries in relation to their source documents.

Contents

1 Introduction

This thesis focuses on automatic text summarization and its evaluation from a human-centered perspective within the domain of user-generated discourse. Specifically, it focuses on textual sources like social media posts, argumentative texts, and forum discussions in particular. This chapter gives a concise background on the domain of user-generated discourse, describes the role of automatic text summarization, and highlights the research gaps that my thesis aims to address. Next, it outlines the research questions tackled and the corresponding contributions grouped into three aspects of summarization research namely data, methodologies, and evaluation. The chapter then concludes with the publication record that is the basis of this thesis as well as publications that represent my broader research interests.

1.1 Understanding User-Generated Discourse

A significant portion of the information we consume on a daily basis comes from various user-generated sources like social media, for example, Reddit posts or Tweets, opinionated texts such as news editorials, and forum discussions facilitated by social media platforms and debate portals. It is estimated that by 2023, 3.5 trillion pieces of content including media formats such as audio, video, images will be created and shared on the web, each month [196]. Collectively known as *user-generated content*, these contributions from active internet users with diverse backgrounds and demographics aim to benefit both individuals and society as a whole, by enabling the sharing of knowledge on various topics of interest.

User-generated *discourse*, a subclass of user-generated content, covers a broad spectrum of topics, including politics, societal changes, science, entertainment, technology, and product (consumer) reviews providing us with unfiltered opinions and diverse viewpoints. As a result, it becomes an essential source of information for understanding public opinions on different matters. For instance, 74% of consumers rely on content from other customers to make purchasing decisions [298]. This makes user-generated discourse a crucial part of driving business decisions, such as product development and marketing strategies, as it provides valuable insights into the needs and preferences of the target audience. Likewise, discussions on Reddit about controversial topics and relevant events invoke active participation from the community. Participants in these discussions usually introduce diverse perspectives of these issues, for instance, on ChangeMyView [1]. Nonetheless, the sheer volume of user-generated discourse on the web poses a challenge in keeping up with the latest developments in our areas of interest. The ever-evolving nature of this content, for instance, ongoing controversial discussions with arguments and counter-arguments, in contrast to static sources like news reports or scientific papers, makes it hard to stay updated. Automatic summarization has a strong potential to tackle this challenge and facilitate effective information processing.

Characteristics of User-Generated Discourse

Before we delve into the process of developing automatic summarization methods, it is crucial to consider the underlying characteristics of user-generated discourse that must be taken into account when designing such methods. In this section, I highlight three key aspects of user-generated discourse that are relevant to the summarization task:

1. **Structure**: User-generated discourse differs from other longform texts on the web, for instance, news articles, in terms of content ordering. Unlike news articles that typically present the gist at the beginning in the lead paragraph, user-generated discourse often follows a more narrative style involving multiple arguments, scattering important details throughout the document. This heterogeneity makes the task of summarizing this register challenging by comparison, as no heuristic can be formulated for effectively identifying the most vital information within the source document.

[1] `https://www.old.reddit.com/r/changemyview/`

2. **Style**: Informality is a common characteristic of user-generated content. The usage of abbreviations, colloquial phrases, and platform-specific slang (as well as offensive language), which may evolve over time, can be unfamiliar to many readers. As a result, crafting coherent and self-contained summaries demands extra effort to ensure readability and accessibility.

3. **Summary Quality**: Evaluating the quality of summaries can be a complex task. For instance, the lead paragraph in a news article often serves as a reliable summary for the rest of the article, due to the inverted pyramid writing style that places crucial information in the initial sentences [237]. However, news editorials deviate from this norm. In these pieces, the lead content is typically crafted to engage the reader's interest, rather than to offer a succinct summary of the editorial. As such, it is essential to establish clear guidelines that take into account the desired quality dimensions for the collection and assessment of high-quality summaries of user-generated discourse.

1.2 The Role of Automatic Summarization

Automatic summarization employs computational methods to condense long pieces of information to their core content, extracting or abstracting the most important (and relevant) details. A high-quality summary saves its readers time in understanding the main takeaways and may even replace the original document to some extent. Therefore, summaries serve a crucial purpose in facilitating faster communication and sharing of ideas, knowledge, and opinions with others. For long discourses, this applies too, especially if only parts of a discussion are relevant or interesting to the reader.

The process of developing automatic summarization models tailored to a specific domain, such as user-generated discourse, encompasses three key stages as depicted in Figure 1.1. The initial stage involves the collection of a large-scale dataset of high-quality ground truth, which serves as a basis for training a summarization model. For unsupervised methodologies, this stage is simplified to compiling a small sample of summaries in advance (test set), or alternatively, conducting a manual evaluation of the generated summaries. The second stage requires the design of an apt model architecture capable of learning to generate summaries from the compiled dataset, and if necessary, incorporating additional features. The final stage involves a comprehensive evaluation of the model to ascertain its ability to produce high-quality summaries. These stages are detailed below.

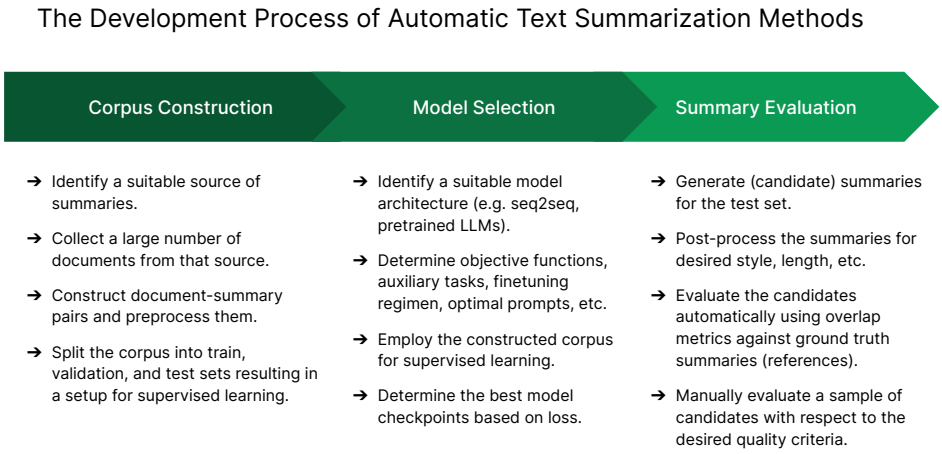


Figure 1.1: Overview of the process of developing automatic text summarization models via supervised learning. The first step is to construct suitable corpora for training the summarization model. The second step involves training the model using the collected corpora. Finally, the trained model is evaluated automatically and manually to ensure that it generates high-quality summaries. This thesis contributes valuable resources at each step of this process.

Corpus Construction The initial and crucial step in devising supervised methods for automatic summarization is the collection of a large-scale dataset comprising high-quality ground truth. This dataset serves as a benchmark for training the summarization model. However, obtaining such high-quality summaries on a substantial scale poses a significant challenge. Past attempts to compile summarization datasets have predominantly relied on scraping news websites to automatically extract accompanying leads as summaries for news articles [132, 206]. Regrettably, this approach often results in noisy and/or incomplete summaries, which may not fully encapsulate the gist of the news article [167]. Furthermore, the summaries are often extractive in nature, which is not ideal for abstractive summarization models. The models trained on such datasets are biased towards extracting sentences rather than abstracting over the text, which may not be suitable for summarizing user-generated discourse.

A more rational and effective alternative is to seek for *abstractive* summaries provided by the authors of the source content. In contrast to simply listing the important sentences from the source text, these summaries compress information from different parts of the text into a coherent and fluent summary. In order to obtain such abstractive summaries from less struc-

tured content than news articles, like social media posts, it is crucial that one exploits some consistent author-provided signal indicating that they are writing a summary, or else one can only resort to manually labeling summaries for each individual document, which quickly becomes infeasible as the number of documents grows.

In the case of unsupervised methods for summarization, there is no strict requirement for a large-scale labeled collection of document-summary pairs. The summaries can be written for a small sample of documents beforehand to serve as a test set for automatically evaluating the unsupervised methods. Alternatively, the generated summaries can be directly evaluated through a manual assessment of their suitability for a given task (i.e., the purpose of the summary).

Model Selection The next stage involves selecting the appropriate model architectures that can process a document and generate a high-quality summary. For supervised models, we utilize the corpus developed in the previous stage and train the model to create summaries. The document and ground truth summary pairs from the corpus constitute labeled training examples that supervise the learning process. Most supervised abstractive summarization methodologies are founded on the sequence-to-sequence architecture [312]. This architecture comprises an encoder that processes the source document and generates a representation of it, and a decoder that uses this representation to create the summary. The generation is based on conditional language modeling where the decoder generates the next word in the summary conditioned on the previously generated words. The encoder-decoder architecture is trained to minimize the cross-entropy loss between the generated summary and the ground truth summary.

Summary Evaluation Finally, the trained model must be evaluated to ensure that it generates high-quality summaries. A common practice of evaluating the generated *candidate* summaries is to automatically compare them with the ground truth *reference* summaries. This comparison is primarily operationalized by researchers by measuring the lexical overlap between the candidate and reference summaries using metrics such as ROUGE [176] and BLEU [225]. These metrics measure the precision and recall of only lexical units such as words or n-grams between the candidate and reference summaries. Also, they do not consider the semantic similarity between the two summaries. Evidently, this approach has its limitations when it comes to *abstractive* summaries, which may include semantically similar words not found in the document. Consequently, true qualitative assessment of sum-

maries requires manual inspection by multiple experts considering clearly defined quality dimensions [71]. Nevertheless, manual evaluation poses its own set of challenges, particularly the absence of instructions to the evaluators that the *purpose* of the summary must also be considered during assessment. For instance, how well do the generated snippets of web pages help to quickly identify relevant documents in a search engine results page [59]. The purpose of a summary, as highlighted by Jones et al. [149], significantly influences automatic summarization and, by extension, guides the qualitative assessment of automatically generated summaries. Chapter 2 provides a detailed discussion on the challenges associated with evaluating summaries as identified in the literature.

The Purpose of Summarizing User-generated Discourse

User-generated discourse represents an ever-evolving text register, driven by the active participation of a diverse range of contributors. These individuals share their unique insights and expertise on a multitude of engaging topics. Furthermore, each new generation of users, spanning various age groups, introduces their own slang and vernacular, which continually morphs in response to the latest trends in social communication. The majority of the audience engages with this content with an information-seeking perspective, aiming to make informed decisions [121]. For instance, users often search for insights on the pros and cons of new technologies or the advantages and disadvantages of specific policies [38, 317]. However, this information-seeking process can be time-consuming, especially when users are looking for specific perspectives or for an overview of all perspectives, amidst a large number of posts.

Thus, the purpose of automatically generated summaries of these discourses is to alleviate information overload, in particular, facilitating quick comprehension of diverse perspectives on a given topic. Ideally, such summarization supports users in their decision-making process, allowing them to effectively comprehend and efficiently navigate through the content to gain valuable insights. This is precisely the goal of this thesis, which contributes unique data, effective methods, and contextualized evaluation for automatic summarization of user-generated discourse.

Research Gaps

While automatic summarization has been a long-standing task (Chapter 2), it has not been extensively studied for various types of user-generated web

discourse. Designing data-driven methods for automatic summarization in this context poses multiple challenges, including:

1. **Lack of suitable datasets**. The majority of datasets utilized for training summarization models are derived from news articles. The extractive highlights accompanying these articles, their headlines, or simply the top few sentences (lead paragraph) are deemed as the ground truth summaries. This is the case for widespread summarization datasets such as CNN/DailyMail [206], Gigaword [114], and XSum [209]. Since, news articles typically follow an inverted pyramid writing style, placing the most crucial information in the initial sentences. Consequently, models trained solely on news summarization datasets display a significant bias towards selecting the opening sentences as the document's summary [158]. This bias hampers their ability to generalize to other domains, such as user-generated discourse. In these domains, the narrative writing style often disperses important information throughout the document.

2. **Insufficient evaluation**. The prevalent method for evaluating automatic summarization involves utilizing metrics like ROUGE [176] or BLEU [225], which quantify the lexical overlap between the reference and the candidate summaries. While these metrics are apt for evaluating extractive summaries, they fall short when it comes to abstractive summaries, as they fail to capture semantic overlap of abstracted information. Although there are several semantic similarity metrics available, such as MoverScore [344], BERTScore [341], and BARTScore [336], they also have limitations. Specifically, they do not evaluate summaries on qualitative aspects such as informativeness, coherence, and redundancy. Consequently, a human-centric evaluation of summaries becomes essential for a comprehensive assessment of model effectiveness.

3. **Purposeless task design**. Evaluating the quality of summaries often involves human assessment, guided by well-defined quality dimensions [71]. Yet, the design of the annotation task typically does not explicitly specify the *purpose* of the summaries under evaluation. The purpose of a summary is a vital context factor that affects its appropriateness for a specific downstream task [149]. While the purpose may be implicitly encoded in reference summaries, explicitly stating the purpose of the summary can aid in obtaining more reliable quality judgments from the annotators.

Addressing these challenges is essential for advancing automatic summarization techniques in the domain of user-generated web discourse and improving the overall quality and reliability of generated summaries. The research questions corresponding to each of these challenges as well as the contributions of this thesis are discussed in the following section.

1.3 Research Questions and Contributions

The primary goal of my thesis is to address the aforementioned research gaps in a systematic manner by contributing novel methods and resources towards improving the state-of-the-art in summarization of user-generated discourse. Table 1.1 summarizes these contributions. This section presents the main research questions I formulated and tackled in the thesis; some of the chapters address multiple research questions. The sections below outline the research questions and contributions according to the three aspects of summarization research: data, methods, and evaluation.

Data

1.3.1 How can high-quality summaries of argumentative texts be defined and operationalized?

Identifying what constitutes a good summary can be a complex task, primarily due to its subjective nature, which is influenced by the intended audience and the summary's purpose. In some documents, the structure can inherently suggest which content is worthy of inclusion in the summary. For instance, the initial sentences of a news report, also known as the lead, or a scholarly document's abstract, are designed to function as summaries. However, for user-generated discourse, such a structure, indicating summary-worthy content, often does not exist. The structure of such content can range from resembling an essay with a thesis statement to being completely unstructured, like a social media post.

In Chapter 3, we devise a definition of a high-quality summary exemplifying news editorials as the target domain. The choice of editorials is strategic; being news articles, they are usually structured like well-formed essays and represent the only news register not following the conventional inverted pyramid writing style. Unlike standard news articles that primarily aim to inform, editorials seek to persuade readers or call for a specific action, making the lead paragraph inadequate as a summary. Editorials are designed more to pique interest than to encapsulate the topic, necessitating a clear definition of what constitutes a high-quality summary in this con-

text. Moreover, since editorials are argumentative in nature, they also form a connection to user-generated discourse, providing a discourse model with a well-formed style and structure.

In response to this, we establish and implement an annotation process for compiling *high-quality* summaries of news editorials. The resulting Webis-EditorialSum-20 corpus comprises 1330 summaries derived from 266 news editorials, providing a robust dataset for the qualitative assessment of future research in this domain. We identified structural differences between high- and low-quality summaries and investigated the effectiveness of existing state-of-the-art models in summarizing opinionated text.

1.3.2 How can summary ground truth for user-generated discourse be collected at scale?

Unlike news articles, which often feature extractive summaries in the form of highlights or titles, user-generated discourse, excluding news editorials, typically lacks such structure. In Chapter 4, I delve into our innovative use of Reddit users' habit of including a `TL;DR` (Too Long, Didn't Read summary) with their posts. Social media communities identified the need for summarization when users often responded to long posts from fellow users with a "TL;DR" indicating that they did not have the time or interest to read the entire content and desired a short summary of their posts. Consequently, providing a "TL;DR" evolved into a more generally accepted way of summarizing information, even in contexts outside of online forums.

We exploited this signal, for the first time, to extract highly abstractive summaries, leading to the creation of the Webis-TLDR-17 corpus. This dataset, comprising around 2 million <`post, tl;dr`> pairs, is a pioneering resource for training abstractive summarization models specifically for social media posts.

Subsequently, in Chapter 5, I introduce a novel approach to identifying summary ground truth with a specific purpose for informal argumentative texts. We capitalized on another emergent structure that evolved in the specific community of ChageMyView: providing concise titles that resemble *conclusions* of their subsequent reasoning. This approach not only helped identify the discussion's target from the conclusion but also indicate the author's stance on the topic.

Methods

1.3.3 How can the standard sequence-to-sequence model be extended for controlled summarization?

The conventional sequence-to-sequence model consists of an encoder, which processes the source document and generates a representation of it, and a decoder, which uses this representation to produce the summary. However, this model lacks the ability to explicitly control the characteristics of the generated summary, such as the inclusion of specific words or the incorporation of additional knowledge to enhance the model. In Chapter 5, we introduce control codes as a viable method to integrate external argumentative knowledge into the summarization model. This approach allows us to generate informative conclusions of argumentative texts, specifically informing the model about the discussion topic, the target entity of the conclusion, and the author's stance on it.

1.3.4 How can summaries that capture the various perspectives of a long discussion be generated?

Online discussions are a rich source of information that can help us understand the various perspectives on a given topic, also known as *frames*. Given the hundreds of arguments, an effective summary must provide an overview these perspectives. However, current methods for summarizing these discussions often fall short as they generate a single summary, which does not adequately capture the variety of perspectives. To address this, in Chapter 6, I present a novel way of summarizing discussions using a predefined inventory of frames that serve as anchors to structure extensive discussions. This approach allows for the creation of multiple frame-specific summaries for a discussion, instead of condensing all critical information into one summary. This ranking-based approach is entirely unsupervised and leverages retrieval models for a joint ranking of the arguments for frame relevance, topic relevance, and informativeness.

1.3.5 How can summarization aid in navigating long discussions?

Extending the task of discussion summarization, I delve into the task of generating *indicative* summaries which facilitate effective navigation of the discussions for identifying interesting perspectives as well as to contribute own arguments. In Chapter 7, I present a novel unsupervised approach for crafting indicative summaries for long discussions, which essentially function as tables of contents. Our approach clusters argument sentences,

generates cluster labels using a large language model (LLM) as abstractive summaries, and classifies these labels into a generic frame inventory. Based on an extensively optimized cluster-then-prompt approach, we evaluate 19 state-of-the-art prompt-based LLMs for generative cluster labeling and frame classification. To evaluate the usefulness of our indicative summaries, we conduct a user study: It shows that our summaries serve as a convenient navigation tool to explore long discussions.

Evaluation

1.3.6 How can the qualitative evaluation of abstractive summarization be improved?

Multiple evaluation metrics have been proposed to integrate semantic overlap between the generated summary and the reference, overcoming the drawbacks of lexical overlap metrics such as ROUGE and BLEU. However, these metrics still fall short in reliably assessing the quality of abstractive summaries as they do not reveal the relation between the generated summary and the source document. Truly abstractive summaries are designed to use new words, paraphrase, and coherently combine information from the source document to produce more human-like summaries. In this process, they often create factual errors by wrongly editing existing facts, or adding new facts that do not exist in the source document. Therefore, evaluating the quality of an abstractive model using *only* automatic metrics is insufficient. To address this issue, we argue that a visual comparison of the summary in the context of the source document is the most effective way to evaluate such summaries.

In Chapter 8, I present Summary Explorer, a visual analytics tool that facilitates multiple analyses of summary quality. It identifies position bias of various models, factuality, hallucinations, and agreement among summaries for a given document. Moreover, it speeds up the evaluation process by visually locating summary provenance, in comparison to reading summaries without any context.

1.3.7 How can the reproducibility of summarization models and metrics be improved?

Besides qualitatively evaluating summaries by comparing *only* with the reference summary, one must also compare qualitatively with the state of the art, both in terms of how a model improves over others or where it does not. However, few researchers do so in practice. One of the reasons is the lack of

supporting tools and especially the lack of availability of (many) other state-of-the-art models in executable form, thus lacking reproducibility. This also applies to new evaluation metrics that are developed for quantitative evaluation of summarization. To address this issue, in Chapter 9, I describe Summary Workbench, a web-based tool that provides a unified access to the state-of-the-art summarization models and supports quantitative evaluation. It includes visual analysis of the lexical and semantic overlap between the summaries and source documents and the correlation between a pair of automatic evaluation metrics for a summarization model. The tool is also deployable locally and emphasizes reproducible artifacts for easy sharing with the research community.

1.4 Thesis Structure

The structure of this thesis is as follows. Chapter 2 offers a comprehensive background on summarization from both a human and machine point of view, emphasizing the cognitive role of summarization in children's learning process. Additionally, it presents an overview of the two major paradigms of automatic summarization, namely extractive and abstractive summarization, its evaluation, and the challenges associated with it.

Building upon the theory of summarization, Chapter 3 then investigates defining and operationalizing high-quality summaries, exemplified for the domain of news editorials. Next, Chapters 4 and 5 present novel datasets that contain author-provided abstractive summaries of user-generated content. These chapters also present supervised methods trained on these corpora for abstractive summarization and their evaluation. In particular, for summarizing social media posts, we discuss the approaches from the participants of the TL;DR Challenge, the first shared task on abstractive summarization, accompanied by human-centered error analyses. Likewise, for conclusion generation, we present a novel approach to incorporating external knowledge into the summarization model for generating informative conclusions of argumentative texts.

The thesis then extends to long discussions, presenting two unsupervised approaches for informative as well as indicative summarization of long discussions. These summaries are designed to assist readers in exploring the various perspectives presented in a discussion, addressing a significant limitation of the conventional concept of a single, informative summary that aims to replace the entire discussion. In Chapter 6, I present a ranking-based approach for summarizing discussions using a predefined inventory of frames that serve as anchors to organize the long discussions, providing

multiple summaries tailored to the perspectives of interest. Following this, Chapter 7 presents an unsupervised approach purely based on LLMs for creating indicative summaries for long discussions, essentially functioning as tables of contents.

In the final part, the thesis addresses the evaluation phase of summarization research. Chapters 8 and 9 introduce visual analytics tools designed to facilitate both qualitative and quantitative evaluation of summarization models as well as to ensure the reproducibility of the models and evaluation metrics. Finally, Chapter 10 concludes the thesis and outlines potential future research directions in this field, especially in the context of prompt-based large language models.

Thesis Contributions		
Data		
Domain	***Summary Type***	***Contributions***
News Editorials	Persuasive Summaries	Webis-EditorialSum-20 (Ch. 3)
Social Media Posts	TL;DR	Webis-TLDR-17 (Ch. 4)
Argumentative Texts	Informative Conclusions	Webis-ConcluGen-21 (Ch. 5)
Long Discussions	Indicative Summaries	Discussion Explorer (Ch. 7)
Methods		
Approach	***Learning Paradigm***	***Contributions***
Distant Learning	Supervised	The TL;DR Challenge (Ch. 4)
External Knowledge Encoding	Supervised	Informative (abstractive) conclusions for persuasive texts (Ch. 5)
Prompt Engineering	Unsupervised	Table-of-Contents for forum discussions (Ch. 7)
Evaluation		
Approach	***Dimension***	***Contributions***
Human-centered Error Analysis	Summary Sufficiency	Annotated error types in ultra-short summaries of social media posts (Ch. 4, 5)
Purpose-based Crowdsourcing	Summary Purpose	Guidelines for evaluation and annotation of summaries emphasizing the *purpose* factor (Ch. 3, 6, 7)
Visual Analytics	Summary Provenance	Summary Explorer (Ch. 8)
Reproducible Models & Metrics	Reproducibility	Summary Workbench (Ch. 9)

Table 1.1: A summary of the core contributions of this thesis categorized by data, methods, and evaluation.

1.5 Publication Record

The following section lists the publications on which this thesis is based. Following this, also included is a list of publications that are not part of this thesis but that study related research areas.

Table 1.2: Overview of the publications included in this thesis. For each publication, the chapters (Ch.) in which the published content is covered is given, as well as the publication venue, the publication type, and the original number of pages.

Ch.	Reference	Venue	Type	Pages
4	Völske et al. [313]	NLP Frontiers@EMNLP	Workshop	4
*Michael Völske and Martin Potthast and **Shahbaz Syed** and Benno Stein. TL;DR: Mining Reddit to Learn Automatic Summarization, 2017.*				
3	Syed et al. [285]	COLING	Conference	12
***Shahbaz Syed** and Roxanne El Baff and Johannes Kiesel and Khalid Al Khatib and Benno Stein and Martin Potthast. News Editorials: Towards Summarizing Long Argumentative Texts, 2020*				
5	Syed et al. [286]	ACL	Conference	11
***Shahbaz Syed** and Khalid Al Khatib and Milad Alshomary and Henning Wachsmuth and Martin Potthast. Generating Informative Conclusions for Argumentative Texts, 2021.*				
6	Syed et al. [290]	SIGDIAL	Conference	11
***Shahbaz Syed** and Timon Ziegenbein and Philipp Heinisch and Henning Wachsmuth and Martin Potthast. Frame-oriented Summarization of Argumentative Discussions, 2023.*				
7	Syed et al. [289]	EMNLP	Conference	11
***Shahbaz Syed** and Dominik Schwabe and Khalid Al Khatib and Martin Potthast. Indicative Summarization of Long Discussions, 2023.*				
8	Syed et al. [287]	EMNLP	Conference	10
***Shahbaz Syed** and Tariq Yousef and Khalid Al Khatib and Stefan Jänicke and Martin Potthast. Summary Explorer: Visualizing the State of the Art in Text Summarization, 2021.*				
9	Syed et al. [288]	EMNLP	Conference	10
***Shahbaz Syed** and Dominik Schwabe and Martin Potthast. Summary Workbench: Unifying Application and Evaluation of Text Summarization Models, 2022.*				

Table 1.3: Overview of peer-reviewed publications not included in the thesis but that represent my broader research interests.

Reference	Venue	Type	Pages
Bondarenko et al. [39]	CLEF	Chapter	29
*Alexander Bondarenko and Maik Fröbe and Johannes Kiesel and **Shahbaz Syed** and Timon Gurcke and Meriem Beloucif and Alexander Panchenko and Chris Biemann and Benno Stein and Henning Wachsmuth and Martin Potthast and Matthias Hagen.* *Overview of Touché 2022: Argument Retrieval, 2022.*			
Alshomary et al. [11]	ArgMining@EMNLP	Conference	5
*Milad Alshomary and Timon Gurke and **Shahbaz Syed** and Philipp Heinisch and Maximilian Spliethöver and Philipp Cimiano and Martin Potthast and Henning Wachsmuth.* *Key Point Analysis via Contrastive Learning and Extractive Argument Summarization, 2021.*			
Alshomary et al. [12]	ACL	Conference	10
*Milad Alshomary and **Shahbaz Syed** and Arkajit Dhar and Martin Potthast and Henning Wachsmuth.* *Argument Undermining: Counter-Argument Generation by Attacking Weak Premises, 2021.*			
Al-Khatib et al. [7]	ACL	Conference	5
*Khalid Al-Khatib and Michael Völske and **Shahbaz Syed** and Nikolay Kolyada and Benno Stein.* *Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness, 2020.*			
Wachsmuth et al. [319]	ACL	Conference	10
*Henning Wachsmuth and **Shahbaz Syed** and Benno Stein.* *Retrieval of the Best Counterargument without Prior Topic Knowledge, 2018.*			
Ziegenbein et al. [346]	ACL	Conference	10
*Timon Ziegenbein and **Shahbaz Syed** and Felix Lange and Martin Potthast and Henning Wachsmuth.* *Modeling Appropriate Language in Argumentation, 2023*			
Al-Khatib et al. [8]	SIGDIAL	Conference	10
*Khalid Al-Khatib and Michael Völske and **Shahbaz Syed** and Anh Le and Martin Potthast and Benno Stein.* *A New Dataset for Causality Identification in Argumentative Texts, 2023*			

2

The Task of Text Summarization

This chapter presents a comprehensive background of text summarization, covering both the human and machine perspectives of the task. Initially, I provide a brief overview on the origins of summarization as a human task. In this context, I describe key insights from psycholinguistic studies on the cognitive processes involved in summarization. Here, I focus on how summarization influences the learning abilities of students, and the difference between novice and expert summarizers based on their summarization strategies. Next, I explore the task of automatic summarization, tracing its history, different summary types, and the various methods developed by the research community, with an emphasis on neural summarization methods. This part aims to provide a clear understanding of the evolution and advancements in automatic summarization techniques. Finally, I provide an overview of the evaluation of automatic summarization systems, discussing the metrics employed to assess the quality of summaries. Additionally, I describe the prominent challenges inherent in conducting qualitative evaluation of summaries, providing readers with a comprehensive understanding of the evaluation process.

2.1 Decoding Human Summarization Practices

The origins of summarization by humans can be traced back to the 15th century [1] from the Latin word *summa* meaning *totality*. A summary was intended to capture the core contents of a document so that knowledge could be easily shared and preserved. For instance, the Greek and Roman people

[1] https://www.etymonline.com/word/summary

created summaries of non-fiction works (epitomes) and fictional plays (hypotheses) that summarized them. Similarly, Egyptians are believed to have created summaries of their legal texts in the form of abstracts [328]. Summarization also plays a vital role in the development of learning capabilities in children as we will see later in this section.

While humans have been summarizing (implicitly and explicitly) for a long time, the era of digital information has formally motivated the need for automatic summarization as a tool. Specifically, for textual information that requires prolonged reading on digital interfaces such as computers and mobile devices, automatic summarization is urgently needed. Previously, the task of condensing information from documents was performed by librarians, who would scan the document and then append an arbitrary number of subject headings that reflected the subject contents of the document [89]. However, the deluge of information in the digital age (as well as from the printing press) has made this task infeasible for humans to perform at scale. This has encouraged scientists to develop computational models for automatic summarization of texts.

To gain a comprehensive understanding of automatic text summarization, it is essential to first delve into the process of human summarization. This involves grasping the cognitive model employed by humans when processing text, understanding the specific steps involved in creating a summary, and recognizing the varying difficulty of these steps based on the age and expertise of the human summarizer. Extensive research in psycholinguistics, focusing on learning and child development, has yielded formal models of summarization applicable to both children and adults.

In the following, I provide an overview of the essential findings from seminal psycho-linguistic studies on summarization. These studies have proposed theoretical models of cognition and discourse comprehension, while also conducting empirical research to examine the development of human summarization capabilities as they mature and the implicit strategies employed in summarizing texts. These studies offer a theoretical framework for distinguishing between high-quality and subpar summaries, as well as establishing a connection between manual and automatic summarization. Specifically, I elaborate on the following aspects: (1) the relationship between the importance of content and its recall, and how this influences summarization; (2) the discourse model of text comprehension and the parameters that affect summarization performance; (3) the differences in summarization strategies between experts and novices, and how inefficient readers can improve their summarization skills; and (4) the various

factors that influence the quality of summaries, such as text complexity, reader motivation, and the purpose of the summary.

2.1.1 Recall as a Function of Content Importance

Learning to summarize is a crucial skill that requires understanding the essential events in a text to be able to recall them later. Johnson [148] investigated the relationship between the importance of units and their recall by undergraduate students (of pyschology) at different time intervals. In their study, a folktale was broken down into plausible subunits or linguistic phrases, with instructions provided to the students that these subunits varied in their structural importance to the overall story. The students were then asked to eliminate some subunits to create a summary, and the number of times each linguistic unit was retained in the resulting summary provided a measure of its structural importance. Later, when the students reproduced the folktale, evidence that any portion of their reproduction was determined by one of the units was considered as evidence of the recall of that particular unit. This study found that the structural importance of a unit was significantly related to its recall. Specifically, learners were able to categorize verbal units in narrative texts based on their structural importance. This categorization was not solely due to additional learning time but appeared to occur even without knowledge of the nature of later occurring units. These findings suggest that good learners have an innate ability to identify and prioritize the essential events in a text, which can be useful in the process of summarization.

In a similar vein, Garner and McCaleb [108] studied the impact of text manipulations on the recall of important units by examining three specific operations: (1) *cuing*, the use of cue words to indicate importance in the text. This was further divided into semantic cuing, where explicit topic sentences were provided for paragraphs, and lexical cuing, which used words such as "important", "central", "key", etc., to indicate the importance of a unit; (2) *organization*, placing the most critical pieces of information at the beginning of the text or distributing them evenly throughout the text; and (3) *reduction*, limiting the summary length to a fixed number of sentences.

They concluded that cuing was the most effective of the three manipulations in improving summarization performance. Specifically, providing cues through either semantic or lexical means significantly improved the recall of important units. These findings suggest that cues play an important role in helping learners identify and prioritize essential information in texts, and can aid in the process of summarization.

2.1.2 A Discourse Model of Text Comprehension

In the realm of theoretical models of summarization, Kintsch and Van Dijk [163], Van Dijk et al. [306] proposed a model that describes the mental operations involved in recall and summarization of narrative texts. This model considers the semantic structure of a text to be described at both the local micro-level and the global macro-level. The microstructure refers to the structure of individual propositions and their relationships, while the macrostructure is the structure of the text as a whole. These two structures are related by a set of specific semantic mapping rules known as the *macrorules*. Specifically, four macrorules were proposed for systematically summarizing a text which are codified by the following macro-operators:

1. *Deletion*: The deletion of unimportant or trivial propositions.
2. *Generalization*: The superordination of lists of propositions by a general proposition denoting an immediate superset of the list. For e.g., the list of propositions *nose, hands, ears, legs* can be generalized to the proposition *body parts*.
3. *Selection*: The selection of a topic sentence if one is explicitly provided or indicated in the text.
4. *Invention*: The creation and use of a topic sentence that did not appear in the text but easily could have.

According to this model, successful summarization relies on the ability to identify the macrostructure of a text and the important information contained within it. The model has been influential in the development of subsequent theoretical models of summarization and has contributed to our understanding of the cognitive processes involved in this important skill. It further postulates that readers of narrative texts employ two distinct cognitive processes: (1) *construction-integration* processes, which involve constructing a mental representation of the text and integrating it with existing knowledge, and (2) *retrieval* processes, which involve accessing and retrieving information from memory. These processes are guided by the macrorules listed above, which help readers to identify the main idea and important details of the text.

With regards to operationalizing text comprehension, the model suggests that readers construct a *text base* that is composed of structured units connected through referential coherence, such as a linear or hierarchical organization of propositions with coreferential expressions. The *discourse topic* connects the various structured units in a global fashion. However, a

reader's specific goal is an essential prerequisite for modeling comprehension using the aforementioned macrorules. In practice however, these goals are often poorly defined, making it challenging to evaluate the generated summary, as it becomes highly subjective.

Parameters of the Theoretical Summarization Model

The macrorules model of text comprehension has three key parameters that influence the quality of the produced summary which helps to differentiate between *efficient* and *inefficient* readers. These parameters are: (1) n the input size per cycle, (2) s the capacity of the short-term memory buffer, and (3) p the reproduction probability of a text unit. A *cycle* here is the process of reading a part of the text and applying the macrorules to create a coherent text base. I describe these parameters in detail below.

The number of input propositions per cycle (n) can be affected by the familiarity of the reader with the discourse topic of the text, known as *apperception* which plays a significant role in their ability to comprehend the text. A strong knowledge base (background) allows for better comprehension and the processing of a greater number of propositions per cycle. In contrast, an unfamiliar reader may struggle to derive the same meaning and process fewer propositions.

The capacity of the short-term memory buffer (s) is also a key factor that influences the quality of the summary. The buffer is the working memory that stores the propositions that are currently being processed. Good readers can hold more text in their short-term memory than poor readers. A plausible reason for this is that persons with low verbal abilities are slower in accessing information or struggle due to the difficulty of the text.

Finally, the reproduction probability of a text unit (p) is the probability that a proposition will be reproduced in the summary or recalled at a later date by the reader. This is mainly affected by the importance (salience) of the proposition in the text. Additionally, it depends on the goal of the comprehension (purpose of the summary). If a long text is read with attention focused mainly on gist creation, the probability of storing individual propositions should be lower than when the same text is read with immediate recall instructions.

2.1.3 Summarization Strategies of Expert Readers

The task of summarization has been studied extensively as an indicator of a reader's ability to comprehend a text, particularly in the field of child development. Key findings from relevant research are outlined in Table 2.1.

Study	Findings
Brown et al. [44]	Younger students summarize primarily by deleting unimportant content or copying the contents verbatim. The resulting summary is partially adequate.
Garner [107]	Efficient students include higher proportion of important ideas, synthesize novel and faithful statements. Inefficient students retain unimportant information and but reject inconsistent ideas more often.
Brown and Day [43]	Expert summarizers combine information across paragraphs and invent new topic sentences. Also, they do not proceed sequentially for deleting or copying the contents.
Brown et al. [45]	Mature students outperform younger ones in the application of generalization operator, while the invention operator was the most difficult one to apply in both cases. Mature students also planned better and were more sensitive to fine-grained importance of the contents.
Winograd [329]	Good readers (students) were more aligned with adults in their perception of important content than poor readers. Good readers relized on both contextual (background) and textual cues to identify important contents, while poor readers only relied on textual cues.
Hare and Borchardt [126]	Poor readers were more likely to identify topic sentences based on their usual position (sequential ordering) in the paragraphs while fluent readers were able to locate important ideas throughout the text.

Table 2.1: A summary of key observations from psycholinguistic studies on the role of summarization in the development of learning abilities of children. Each finding aims to differentiate between effective and ineffective (readers) summarizers based on the strategies adopted as well as the use of the macro-operators from the cognitive model of summarization.

These provide an empirical basis for understanding the role of summarization in child development, as well as the common challenges associated with the task. In the following, I expand on some of the observations.

Evaluation of learning abilities in children is typically conducted using a three-step framework. Firstly, researchers select narrative texts that are of reasonable length (up to 800 words) and complexity, covering a range of subjects including geography, folktales, science, psychology, etc. Secondly,

a list of instructions is provided, which are codified by the macro-operators (*deletion*, *generalization*, *selection*, and *invention*) described in Section 2.1.2, as well as instructions for "polishing" the summary [126], such as paraphrasing, using connecting words, and adding introductory or concluding sentences. Finally, the efficiency of the readers is evaluated by comparing the summaries produced by different age groups (school children, undergraduate students, and graduate students) to the original text, with a focus on the usage of the four macro-operators and the development of summarization skills over time.

For instance, a study by Brown et al. [44] found that younger students (seventh grade and junior college students) tend to summarize texts primarily by deleting unimportant information or copying the text nearly verbatim. While this strategy produces a recognizably summary-like product, it is only partially adequate. Garner [107] examined 24 undergraduate students to determine the differences between high- and low-efficient summarization strategies. High-efficient students were found to include a higher proportion of important ideas and synthesized novel, faithful statements, while low-efficient students were more likely to retain unimportant information and reject inconsistent ideas. This suggests a "cost-benefit" phenomenon where effective summarization involves a trade-off between including important statements that increase retrieval performance and tolerating inconsistent ideas.

In a similar direction of research, Brown and Day [43] found that there is a significant difference between the summarization skills of "mature" and "expert" summarizers. While mature summarizers were able to combine information across paragraphs, expert summarizers were able to invent new topic sentences in addition that were not present in the original text, and they did not proceed sequentially for deleting or copying the segments.

Follow up study by Brown et al. [45] also found that younger students were outperformed by mature students in the application of the *generalization* operator, while the *invention* operator was found to be the most difficult operator to apply. This operator requires the summarizer to add a synopsis in their own words of the implicit meaning of the text, which is considered the essence of good summarization. Moreover, mature students were better at planning ahead, were more sensitive to fine-grained importance of textual units, and condensed more important units into the same number of words. This shows that the emergence of strategic planning is a gradual process related to age, as well as, that strategic action of selecting important contents and a semantic understanding of the text are closely related.

Connecting the notions of importance, age, and reading ability, Winograd [329] observed that good readers (students) were more aligned with adults in their perception of important content than poor readers. However, both groups were equally consistent in their judgments of what was important, despite having different views about which ideas in the text were significant. Good readers were able to identify importance using both contextual (their own background about the text's topic) and textual (importance indicators provided by the author) cues, while poor readers relied only on contextual cues. These findings suggest that reading ability plays a significant role in identifying important information in a text.

An often overlooked aspect is the *position* of contents in the text and its impact on the summary. The position of important contents impacts what poor readers chose to include in their summaries. Unlike fluent readers who identified important ideas throughout the text, poor readers relied heavily on the sequential order of the text, becoming less proficient at using their *own* perceptions of importance as the processing load increased.

This bias towards sequential text processing in poor readers was further investigated by Hare and Borchardt [126], who found that poor readers were more likely to identify topic sentences based on their *usual* position in the paragraphs. With regards to *sensitivity* to importance, fluent readers first used textual cues and background knowledge to identify important elements of various granularities which were then used to construct an internal representation of the text, effectively condensing its meaning in a few words. Poor readers had difficulty integrating individual propositions into larger semantic units, a skill that is critical for summarization. Sensitivity to importance thus reflects difficulties in text comprehension in children.

Summary Younger students notably tend to delete unimportant information or copy text verbatim, resulting in partially adequate summaries. Efficient summarizers include important ideas and generate novel, faithful statements, while less efficient ones retain unimportant details, however rejecting inconsistent ideas more often. This trade-off suggests a "cost-benefit" phenomenon. Further research distinguishes between “mature” and “expert” summarizers, with experts inventing new topic sentences and applying operators strategically. The difficulty of the *invention* operator, which involves adding a synopsis of implicit meaning, highlights its essence in good summarization. Reading ability also influences identification of important content, with good readers aligning more closely with adults. The impact of content position on summaries is evident, with poor readers relying heavily on sequential processing. Finally, sensitivity to importance

reflects comprehension difficulties in children as they struggle to integrate propositions into coherent summaries.

2.1.4 Establishing Summary Quality

Defining an *ideal* summary is challenging. According to Kintsch and Kozminsky [162], a good summary is one on which everyone agrees. In other words, if different people produce different summaries of the same text, it suggests a lack of clarity about the *purpose* of summarization. However, achieving consensus on what constitutes a good summary is difficult because it depends on various implicit and often ill-defined (contextual) factors such as the readers' ability, background knowledge, goal, time delay between reading and recall, and the target audience for the generated summary. These factors, as identified by Jones [150], are crucial for both developing and evaluating automatic summarization systems. The goal of human assessment of summary quality is typically to achieve high agreement from multiple judges on the importance of the information included in the summary.

To summarize this section on the human perspective of the task of summarization, expert readers employ several strategies to effectively summarize text such as using cues (lexical and semantic indicators) to identify important contents, combining information from different parts of the text to synthesize novel and coherent summaries. Next, model parameters such as input size, short-term memory capacity, and reproduction probability are also important factors that influence the summarization process. Finally, composing a *good* summary that is agreed upon by *most* people is affected by context factors such as reader's abilities, background knowledge, purpose (goals), and the target audience. These findings from psycholinguistics provide a strong basis for developing and evaluating automatic summarization systems that can emulate the strategies of expert readers.

2.2 Exploring Automatic Summarization Methods

The task of automatic summarization was first developed to create *abstracts* of scholarly documents to help index the rapidly growing literature [25, 186]. The first approaches to automatic summarization were based on a simple yet highly effective heuristic that involved identifying frequent terms (words) and extracting the sentences containing them. Since then, the research community has introduced many robust and novel methods for creating automatic summaries of a variety of documents from diverse

domains. Given the previously described insights from psycholinguistic studies on child development about the differences between "novice" and "expert" readers (summarizers) and how their summaries differ in quality, we can align their strategies to the two broad categories of automatic summarization: *extractive* and *abstractive*.

Novice readers mostly summarize via a *copy-delete* strategy by processing the text in a sequential fashion. This aligns with the *extractive* paradigm of automatic summarization where important sentences are selected from the original text and concatenated to form the summary. In contrast, expert readers are able to first gain an understanding of the text, combine information across paragraphs, and invent novel topic sentences to create a summary. This aligns with the *abstractive* paradigm of automatic summarization where the summary is generated by synthesizing the information in the original text and summarizing it fluently (human-like). Thus, abstractive summarization is a more challenging task than extractive summarization, and is crucial to develop sophisticated artificial intelligence systems that can understand and summarize texts.

The following sections concretely describe extractive and abstractive summarization methods, followed by a brief overview of the neural summarization methods which form the basis of the research focus of my thesis and modern summarization technology in general.

2.2.1 Extractive Summarization

Extractive summarization is the process of identifying important information in a text and producing them verbatim to form a summary. It is mainly concerned with the *content* of a summary rather than its *form*. This is evident by the explicitly stated goals of the seminal works on automatic summarization of technical literature: to produce ''abstracts" or "gists" that can be used for efficiently indexing the vast amounts of technical literature" [25, 89, 186]. These works focused on statistical features of the text such as frequency of terms, sentence position, cue words, document skeleton, and sentence length to identify important sentences. These methods were unsupervised and required domain-specific hand-crafted features to be effective.

With increasing amounts of data being easily available, trainable summarizers were introduced that learnt from the data [15, 169]. These methods could classify important sentences based on features such as term frequency (tf), inverse document frequency (idf), and sentence position. However, these methods were limited by the fact that they were unable to capture the relationships between sentences. To account for this, hidden markov

models (HMMs) were used to model the local dependencies between sentences [69]. In a different approach, Marcu [194] leveraged the discourse rhetorical structure of the documents to identify key sentences based on the relative importance of their rhetorical relations.

Other influential approaches include *centrality* based methods that use the cluster centroids [241] or the graph structure of the text to identify important sentences [201]. These methods are based on the assumption that important sentences are more likely to be connected to other important sentences. The centrality of a sentence is measured by the number of other sentences that are connected to it. These methods are primarily unsupervised and can hence be applied on a wide range of domains, especially in which large amounts of training data is unavailable for developing supervised summarization models. For a more comprehensive overview of all the extractive summarization methods, see [74, 212].

2.2.2 Abstractive Summarization

Abstractive summarization also aims to generate a summary of the source text, however focussing on the *form* of the summary in addition to its content. In contrast to an extractive summary that can be incoherent, poorly structured, or show a significant information loss due to compression, an abstractive summary is expected to be fluent, coherent and semantically accurate. A key feature of abstractive summarization is to introduce *novel* words into the summary either via paraphrasing, using synonyms, or combining information across sentences into concise phrases. This is in contrast to extractive summarization where the summary is a verbatim copy of the original text. Evidently this is a much harder task and did not gain traction in the early research era of automatic summarization.

First approaches to abstractive summarization aimed to generate natural language summaries in two broad ways:(1) using prior information by combining multiple sources, and (2) using natural language generation (NLG) systems. In the first approach, additional information about a certain topic (say a news event) was gathered from multiple related articles to obtain a discourse structure of the events. This allowed generating coherent summaries of these documents where information was first extracted, combined, and then realized into natural language sentences [124, 240]. In the second approach, after identifying and arranging the important content in a preestablished conceptual order, information was merged via discourse markers, deletion, verb transformation etc., to output a natural language summary [145, 258]. Follow up approaches focused on automatic

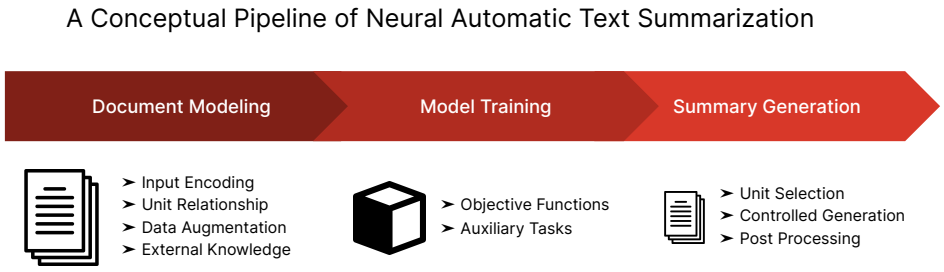


Figure 2.1: A conceptual pipeline of (neural) automatic text summarization. Document modeling considers the various units of selection (words, sentences, paragraphs), the relationships between these units, augmenting the document (paraphrasing, translating), and leveraging external knowledge to enhance the input. Next, model training encompasses objective functions to teach the model to summarize or using pretrained models that perform auxiliary tasks to boost downstream task effectiveness. Finally, summary generation deals with which units to select, constraining the summary to a specific aspect or style, and post-processing the summary to make it fluent.

sentence compression by solving optimization problems to select the words to be dropped from a sentence [67], using lexicalized markov grammars with edit word detection [179], or sub-graph detection to remove redundant information [105].

2.2.3 Neural Text Summarization

Before the advent of deep learning, the task of automatic summarization was dominated by extractive methods [74, 190, 242]. A transformative moment for automatic summarization was the introduction of sequence-to-sequence (`seq2seq`) neural network models [283] that can learn (in an end-to-end fashion) from large amounts of labeled data to transform a sequence of tokens (source document) to another shorter sequence of tokens (its summary). With the large amounts of easily available web documents (primarily news articles) accompanied by automatically extracted summaries, abstractive summarization has seen a significant advancement in the last decade. Figure 2.1 depicts a conceptual pipeline of automatic neural text summarization which is comprised of multiple components that accept a document as input and generate a summary in an end-to-end fashion. Described below are the key components of this pipeline:

1. *Document Modeling* that encodes the source document into a vector representation, modeling the inter-unit relationships where a unit can be a word, sentence, or a paragraph, augmenting the input data with additional (user-specific or style-specific) information, and leveraging external (domain) knowledge to enrich the representation.

2. *Model Training* that includes learning the parameters of the model using a large-scale dataset via custom objective functions, and/or leveraging pretrained models to perform auxiliary tasks and finetuning them to improve the summarization effectiveness (such as missing text prediction, paraphrasing, or textual entailment).

3. *Summary Generation* that includes selecting the summary-worthy units (word, sentence, or a paragraph) from the source document, constraining the summary to a specific aspect/user-style (controlled generation), and finally post-processing to make the summary fluent and/or filter out undesired content.

This conceptual pipeline is in alignment with the sequence-to-sequence paradigm of text generation that forms the core architecture of neural summarization models. Some of the prominent neural summarization models are described in the following section.

2.2.4 Overview of Neural Summarization Methods

Early neural approaches to text summarization were focused on generating abstractive summaries of sentences and paragraphs. Chopra et al. [64], Rush et al. [257] applied `seq2seq` models to automatically generate headlines of news articles. Nallapati et al. [206] introduced the CNN/DailyMail corpus, a collection of news articles accompanied by (mostly extractive) multi-sentence highlights as summaries. This shifted the focus of the summarization community from generating single sentence summaries to multi-sentence summaries. Significant improvements were made to the standard `seq2seq` model by the neural machine translation community that were quickly integrated and adapted for neural text summarization such as: (1) the *attention* mechanism [18] which allows the model to focus on different parts of the source text while generating the summary, (2) the *pointer* network [312] that allows the model to copy words via pointing to source text or a vocabulary of words, (3) the *copying* mechanism [119] that allows the model to copy words from the source text to the summary in case of uncertainty in the generation process, and (3) the *coverage* mechanism [303]

that prevents the model from repeating words in the summary by maintaining a coverage vector (history of attended words) of the source text, (4) the *pointer-generator* network [270] that combines aforementioned mechanisms to generate a summary that allows to switch between the generation of novel words or copying of words from the source text, and finally (5) the *Transformer* model [309] that uses self-attention to efficiently model the relationships between words in a sentence, regardless of their respective position, eliminating the computational drawbacks of sequential processing for large texts.

These augmentations of the standard `seq2seq` model led to a significant improvement in the quality of the generated summaries, overcoming the limitations of the early neural summarization methods such as repetition of words, inability to generate out-of-vocabulary words, and inability to copy words from the source text.

2.2.5 Large Language Models for Neural Summarization

With the introduction of Transformer-based pretrained models such as BERT [83], BART [173], GPT [46], and T5 [245] (to name a few) the summarization community has seen a significant shift towards adapting these models for summarization. These models, also known as "foundational models" [37] are pretrained on large amounts of data and multiple downstream tasks allowing them to learn a richer representation of the source text and its semantics via *transfer learning* [230], in comparison to end-to-end training on a specific corpus. Moreover, they can be fine-tuned on a domain-specific summarization task with a relatively small amount of data, increasing their adoption to low-resource domains. While the text quality of the summaries generated by these models is significantly better than the `seq2seq` models, they often introduce new challenges such as subtle hallucination of facts, improper coreferences, and variable levels of abstraction that impact the suitability of automatic evaluation metrics for evaluating such summaries [140]. Currently, with the rapidly advancing field of large language models, the summarization community is actively exploring the use of these models for zero-shot and few-shot summarization, eliminating the need for constructing large-scale training datasets. For instance, Goyal et al. [113] demonstrated that GPT-3 can generate high quality abstractive summaries of new reports in zero-shot setting (i.e., without any finetuning) that are overwhelmingly preferred by human judges over the summaries generated by the state-of-the-art supervised summarization models.

The majority of the neural summarization methods described above have focused on structured and easily accessible news articles, where *extractive* summaries serve as the primary ground truth. In the context of news reports, the inverted pyramid style is commonly used, presenting key information in the initial sentences, known as the *lead* [237]. This structure simplifies the task for models, as they can easily identify the most important sentences and extract them to form the summary. However, this approach tends to favor the lead and may not work as effectively when applied to other document domains, resulting in lower summary quality [152, 158]. In contrast to news articles, user-generated content, such as social media posts, web pages, argumentative discussions, and opinions, lacks a clear and organized structure that models can readily leverage. This poses a challenge for neural summarization methods in these domains and highlights the importance of investigating their performance to develop robust summarization models.

Evaluating the effectiveness of these models also requires considering summary purpose and the relationship between the source document and summary. Moreover, establishing *true* state-of-the-art in automatic text summarization requires that the methods are extensively evaluated and compared with each other on a large number of datasets. However, the lack of a standardized evaluation framework and the use of different evaluation metrics across different datasets makes it difficult to compare the performance of different summarization methods. In the following section, I discuss the evaluation of summarization methods and the challenges associated with it.

2.3 Evaluation of Automatic Summarization and its Challenges

Evaluating automatic summarization poses significant challenges primarily due to the lack of an objective definition of what constitutes a *good* summary. Moreover, it is often unclear what the intended *purpose* of a summary is and how it will be *used* to solve a downstream task, if one exists. Finally, little is known about who is the target *audience* for the generated summaries. Much of the summarization done by humans is often *reflective* in nature [150], which means that it often serves the default purpose of presenting prominent source content; this implicit goal allows such summaries for any use. In this section, I first describe the various methods and criteria used for evaluating automatic summarization, followed by the various challenges and pitfalls as outlined in Table 2.3.

2.3.1 Methods and Criteria for Evaluation

Early approaches to summarization were predominantly extractive in nature. Consequently, the evaluation of these approaches revolved around assessing the overlap between the words in the generated candidate summary and the reference summary. Reference summaries often originated from professionals, such as in the scientific domain, where authors themselves (or professional summarizers such as librarians) provided condensed abstracts of their research. Pollock and Zamora [232] identified four broad categories of evaluation metrics employed during this period:

1. ***Intuitive***: In this method, human judges are asked to rate the quality of the generated summary.

2. ***Statistical***: In this method, the generated summary is compared to the reference summary using a statistical measure such as the overlap between their words. This is relatively easier to compute via automatic metrics and is less expensive than the intuitive method.

3. ***Computational***: While similar to the statistical method, this category of methods compare generated summary with the information content of the source document.

4. ***Functional***: This category of methods grounds the summary evaluation in specific tasks such as asking humans to answer questions based on reading the summary vs. the source document, index term content and retrieval capabilities, relevance prediction, and if the user needs to read the source document after reading the summary.

The first three categories, namely *intuitive, statistical,* and *computational*, are commonly referred to as *intrinsic* methods of evaluating summaries. These methods assess the quality of summaries in isolation, considering criteria such as "coherence", "readability", "informativeness", "coverage", "concept capture", "faithfulness", and "factuality".

On the other hand, the *functional* category is recognized as an *extrinsic* evaluation approach. It evaluates summaries in the context of a downstream task, such as information retrieval or relevance prediction. These tasks involve comparing the speed and accuracy with which users can identify relevant items from a list of retrieved documents based solely on their summaries, as opposed to reading the full text. Other related tasks may include question answering or assessing text comprehension based on the

generated summaries. The quality criteria assessed in these evaluations include the "usefulness" or "responsiveness" of the summary to the specific task at hand [71, 191, 299].

Automatic Evaluation

Intrinsic evaluation as described above lends itself to automatic (*quantitative*) evaluation via repeatable and computationally efficient metrics that offer quick feedback on summary quality to developers. The research community has introduced several metrics that compare the generated summaries (candidates) to the reference summaries (or the source documents). These comparisons encompass the following approaches: (1) computing lexical or semantic overlap of the content, (2) greedy alignment of tokens, words, or phrases to maximize similarity scores, (3) answering questions using the summary about important concepts (derived from the source document or reference), and (4) estimating the conditional probabilities of large (pretrained) language models to generate the candidates, given the source document or the reference summary. Table 2.2 outlines the various metrics employed for quantitative evaluation of text summarization.

Human Evaluation

While automatic metrics provide quick approximations of the information quality captured by the generated summaries, they are often insufficient at assessing the text quality of the summary [71]. Furthermore, some of these metrics demonstrate poor correlation with human judgments of summary quality or fail to provide a complete picture [29, 97]. Thus, it is necessary to incorporate humans, specifically end users (if possible) in the evaluation process to obtain a *qualitative* assessment of summary quality. Qualitative evaluation can be conducted in both intrinsic and extrinsic scenarios. Often, crowdsourcing of (curated) non-experts is employed as a viable alternative to collecting expensive expert judgments.

In intrinsic evaluation, human judges assess the linguistic quality criteria outlined by the Document Understanding Conferences (DUC) [71, 72, 73]. While DUC initially focused on multi-document summarization, their human evaluation criteria are also applicable to the single-document summarization setting. Specifically, summary quality is manually evaluated along the following dimensions:

1. ***Grammaticality***: The summary should have no system-internal formatting errors or obviously ungrammatical sentences that affect its readability.

2. ***Non-redundancy***: The summary should not contain unnecessary repetition such as entire sentences, facts, or nouns.

3. ***Referential clarity***: The summary should not contain pronouns or other referring expressions that are unclear or ambiguous.

4. ***Focus***: The summary should only contain information that is related to the rest of the summary.

5. ***Structure and Coherence***: The summary should be well-structured and well-organized instead of being a heap of related information.

In addition to these criteria, researchers also manually evaluate other (often vaguely defined) dimensions such as "informativeness", "relevance", "overall quality", "conciseness", "truthfulness" to name a few [138].

2.3.2 Challenges in Evaluation

Despite the several kinds of quantitative and qualitative evaluation methods described above, evaluation of summarization systems is still an open problem. Table 2.3 outlines the key challenges as described by literature. Relevant excerpts from the literature are provided below for each challenge:

1. ***Subjectivity and Effort***: Human evaluation is subjective, time consuming, expensive and does not scale to a large number of summaries [191]. There are two primary sources of variation in summaries written by humans: content variation such as summary focus and style, and a general disagreement as to how well a system summary covers the human summary [128]. Humans are quite consistent with respect to what they perceive as being the most important and the most unimportant but less consistent at what they perceive as being *less* important [146].

2. ***Inter-Annotator Agreement***: Inter-annotator agreement (IAA) decreases as the summary length increases [191]. Low IAA is seen as a barrier to reproducible research and to drawing generalizable conclusions. However, this is not necessarily true and is highly dependent on the subjectivity of the task, for instance, judging relevance often leads to low kappa values [178].

3. ***Incomplete Ground Truth***: There can be multiple good summaries for a given document and it is difficult to define a single gold standard summary; the set of reference summaries is *necessarily incomplete*. Moreover, in the absence of a clear goal, the machine-generated summary is possibly a good summary that is quite different from any human summary used to evaluate it [191]. Reference summaries used for automatic evaluation can also be of low-quality containing extraneous information such as hyperlinks and click-bait [97]. Prescriptive attempts to define what a good answer or summary *should be* will lead to systems that are not useful in real-world settings [178]. Most human generated summaries were incomplete or included redundant/irrelevant information. [183]

4. ***Variable Summary Lengths***: As summarization is a compression task, machine-generated summaries must be evaluated at different compression rates. However, this increases the scale and complexity of the evaluation task [150]. For an "ideal" summary based evaluation, accuracy decreases as summary length increases, while for task based evaluations summary length and accuracy on information retrieval appear to correlate randomly. Evaluation results for the same summarization system can be significantly different if summaries are cut at different length [146]. Summaries of different length produced by the same system have a clear non-linear pattern of quality as measured by ROUGE: initially improving steeply with summary length, then starting to gradually decline. Neural models produce summaries of different length, possibly confounding improvements of summarization techniques with potentially spurious learning of optimal summary length. Humans prefer shorter summaries in terms of the verbosity of a summary but overall consider longer summaries to be of higher quality [282].

5. ***User Needs***: The user's or application's needs must be taken into account which complicates the evaluation task [150].

6. ***Quantification of Content***: The notion of information content is hard to quantify and involves answering questions such as: (1) In what units should the information be expressed? (2) How should the information be weighted (old, new, trivial, important)? (3) Should only explicit information be counted or should implicit (logically entailing) information also be counted? (4) How would the numerical value for each content be computed? (5) Do the quantitative values depend on

the reader or the author of the document that is being summarized? [232]

7. ***Annotation Guidelines***: Clear instructions and interventions are necessary for reliable expert evaluation [142]. Attempting to raise agreement by rigid assessment guidelines, may do more harm than good [178]. Annotators necessarily do not follow the guidelines even if they are clearly defined and exhibit a random behavior in creating summaries [183].

8. ***Annotator Expertise***: Annotation expertise affects the label quality as non-experts tend to disagree significantly with the experts [111, 142, 183]. Using crowdsourcing for collecting ground truth summaries may be inefficient in two ways: first a lot of time needs to be invested in verifying that the results are in fact summaries, and second most workers do not care about the task and may be more focused on the money, trying to complete the task as quickly as possible [183].

9. ***Annotation Design and Reporting***: The standard setting of three annotators per label is insufficient in case of crowdsourced evaluations with non-experts [142, 280]. There is very little shared practice in human evaluation in NLG, in particular on naming the quality aspects to be evaluated and how to define them [138]. Higher compensation of crowdsourced workers may yield lower quality work as it attracts people wanting to make quick money. Non-experts have difficulty distinguishing linguistic quality from content. Readability is more important to non-experts, or at least easier to identify [111, 183]. Evaluation study parameters such as the overall number of annotators, distribution of annotators to annotation items are often not fully reported. Subsequent statistical analysis ignores grouping factors arising from one annotator judging multiple summaries [280]. Annotators are biased in favor of anything that makes scoring easier such as the extractiveness of the summary and length of the summary [284, 286, 347]. After reading a summary, an annotator may choose not to review carefully the whole text, but to consider in detail only the parts that look most similar to the summary from the text [307].

10. ***Scoring Range***: The *width* of the scoring range (low, average, high) affects inter-metric correlation. Metrics agree in ranking summaries from the full scoring range but disagree in ranking summaries from low, average, and high scoring ranges when taken separately [28].

11. ***Summary Type***: Inter-metric correlation is higher for ranking extractive summaries but lower for abstractive summaries [28]. Extractive summaries also suffer from unfaithfulness problems such as incorrect/incomplete coreference, incorrect/incomplete discourse, and misleading by selection resulting in media bias [340].

12. ***Corpus Type***: With new datasets and methodologies constantly evolving, conclusions about old metrics such as ROUGE do not hold anymore. Different metrics are better suited for different corpora [29].

13. ***System Rankings***: There is low inter-annotator correlation of system rankings based on recall measures from non-identical reference summaries [84]. Metrics cannot reliably quantify the improvements made by one system over the others, especially for the top few systems across all datasets [29].

14. ***Time Variance***: Scoring responsiveness or summary quality is *time variant*. Humans give different scores to the same summaries over a period of time [68].

15. ***Correlation with Human Judgments***: Confidence intervals for correlations of automatic metrics with human judgments are rather large, implying a large amount of uncertainty about their reliability and precision. Moreover, metric correlations are not evaluated in a realistic setting where multiple similar quality systems are compared. Correlation of ROUGE to human judgments is *near zero* in realistic scenarios for cases where systems are separated only by a small difference in their automatic scores (as is commonly observed in practice) [81, 82]. They also have weak or moderate correlation with the relevance dimension due to the difficulty in defining the concept of relevance and collect reliable human judgments for it [97].

16. ***Metric Sufficiency***: Scores from popular metrics such as ROUGE and BERTScore largely cannot be interpreted as measuring information overlap. Rather they are better estimates of the extent to which the summaries discuss the same topics. Thus these metrics cannot measure if summaries contain high-quality information or not [79]. Most metrics correlate poorly (weak or moderate) with the coherence dimension. This is because majority of the metrics rely on hard or soft sequence alignments, which do not measure well the interdependence between consecutive sentences [97]. ROUGE fails to accurately measure factual inconsistency across domains [102]. Com-

monly used variants of ROUGE may be sub-optimal for automatic evaluation. Combining different variants together results in an evaluation metric that is extremely competitive [115, 248]. Using ROUGE for evaluating extractive systems forces SOTA to strive towards perfect scores that are theoretically and computationally hard to achieve [266]. Current automatic reference-based metrics cannot be used to reliably measure summary quality under the zero-shot prompting paradigm; same with reference-free metrics. They cannot produce a ranking similar to human preferences in the zero-shot setting with GPT3 that exhibits different properties with respect to the summary style compared to supervised models [113].

17. ***Sample Size***: System level correlations of automatic metrics with human judgments are significantly affected by the sample size used to evaluate them. More accurate estimates of metric correlations need collecting more high-quality human judgments of summaries [82].

In Chapters 8 and 9, I propose tools to mitigate some of these challenges by leveraging visual analytics for a more unified as well as transparent evaluation process. Summary Explorer (Chapter 8) contextualizes corpus-based evaluation of multiple system summaries in relation to the source document. This allows evaluators to easily identify hallucinations, position bias, and informativeness of the summaries. Summary Workbench (Chapter 9) provides a unified interface for generating and evaluating summaries of any text via reproducible artifacts of the state-of-the-art summarization models. Additionally, it provides easy access to a suite of automatic evaluation metrics for easily reporting results and comparing models.

2.4 Summary

This chapter offered a comprehensive exploration of text summarization from both human and machine perspectives. Drawing on seminal psycholinguistic studies, the distinctions between expert and novice summarizers are highlighted, revealing how different strategies are employed to comprehend and convey the core content of a text. Experts excel in identifying crucial information throughout the text and skillfully synthesizing it into original and faithful topic statements. In contrast, novices tend to employ a copy-and-delete approach, resulting in partially effective summaries and susceptibility to positional biases in the text. These insights lay a robust foundation for the development and assessment of automatic summarization systems that emulate expert reading strategies. The discussion

then delved into automatic summarization within the context of two main categories: extractive and abstractive summarization. Notably, abstractive summarization closely aligns with expert strategies, making it a more intricate task than extractive summarization. Finally, an overview of evaluation methodologies and their associated challenges is provided, revealing the necessity for a unified and transparent evaluation process that can harness visual analytics to address the limitations of automatic evaluation metrics.

Building on the insights gained from this overview of automatic text summarization, the subsequent chapters of this thesis delve into my contributions focused on summarizing user-generated discourse, a domain that has received comparatively limited attention in summarization research. These chapters address the challenges previously discussed by introducing innovative resources such as datasets, methodologies, and evaluation approaches. The primary goal of these contributions is to improve the effectiveness and practicality of neural summarization techniques while offering meaningful perspectives for the ongoing progress of this field.

Metric	Content Unit	Reference
Lexical Overlap		
BLEU [225]	n-gram	✓
ROUGE [176]	n-gram	✓
METEOR [20]	unigram	✓
Basic Elements [137]	phrase	✓
Pyramid [213]	phrase	✓
CIDEr [310]	n-gram	✓
CHRF [235]	n-gram	✓
Semantic Similarity		
Greedy Matching [256]	word	✓
ROUGE-WE [215]	n-gram	✓
MoverScore [344]	n-gram	✓
Sentence Mover's Similarity [65]	sentence	✓
BERTScore [341]	token	✓
SUPERT [106]	token	✗
BARTScore [336]	–	Optional
Question Answering		
QA-based Evaluation [58]	text	✓
APES [96]	–	✓
SummaQA [268]	–	✗
FEQA [88]	–	✗
QAGS [323]	–	✗
QAEval [80]	–	✓
QuestEval [269]	–	✗
LLM-based		
BLEURT [271]	–	✓
BLANC [308]	–	✗
GPTScore [100]	–	Optional
G-Eval [181]	–	Optional

Table 2.2: Automatic metrics for evaluating summarization categorized by the respective paradigm adopted for measuring content overlap between the reference and the generated summary. Detailed description of each metric is provided in Appendix Table A.3.

Dimension	Mode	Orientation	Quality Criteria
Subjectivity & Effort	Qualitative	Intrinsic	Overall Quality
Inter-Annotator Agreement	Qualitative	Intrinsic	Importance, Relevance
Incomplete Ground truth	Both	Intrinsic	Informativeness, Overall Quality
Variable Summary Lengths	Both	Intrinsic	Informativeness, Abstractiveness, Overall Quality
User Needs	Quantitative	Intrinsic	Overall Quality
Quantification of Content	Quantitative	Intrinsic	Informativeness, Coverage
Annotation Guidelines	Qualitative	Intrinsic	Coverage, Overall Quality
Annotator Expertise	Qualitative	Intrinsic	Coverage, Readability, Overall Quality
Annotation Design & Reporting	Qualitative	Intrinsic	Overall Quality
Scoring Range	Quantitative	Intrinsic	Coverage
Summary Type	Quantitative	Intrinsic	Abstractiveness
Corpus Type	Quantitative	Intrinsic	Coverage
System Rankings	Quantitative	Intrinsic	Overall Quality
Time Variance	Quantitative	Intrinsic	Responsiveness, Overall Quality
Correlation with Human Judgments	Quantitative	Intrinsic	Overall Quality
Metric Sufficiency	Quantitative	Intrinsic	Informativeness, Coherence, Relevance, Abstractiveness
Sample Size	Quantitative	Intrinsic	Overall Quality

Table 2.3: Challenges in evaluating summaries categorized by the mode of evaluation (**qualitative, quantitative**), the orientation for evaluation (**intrinsic, extrinsic**), and the targeted summary quality criteria.

3

Defining Good Summaries: Examining News Editorials

This chapter builds upon the summarization theory described so far by taking the first step in the process of automatic summarization: defining what an *ideal* summary must look like for a given document. This entails clearly defining the quality dimensions that a summary must adhere to, and then operationalizing this definition in a manner that allows human annotators to create high-quality ground truth with ease. To exemplify this process, the present chapter focuses on the task of defining and gathering high-quality summaries of news editorials, which are opinionated texts that typically do not come with an author-provided summary, but rather a *lead* that intend to attract the readers' attention. We first establish a set of quality dimensions for editorial summaries and then proceed to collect a corpus of high-quality summaries of news editorials via crowdsourcing. Following this, we conduct a detailed analysis of the corpus to discern content-specific differences between high-quality and low-quality summaries. The chapter concludes with an evaluation of two unsupervised extractive summarization models on the newly collected corpus.

3.1 Key Characteristics of News Editorials

News summarization has been, and still is subject of active research to this day [177, 275, 334]. However, the inverted pyramid structure of news reports, where the lead paragraphs often have a summarizing quality in and of themselves [237], induces a bias in many recent news summarization approaches to just copy the opening sentences [152, 158]. Hence, these

approaches fail at argumentative news articles, such as editorials, whose structure differs from that of news reports.

Editorials represent the views of an organization (newspaper) on long-standing societal issues and aim to shape public opinion [99, 103]. Compared to news reports, which aim to inform objectively about current events, editorials subjectively assess a controversial topic in order to persuade its audience of a specific stance toward it [141, 305]. This difference in their goals leads to a difference in linguistic choices. An editorial is usually composed of three discourse parts, namely lead, body, and conclusion [252, 304]. The lead introduces the issue at hand by starting with an anecdote or a question. The body elaborates on arguments and background information, while the conclusion provides an evaluation as well as (possibly) implicit suggestions and calls to action [35]. The summary of an editorial must hence be constructed with care in order to preserve its argumentative structure and its persuasive means. Research on automatic (news) summarization has so far neglected argumentative texts in general, and the genre of editorials in particular. With this paper we contribute to closing this gap, taking effective steps toward editorial summarization:

1. We define an annotation scheme tailored to editorial summaries, requiring a high-quality summary to be thesis-indicative, persuasive, reasonable, concise, and self-contained.

2. We create a corpus of 1330 summaries for 266 news editorials (five summaries each), manually acquired and evaluated by operationalizing the proposed annotation scheme.

3. We analyze each summary of the corpus with respect to content overlap, distribution of evidence types, adherence to the editorial structure, and annotator indications regarding summary quality.

4. We evaluate two unsupervised, extractive summarization models (four variants total) in comparison to the acquired references, and their potential to identify an editorial's core message (the thesis).

The evaluation indicates a high suitability of the corpus for research and development: For 90% of the editorials there are at least three high-quality summaries, and for 52% all five are. The analyses also reveal that multiple summaries can be collected for an editorial with low content overlap, that good summaries include more *third party evidence* to justify an editorial's thesis, and that editorials' summaries have a distinct structure compared to

those of news reports, with a specific contribution from each editorial discourse unit (lead, body, and conclusion). The corpus and other resources are publicly available.[1]

Related Work

The summarization of argumentative text has hardly been studied: Egan et al. [90] automatically summarize online political debates by extracting key content from their arguments as "points" (verbs and their syntactic arguments). Similarly, Bar-Haim et al. [22] propose to map crowdsourced arguments to "key points." They created the ArgKP corpus, containing 24,000 ⟨argument, key point⟩ pairs, extracted from the IBM-Rank-30k dataset [116]. To the best of our knowledge, no argument corpus comprises summaries of long-form monological argumentative text.

Most of the commonly used corpora for automatic news summarization, such as the NYT corpus [259], Gigaword [208], CNN/DailyMail [132, 206], XSum [209], and NEWSROOM [118], primarily consist of (non-argumentative) news reports and only one ground truth summary per report. Although the DUC shared task datasets [220] provide multiple summaries per document (500 news reports), they are very short (up to 14 words or 75 bytes), similar to Gigaword and XSum (up to two sentences). These corpora, stemming from the news domain, may contain some editorials; the ones in the NYT corpus were studied by Li et al. [174], Al-Khatib et al. [6], El Baff et al. [91], and El Baff et al. [92] for tasks such as summarization, analysis of rhetorical strategies, and argumentation quality assessment. But Li et al. [174] observes that the accompanying summaries in this corpus are teasers rather than actual summaries. In our work, we focus on composing summaries exclusively for news editorials by aiming to capture their core argumentation, providing multiple and comparably longer ground truth summaries (20% of an editorial's segments) for each editorial.

A scheme for annotating argumentative roles of sentences for summarizing research articles was presented by Teufel et al. [295]. However, they only analyzed the effectiveness of this scheme and did not collect or evaluate any summaries. The key difference between other news summarization corpora and ours is the use of a (genre-specific) annotation scheme that unifies the summary acquisition and evaluation. Other summarization corpora lack such unification and only adopt the notion of "salience" (importance) of sentences in a text [231] to automatically extract summaries or crowdsource their acquisition [93]. In the absence of a human-written ground truth,

[1] https://webis.de/publications.html#?q=COLING+2020

parts of a text, such as the title, highlighted sentences, or lead sentences, are used as proxies for summaries of the source documents. While such heuristics help create large corpora, the infeasibility of evaluating all the ground truth summaries leads to increased noise in the datasets, severely limiting the task of summarization and its evaluation [167]. We evaluated each summary in our corpus for its quality and provide labels for high (low) quality per quality dimension defined in our annotation scheme.

3.2 Operationalizing High-Quality Summaries

As per Hidi and Anderson [133], humans produce two types of summaries: writer-based ones and reader-based ones. A writer-based summary is produced to facilitate one's own comprehension of a text. A reader-based summary intends to inform others about a text's core message, possibly to evoke further interest in the reader. News, in particular, may also be accompanied by a teaser, namely an incomplete summary that aims to attract people to read the entire news article (report or editorial) [174]. In extreme cases, teasers can become clickbait [236], constructed to manipulate their readers to visit an online news article (e.g., by invoking strong curiosity). For our corpus, we strive for reader-based summaries.

The intention of a reader-based summary is often to substitute the original text. For informational texts, such as news reports, this is roughly performed by the omission of irrelevant sentences (deletion), the subsumption of details into higher-level categories (generalization), and the integration of details into topic sentences (construction) [163]. Reorganization and rewording are possible [147], but new ideas must not be introduced [43, 163]. In this regard, an editorial aims to persuade its readers of one central claim (thesis) through its monological argumentation [5, 318]. It is composed of argumentative discourse units (ADUs, typically statements) that form arguments to support the thesis [226, 279]. These arguments implement the author's strategy, incorporating not only logical, but also emotional and credible means of persuasion [16]. The core message of an editorial corresponds to its thesis and its most persuasive segments; thus, an editorial's summary—and that of long argumentative texts in general—should aim to preserve both. We propose an annotation scheme tailored for editorial summaries, defining five quality dimensions that emphasize argumentation as well as summarization quality:

1. *Thesis-indicativeness..* The thesis of an editorial can be stated as a call for action or as an opinion [304]. The summary should thus explicitly contain the thesis or indicate it.
2. *Persuasiveness..* As the goal of an editorial is to persuade, the same applies to its summary. As per Wachsmuth et al. [315], the summary should aim to be effective (i.e., aim to persuade the target audience of its thesis).
3. *Reasonableness..* The summary should help its audience to reach the thesis and rebut plausible counter-arguments to it.
4. *Conciseness..* A summary should be significantly shorter than the editorial and lack any superfluous phrasing or information.
5. *Self-containedness..* A summary should be comprehensible with general knowledge, without referring to additional resources.

Altogether, we strive to compile a corpus of editorial summaries that come close to the following definition:

> A high-quality summary of an editorial *indicates its thesis*, argues for this thesis in a *persuasive* and *reasonable* manner, and is *concise* yet *self-contained*.

Defining such an annotation scheme as a prerequisite allowed us to collect high-quality summaries and evaluate editorial summarization approaches in a unified manner. Nevertheless, just as for other kinds of summarization, it is subjective to determine the "core" parts of a text [329]. This circumstance is prevalent in editorials, where the argumentative structure and even the thesis might not be explicitly stated, leaving room for interpretation. Therefore, both the data collection and evaluation must not rely on a single ground truth summary. Below, the operationalization of our annotation scheme is described.

Based on a corpus of news editorials that have previously been annotated with regard to argumentative discourse units, we crowdsource the generation of multiple reader-based summaries using Amazon's Mechanical Turk. The summarization is framed as an annotation / extraction task, where segments from an editorial are selected to compose a summary.

Data Source The news editorials corpus of Al-Khatib et al. [5] forms the data source of our study. It comprises 300 editorials from three different news portals: Al Jazeera, Fox News, and The Guardian. After reviewing them, we omitted 34 ones falsely labeled as editorial, and very short

ADU type	Example
Assumption	Many have simply lost faith in global climate negotiation summits such as COP 20 starting in Lima, Peru, today.
Anecdote	We were in-between lessons during our first class, when we suddenly heard the sound of shooting.
Common-Ground	Politicians are meant to act in the interests of their people.
Statistic	In the early 1900s, Argentina ranked among the world's top 10 in per capita income.
Testimony	"I saw my brother drown in front of my eyes," said Hamid.

Table 3.1: Examples of ADUs (evidence types) selected from different editorials.

ones. Each editorial has been segmented and annotated via crowdsourcing with the following argumentative discourse unit (ADU) types: anecdote, assumption, common ground, statistics, testimony, and other (collectively "evidence types", see Table 3.1). We adopt these ADU segments as our selection units for creating summaries (see Figure 3.1), since they are (mostly) well-formed texts by definition [5]. We choose this corpus because its annotations enable our detailed argumentation analysis of the acquired summaries as well as our evaluation of automatic summarization models. Although each editorial in the corpus is accompanied by a very short summary (one sentence), extracted automatically from its web page, these summaries are insufficient to study their argumentative nature.

Annotation Task The manual selection of summary segments often relies on the concept of importance or salience, i.e., the importance of each segment decides whether it is to be included in the summary [93, 125]. However, alongside capturing important content, the summary should also adhere to our annotation scheme. To operationalize this in the summary acquisition process, we specifically asked the workers to label each editorial segment as one of:

1. *Thesis.*: segments that represent what the author wants to persuade the reader of.
2. *Justification.*: segments that support the thesis.
3. *Background.*: segments that provide background information to the reader.

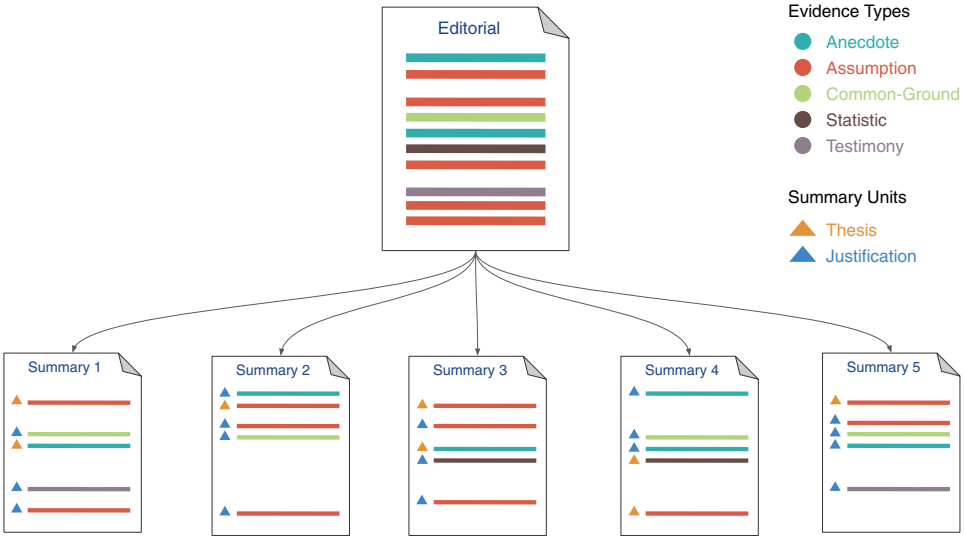


Figure 3.1: An editorial is comprised of multiple ADUs (evidence types), a subset of which are extracted to compose summaries. Each extracted ADU serves either as thesis or justification in the summary (summary units). Annotators can select up to two ADUs as the thesis.

4. *Not-in-summary.*: segments that should not be in the summary.

Summary length was limited to be 20% of an editorial's segments. In many editorials, the thesis may not be explicitly stated but rather implied by the author. For this reason, we allowed up to two segments to be labeled as thesis, which allows for inspecting the worker agreement on the editorial's core message.

We also asked each worker to self-assess (1) their prior knowledge of the editorial's topic (background), (2) if they agreed with the author's opinion (stance), (3) their general interest in the topic (interest), and (4) the persuasiveness of the editorial (persuasiveness). These questions constitute a profile of the worker, which allows for a profile-dependent analysis of the summaries. A similar profile was considered in evaluating spoken argumentation by Jovičić [151], where the effectiveness of the conveyed arguments depended on the audience. In our case, it turns out that the quality of our summaries is indirectly influenced by the workers (more details below).

Pilot Study We carried out a pilot study with 25 editorials, one editorial per human intelligence task (HIT), to check if our initial guideline required any revisions. Each editorial was annotated by five workers, resulting in 125 summaries. We did not show the segments' evidence types (from our

data source) to avoid any selection bias. However, we excluded the segments of an editorial annotated as "Other" as these segments are not argumentative. Although this somewhat affects the readability, most argumentative segments are well-formed, and our emphasis on the composition of useful and self-contained summaries in the guideline mitigates this problem to some extent.

From the 125 summaries, we obtained a total of 1180 summary segment labels: 19.66% thesis, 54.15% justification and 26.19% background. We found that 28% of unique summary segments were annotated interchangeably as justification or background. Thus, to simplify the final annotation, alongside the thesis label, we only used the justification label to annotate the segments that either support the thesis or provide background information.

Final Annotation Using the three labels thesis, justification, and not-in-summary (to undo previous selections), we acquired summaries for the remaining 241 editorials. Similar to the pilot study, each editorial was annotated by five workers, and to ensure quality, we chose workers with an approval rating of at least 98% and 1000 accepted HITs from three native English speaking countries (US, UK, and Canada). We chose these countries, in particular, to render the task more relevant to workers, since most editorials discuss topics related to these regions. This is further reflected in the self-assessment questionnaire, where 76.11% of the workers stated that they have sufficient background knowledge about the editorial topics they annotated.

Thesis Agreement In general, the agreement among summaries is expected to be low, not necessarily due to poor annotations, but due to the subjectivity of the importance notion [125, 190] and argumentative text perception. Besides, agreement tends to further decrease as the length of the summary increases [146]. However, the agreement on the thesis segment(s) indicates that the workers agree on the core message of the editorial. Hence, we consider worker agreement only on their selected thesis segments. As workers can label up to two segments as thesis, we consider full (two common segments) and partial (one common segment) agreement. As shown in Table 3.3a, the 61% majority agreement is promising, considering the challenging nature of the task, especially that a thesis can be indirectly implied when not explicitly stated in the editorial.

Our corpus consists of 1330 summaries having 12,806 labeled segments, with 14.7% labeled as thesis and 85.3% as justification. Table 3.3b shows the summary lengths in terms of segment and word counts.

Dimension	Explanation	%Maj.	% Summ.
Thesis-relevance	The thesis is relevant to the title, i.e., it could be the main point(s) of the editorial with the given title.	76%	81.7%
Persuasiveness	A persuasive summary aims to convince its readers to take a stand on a particular topic. To this end, it uses persuasion techniques such as: providing logical arguments to support its stand, invoking certain emotions on the readers, and/or using effective phrases.	76%	86.5%
Reasonableness	A reasonable summary adequately supports its thesis, i.e., the thesis is supported by a sufficient number of arguments.	79%	89.8%
Self-containedness	A self-contained summary is understandable by most of the readers, i.e., no need for additional information to get its thesis and follow its argumentation. Also, a self-contained summary refers to entities (people, locations, events, etc.) without any confusion in the usage of pronouns.	74%	84.1%
Overall score	–	74%	82.4%

Table 3.2: Summary quality dimensions and the guideline given to workers, derived from our annotation scheme to render it comprehensible to non-experts. Thesis-relevance is an indirect assessment of *thesis-indicativeness*. Third column (% Maj.) shows percentages of at least 2/3 agreement (majority) on each dimension, and for the overall score. The last column shows the percentage of summaries per editorial (averaged over all 266), which satisfy the corresponding quality dimension. On average, 82.4% of the summaries are high-quality per editorial, i.e., they satisfy at least three quality dimensions.

3.3 Evaluating and Ensuring Summary Quality

Adherence to the DUC Guideline Manual qualitative evaluation of summaries is often carried out according to the DUC guideline [71]: summaries must be grammatical, non-redundant, exert referential clarity, and have focus, as well as structure and coherence. Gillick and Liu [111] found this to be an expensive and a rather difficult task for non-experts. Thus, many summarization studies either avoid manual evaluation completely, or carry out only partial studies, rendering comparisons across papers difficult [125]. Even ground truth summaries themselves are rarely evaluated. To ensure the quality of our corpus, we therefore thoroughly evaluate each acquired summary for quality.

We argue that, by the construction of the summaries, their grammaticality and non-redundancy are sufficiently fulfilled, and that, by definition of our annotation scheme, the remaining DUC criteria are covered. Our summaries inherit the grammaticality of the editorials they were derived from. They are sequences of ADUs extracted from the editorials, and although ADUs may be part of longer sentences, they do form complete sentences in and off themselves [5]. Similarly, non-redundancy is inherited from the editorials; given their high writing quality, we can expect less redundant text, whereas if a certain point is repeated in an editorial to emphasize it, it stands to reason its summary may do so. Local redundancies, such as repeated names where an anaphora would suffice, cannot be avoided, since we did not ask the crowd workers to revise the summaries. Referential clarity, focus, structure, and coherence form part of our annotation scheme: Assessing the reasonableness of a summary includes checking for justifications to support its thesis, which is an indirect judgment of a summary's focus, i.e., the summary contains only related segments that together support its thesis. Likewise, assessing if a summary is self-contained considers referential clarity (i.e., no confusing usage of pronouns) as well as structure and coherence (i.e., the summary is well-organized).

Evaluation Task Similar to the acquisition of the summaries, we crowdsourced their qualitative evaluation. Table 3.2 shows how we explained the different quality dimensions of our annotation scheme to the crowd workers. The judgments for each dimension were made on a four-point scale (strongly disagree, weakly disagree, weakly agree, strongly agree). For persuasiveness, owing to the infeasibility of measuring this for some (often unknown) target audience [315], we restrict this dimension to being persuasive in general. Similarly, reasonableness of argumentations in theory also includes their acceptability by the target audience [315]. Again, due to the infeasibility of measuring this for our summaries, we restrict this dimension to having adequate justifications for their thesis. All (five) summaries of an editorial were evaluated in one HIT, each performed by three workers with the same selection criteria as in the summarization task. We only showed an editorial's title alongside each summary,[2] with its thesis emphasized in bold.

Pilot Study. To test and revise our guideline, we carried out another pilot study to evaluate the 25 editorials and their summaries from the sum-

[2]Reading titles only instead of the whole editorials significantly reduced the time taken for a HIT.

mary acquisition pilot study. Regarding thesis-indicativeness, for each summary's thesis, workers judged its relevance to the shown title, rather than reading the whole editorial. This design decision was backed by manually inspecting each title to ensure that it sufficiently indicates the issue discussed in the corresponding editorial. Acknowledging worker feedback, we included examples to help judge reasonableness and self-containedness, while only the description shown in Table 3.2 sufficed for judging persuasiveness.

For a sanity check, we exploit the fact that each of an editorial's five summaries is supposed to have the same or at least a similar thesis (Table 3.3a). Specifically, we asked workers to judge how similar in meaning is the thesis of a particular summary to that of the remaining summaries in a HIT. Then, given two summaries comprising similar thesis segments, we rejected the submissions of workers who judged them to be dissimilar.[3]

Annotator Agreement To compute annotator agreement, we first mapped all judgments to numeric scores (strongly-disagree: -2, weakly-disagree: -1, weakly-agree: 1, strongly-agree: 2). Then, we computed an overall score for a summary by averaging the numeric scores of all its quality dimensions. Thus, a summary with multiple quality dimensions gets a higher score. Table 3.2 shows the majority agreement for each quality dimension, as well as for the high-quality summaries. We note sufficient agreement on all quality criteria with the highest value for the reasonableness dimension.

In Table 3.3c, we also report significant correlations (at $p < 0.05$) between quality dimensions and overall score. We observe that: (1) A reasonable summary is also self-contained. By including justifications that build upon its thesis, a reasonable summary mitigates any distractions in its argumentation flow, thus rendering it understandable. (2) A reasonable summary is also persuasive. A key persuasion technique by authors is providing logical arguments to support their stances, i.e., reasonable summaries where a sufficient number of justifications is provided are more likely to be persuasive.

Quality Groups We distinguish summaries as being high or low quality based on the workers' assessments. We assert that a high-quality summary has at least three quality dimensions defined in our annotation scheme (Section 3.2) as judged by workers. As each summary is assessed by three workers, they may disagree on which dimensions it has. Thus, we use

[3]This sanity check was repeated multiple times to ensure reliable judgments.

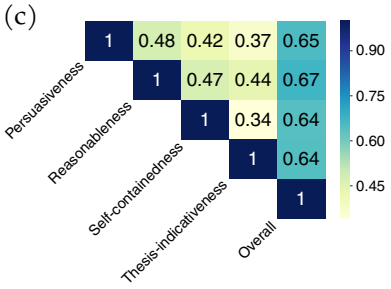
(a)

Workers	Editorials
2/5	96%
3/5	61%
4/5	25%

(b)

Length	Min	Mean	Max
Words	71	209.3	492
Segments	4	9.6	26

(c)



(d)

Position	Summary Segments		
	Thesis	Justif.	Combined
Lead	73.2%	20.2%	28.3%
Body	21.5%	67.6%	60.6%
Conclusion	5.2%	12.2%	11.1%

Table 3.3: (a) Agreement on thesis segment(s) among five workers. Values are computed on all 266 editorials (1330 theses). (b) Length statistics of all the summaries. (c) Correlation among judgments (Kendall's τ) for various quality dimensions including overall summary score. All values are significant at $p < 0.05$. (d) Percentages of summary segments (by function as thesis or justification and combined) extracted from lead, body and conclusion.

the majority vote to label a summary as "high-quality," i.e., if at least two workers agreed that it has at least three quality dimensions. We similarly distinguished the summaries per quality dimension (e.g., high/low-thesis-indicativeness). The distribution of high-quality summaries per editorial (on average, by quality dimension) is shown in Table 3.2. As for the quality dimensions, we observe that all dimensions are (almost) equally distributed in high-quality summaries, with reasonableness dimension favored slightly more (26.25%) (thesis-indicativeness: 24.08%, persuasiveness: 24.98%, self-containedness: 24.69%). Totally, we have 1096 high-quality and 234 low-quality summaries as per our manual evaluation.

In this section, we present a thorough analysis of our corpus, exploring (1) summary content overlap (i.e., the annotator agreement), (2) distribution of evidence types, (3) adherence of summaries to the editorial's structure, and (4) annotators' profiles and their impact on the quality of summaries.

ADU type	Editorials				HQ Summaries			LQ Summaries		
	L	B	C	Comb.	Thesis	Justif.	Comb.	Thesis	Justif.	Comb.
Assumption	61.7	66.9	79.2	68.0	70.4	64.2	65.1	66.5	66.3	66.3
Anecdote	25.1	18.9	8.5	18.2	18.8	18.9	18.9	23.3	20.7	21.1
Common-Grnd.	2.0	1.8	1.4	1.7	1.1	1.4	1.4	1.6	1.1	1.2
Statistics	2.9	3.1	1.6	2.9	1.8	4.4	4.0	1.6	2.5	2.4
Testimony	7.6	8.5	6.0	8.0	7.5	10.6	10.1	6.4	8.3	8.0

Table 3.4: Comparison of ADU distributions (in %) in editorials (by occurrence in **L**ead, **B**ody, or **C**onclusion and combined) as well as in groups of high-quality (**HQ**) and low-quality (**LQ**) summaries (by function as thesis or justification and combined).

Summary Content Overlap Here, we inspect the overlap among the five summaries per editorial. We first computed the Jaccard index[4] between each pair of summaries, in which the Jaccard index measures the intersection over the union between a pair's segments. Then, we averaged the indices over all the summary-pairs for an editorial (five summaries, ten pairs). We found that the average of Jaccard indices over all editorials is 0.2, which speaks for a low overlap between summaries. Although the workers agreed on the thesis (Table 3.3a), they still chose different justifications, leading to diverse summaries.

Distribution of Evidence Types We examine the distribution of evidence types in our summaries by obtaining the ADU labels from our data source. Table 3.4 shows this distribution in the three discourse parts (i.e., lead, body, and conclusion),[5] and in high-quality and low-quality summaries according to their role as thesis, justification, and combined.

The table reveals two key insights into the summaries' evidence types: (1) High-quality summaries have more statistics than the low-quality ones. Statistics is an evidence type stating or quoting the results or conclusions of quantitative research, studies, empirical data analyses, or similar [5]. (2) High-quality summaries have more testimony than the low-quality ones. Testimony is an evidence type that either states or quotes propositions made by some expert (person or organization) other than the author [5]. Although the overall percentage of statistics and testimony is less compared

[4]Jaccard index has an interval of [0,1]; values close to 1 indicate high overlap between two sets.

[5]After inspecting the length of *online lead paragraphs* from the NYT corpus [259], which are on average 12% of the article's length, we considered 15% of the top and bottom segments of an editorial as its lead and conclusion.

to other evidence types, such as assumption or anecdote, it is interesting to note that workers preferred more third-party evidence in their summaries. In conventional news summarization, these contents may been seen as extraneous details that need not be in a summary, whereas for editorials, they play a crucial role of supporting the thesis as justification.

Adherence to Editorial Structure We argue that constructing editorial summaries requires considering the specific contributions of its discourse parts to the argumentation. This means that unlike news reports where the summary is condensed primarily in the lead [327], important contents are distributed throughout the editorial. As shown in Table 3.3d, the majority of thesis segments are extracted from the lead (73.2%), but are insufficient to fully summarize the editorial, comprising only 28.3% of the combined summary segments. Furthermore, we note that each discourse part contributes proportionally to its summary (28.3%, 60.6% and 11.1% for lead, body, and conclusion).

Worker Profiles' Impact on Quality All workers employed in the acquisition task were assessed on a five-point scale regarding their background knowledge of the editorial's topic, if they agree or disagree with the editorial's stance, their interest in the topic, and if they find it persuasive.

To understand the impact of the workers' profiles on the quality of their summaries, we computed the correlation (Kendall's τ) between each aspect of their profile and the summaries' quality dimensions.[6] With a significant positive correlation ($p < 0.05$), we found that the workers who have more background knowledge of an editorial composed more persuasive summaries. On the other side, we did not find any significant correlation between the workers' stance toward a topic and the persuasiveness of a summary (as well as the overall quality) of their summaries.

3.4 Automatic Extractive Summarization of News Editorials

In this section, we investigate the capability of automatic summarization technology for generating high-quality summaries for editorials. Specifically, we implemented two unsupervised extractive summarization models (TextRank and ExtSum) and evaluated their output based on the Webis-EditorialSum-2020 corpus. These models emulate the manual summary acquisition setting, i.e., extracting segments within a given length budget. The

[6] We converted each judgment to a numerical score as in Section 3.3.

Model	Length (words)	Position			Thesis Cov.						Summary Cov.
		L	B	C	1	2	3	4	5	Maj.	
TextRank-Lex	79.9	13.6	**69.3**	**17.1**	67.3	23.1	7.7	1.9	0.0	9.6	11.9
TextRank-Entity	141.4	**40.6**	58.5	0.9	41.0	21.6	14.9	13.4	9.0	**37.3**	**20.1**
ExtSum-XLNet	151.0	23.4	60.5	16.2	34.4	29.7	23.6	8.5	3.8	35.8	16.4
ExtSum-DistilBERT	155.0	22.5	64.8	12.7	36.2	30.5	21.9	7.6	3.8	33.3	16.2
References	**209.3**	21.6	65.3	13.1							

Table 3.5: Average summary length in words. Average distribution of summary segments extracted by models from **L**ead, **B**ody and **C**onclusion in comparison to references. Percentage of (reference) theses and summary segments covered by models. For thesis coverage, we inspected if a thesis is completely included in the automatic summary (i.e., both segments). Accordingly, as each editorial has five theses, we also show coverage by number of theses completely captured in the model's summary. Summary coverage is the percentage of unique summary segments (from all five summaries of an editorial) captured.

input for each summarization model was the argumentative segments in an editorial (without any information about their evidence type). We set the same summary length (20%) for the automatic summaries as the ground truth ones.

3.4.1 Extractive Summarization Models

Our first summarization model is based on TextRank [201], an unsupervised summarization model based on PageRank [42]. Petasis and Karkaletsis [229] demonstrated that TextRank is able to identify argument components in a text. By comparing the connections among sentences with those between claims and premises, they established TextRank as a suitable model for argument mining. Accordingly, we leverage this to create extractive summaries of the editorials. TextRank first constructs an undirected graph of the entire editorial with the segments as nodes. For weighing the connecting edges, we investigated two similarity functions resulting in two variants of TextRank: `TextRank-Lex` which uses lexical overlap among segments and `TextRank-Entity` which uses the number of common named entities[7] between two segments.

As our second summarization model, we adopt an extractive summarization model based on BERT [83] that clusters (using K-Means) the contextual embeddings of an editorial's segments and selects those that are closer to its centroid as the final summary [202]. To encode the editorial segments, we

[7] We used Spacy's *en-core-web-md* model for tagging named entities.

chose contextual embeddings from two distinct architectures: `ExtSum-XLNet` based on XLNet [333], an autoregressive language model that outperforms BERT on several tasks, and `ExtSum-DistilBERT` based on DistilBERT [260]. DistilBERT is an efficient language model that leverages knowledge distillation to achieve similar performance as BERT but with significantly fewer resources and increased speed compared to XLNet.

3.4.2 Model Evaluation

We compare each model's summary for an editorial with its multiple references in terms of its adherence to the editorial's structure and coverage of unique summary (and theses) segments.

Regarding the structure of the automatic summaries, the distribution of segments from lead, body, and conclusion is shown in Table 3.5. We see that `TextRank-Lex` extracts more segments from the body and the conclusion. However, it extracts much shorter segments than those in the references. In contrast, `TextRank-Entity` extracts more segments from the lead of an editorial and produces longer summaries. This is because the actors of an editorial (named entities) are usually introduced in the beginning. Both the ExtSum variants have almost a similar distribution of extracting segments from the editorial's discourse parts; besides that, embeddings from the smaller DistilBERT produce relatively longer summaries. Still, all the automatically produced summaries are shorter than the references in terms of word count.

The coverage of the theses and summary segments of the references by the automatic summaries is shown in Table 3.5. We observe that `TextRank-Entity` has the highest coverage of the reference summary segments. Despite producing shorter summaries than the ExtSum models, it also consistently captures a majority of theses. This reveals a plausible segment extraction strategy followed by workers in the summary acquisition task, where argumentative segments connecting different actors are often selected. Among the ExtSum models, `ExtSum-DistilBERT` has a similar distribution of segments from the discourse parts as the references, with `ExtSum-XLNet` having a slightly higher coverage of the unique summary segments from the references.

3.5 Summary

This chapter presented the first steps towards summarizing news editorials, a type of long form argumentative text, by defining and operationalizing

high-quality summaries. We introduced an annotation scheme tailored to editorial summaries, which we employed to acquire and evaluate the Webis-EditorialSum-2020 corpus; the first corpus for news editorial summarization containing five summaries per editorial (1330 summaries in total). Our annotation scheme defines multiple quality dimensions grounded in argumentation quality studies. Through detailed corpus analyses, we found that editorial summaries have a distinct structure compared to those of news reports; that third party evidence in summaries improves their overall quality; that background knowledge of workers is positively correlated to the persuasiveness of their summaries; and, that some automatic models can at least capture an editorial's thesis. We consider our corpus a useful resource for promoting research in automatic summarization and computational argumentation. For instance,it can be used to learn automatically classifying evidence types in other long-form argumentative texts such as debates, social media posts, and student essays.

4

Mining Social Media for Author-provided Summaries

Although it is crucial to define the characteristics of reference summaries as done in the previous chapter, it is often impractical to gather such summaries on a large scale through crowdsourcing. A practical alternative is to seek human indications of summary-worthy content in a given text. This can help in automatically identifying pairs of documents and summaries, thereby facilitating the creation of appropriate training datasets. Accordingly, this chapter introduces an extensive dataset of social media posts complemented by author-provided, highly abstractive summaries. To create this dataset, we leveraged the common practice of social media users summarizing their own posts as a courtesy to their readers. We collected a substantial number of such author-provided summaries from the popular social news aggregation and discussion website Reddit. The chapter outlines the data collection and preprocessing procedures, providing a comprehensive analysis of the resulting dataset. The uniqueness of this dataset lies in its inclusion of author-provided, abstractive summaries, making it a pivotal resource for investigating the widely adopted paradigm of reinforcement learning from human feedback (RLHF) [281], which currently plays a dominant role in the operationalization of large language models.

Furthermore, the chapter presents the results from the first shared task on abstractive summarization, conducted using this dataset. It includes a human-centered qualitative evaluation and error analysis of the candidate summaries produced by the participants. These insights offer valuable perspectives on the performance and capabilities of various summarization ap-

Corpus	Genre	Training pairs
English Gigaword	News articles	4 million
CNN/Daily Mail	News articles	300,000
DUC 2003	Newswire	624
DUC 2004	Newswire	500
Webis-TLDR-17	**Social Media**	**4 million**

Table 4.1: Top rows: commonly used English-language corpora; bottom row: our contribution.

proaches, shedding light on the strengths and limitations of current techniques.

4.1 Leveraging Human Signals for Summary Identification

Given a document, automatic summarization is the task of generating a coherent shorter version of the document that conveys its main points. Depending on the use case, the target length of a summary may be chosen relative to that of the input document, or it may be limited. Either way, a summary must be considered "accurate" by a human judge in relation to its length: the shorter a summary has to be, the more it will have to abstract over the input text. Automatic *abstractive* summarization can be considered one of the most challenging variants of automatic summarization [104]. But with recent advancements in the field of deep learning, new ground was broken using various kinds of neural network models [64, 139, 257, 270].

The performance of these kinds of summarization models strongly depends on large amounts of suitable training data. To the best of our knowledge, the top rows of Table 4.1 list all English-language corpora that have been applied to training and evaluating single-document summarization networks in the past two to three years; only the two largest corpora are of sufficient size to serve as training sets by themselves. At the same time, all of these corpora cover more or less the same text genre, namely news. This is probably due to the relative ease by which news articles can be obtained as well as the fact that the news tend to contain properly written texts, usually from professional journalists. Notwithstanding the usefulness of existing corpora, we argue that the apparent lack of genre diversity currently poses an obstacle to deep learning-based summarization.

In this regard, we identified a novel, large-scale source of suitable training data from the genre of social media. We benefit from the common practice of social media users summarizing their own posts as a courtesy to their

readers: the abbreviation TL;DR, originally used as a response meaning "too long; didn't read" to call out on unnecessarily long posts, has been adopted by many social media users writing long posts in anticipatory obedience and now typically indicates that a summary of the entire post follows. This provides us with a text and its summary—both written by the same person—which, when harvested at scale, is an excellent datum for developing and evaluating an automatic summarization system. In contrast to the state-of-the-art corpora, social media texts are written informally and discuss everyday topics, albeit mostly unstructured and oftentimes poorly written, offering new challenges to the community. Thus, we endeavored to extract a usable dataset specifically suited for abstractive summarization from Reddit, the largest discussion forum on the web, where TL;DR summaries are extensively used. In what follows, we discuss in detail how the data was obtained and preprocessed to compile the Webis-TLDR-17 corpus.

4.2 Constructing a Corpus of Abstractive Summaries

Reddit is a community centered around social news aggregation, web content rating, and discussion, and, as of mid-2017, one of the ten most-visited sites on the web according to Alexa.[1] Community members submit and curate content consisting of text posts or web links, segregated into channels called *subreddits*, covering general topics such as Technology, Gaming, Finance, Well-being, as well as special-interest subjects that may only be relevant to a handful of users. At the time of writing, there are about 1.1 million subreddits. In each subreddit, users submit top-level posts—referred to as submissions—and others reply with comments, reflecting, contradicting, or supporting the submission. Submissions consist of a title and either a web link, or a user-supplied body text; in the latter case, the submission is also called a *self-post*. Comments always have a body text—unless subsequently deleted by the author or a moderator—which may also include inline URLs.

Large crawls of Reddit comments and submissions have recently been made available to the NLP community.[2] For the purpose of constructing our summarization corpus, we employ the set of 286 million submissions and 1.6 billion comments posted to Reddit between 2006 and 2016.

Given the raw data of Reddit submissions and comments, our goal is to mine for TL;DR content-summary pairs. We set up a five-step pipeline of consecutive filtering steps; Table 4.2 shows the number of posts remaining after each step.

[1] `http://www.alexa.com/siteinfo/reddit.com`
[2] `http://files.pushshift.io/reddit/`

Filtering Step	Subreddits	Submissions	Comments
Raw Input	617,812	286,168,475	1,659,361,605
Contains tl.{0,3}dr	37,090	2,081,363	3,755,345
Contains tl;dr[3]	34,380	2,002,684	3,412,371
Non-bot post	34,349	1,894,094	3,379,287
Final Pairs	32,778	1,667,129	2,377,372

Table 4.2: Filtering steps to get the TL;DR corpus.

An initial investigation showed that the spelling of TL;DR is not uniform, but many plausible variants exist. To boil down the raw dataset to an upper bound of submissions and comments (collectively posts) that are candidates for our corpus, we first filtered all posts that contain the two letter sequences 'tl' and 'dr' in that order, case-insensitive, allowing for up to three random letters in-between. This included a lot of instances found within URLs, which were thus ignored by default. Next, we manually reviewed a number of example posts for all of the 100 most-frequent spelling variants (covering 90% of the distribution) and found 33 variants to be highly specific to actual TL;DR summaries,[3] whereas the remaining, less frequent, variants contained too much noise to be of use.

The Reddit community has developed many bots for purposes such as content moderation, advertisement or entertainment. Posts by these bots are often well formatted but redundant and irrelevant to the topic at hand. To ensure we collect only posts made by human users—critically, some Reddit users operate TL;DR-bots that produce automatic summaries, which may introduce undesirable noise—we filter out all bot accounts with the help of an extensive list provided by the Reddit community,[4] as well as manual inspection of cases where the user name contained the substring "bot."

For the remaining posts, we attempt to split their bodies at the expression TL;DR to form the content-summary pairs for our corpus. We locate the position of the TL;DR pattern in each post, and split the text into two parts at this point, the part before being considered as the content, and the part following as the summary. In this step, we apply a small set of rules to remove erroneous cases: multiple occurrences of TL;DRs are disallowed for their ambiguity, the length of a TL;DR must be shorter than that of the content, there must be at least 2 words in the content and 1 word in TL;DR. The last rule is very lenient; any other threshold would be artificial (i.e., a

[3] tl dr, tl;dr, tldr, tl:dr, tl/dr, tl; dr, tl,dr, tl, dr, tl-dr, tl'dr, tl: dr, tl.dr, tl ; dr, tl_dr, tldr;dr, tl ;dr, tl\dr, tl/ dr, tld:dr, tl;;dr, tltl;dr, tl~dr, tl / dr, tl :dr, tl - dr, tl\\dr, tl. dr, tl:;dr, tl|dr, tl;sdr, tll;dr, tl : dr, tld;dr

[4] https://www.reddit.com/r/autowikibot/wiki/redditbots

Example Submission
Title: Ultimate travel kit **Body:** Doing some traveling this year and I am looking to build the ultimate travel kit ... So far I have a Bonavita 0.5L travel kettle and AeroPress. Looking for a grinder that would maybe fit into the AeroPress. This way I can stack them in each other and have a compact travel kit. **TL;DR:** What grinder would you recommend that fits in AeroPress?
Example Comment (to a different submission)
Body: Oh man this brings back memories. When I was little, around five, we were putting in a new shower system in the bathroom and had to open up the wall. The plumber opened up the wall first, then put in the shower system, and then left it there while he took a lunch break. After his break he patched up the wall and left, having completed the job. Then we couldn't find our cat. But we heard the cat. Before long we realized it was stuck in the wall, and could not get out. We called up the plumber again and he came back the next day and opened the wall. Out came our black cat, Socrates, covered in dust and filth. **TL;DR:** plumber opens wall, cat climbs in, plumber closes wall, fucking meows everywhere until plumber returns the next day

Table 4.3: Examples of content-summary pairs.

10 word sentence may still be summarizable in 2 words). However, future users of our corpus probably might have more conservative thresholds in mind. We hence provide a subset with a 100 word content threshold.

Reddit allows Markdown syntax in post texts, and many users take advantage of this facility. As this introduces some special characters in the text, we disregard all Markdown formatting, as well as inline URLs, when searching for TL;DRs.

After filtering, we are left with approximately 1.6 million submissions and 2.4 million comments for a total of 4 million content-summary pairs. Table 4.3 shows one example each of content-summary pairs in submissions and comments. Table 4.9 shows the comparison of the abstractive summary from our corpus against an extractive summary from the CNN-DailyMail corpus. The development of the filtering pipeline went along with many spot-checks to ensure selection precision. As a final corpus validation, we reviewed 1000 randomly selected pairs and found 95% to be correct, a proportion that allows for realistic usage. Nevertheless, we continue on refining the filtering pipeline as systematic errors become apparent.

4.2.1 Corpus Statistics

For the 4 million content-summary pairs, Table 4.4 shows distributions of the word counts of content and summary, as well as the ratio of summary to content word count. On average, the content body of submissions tends to be nearly twice as long as that of comments, whereas the fraction of the total word count in the summary tends to be higher for submissions (about 11% being typical) than for comments (8%). As the length of a post increases,

	Min	Median	Max	Mean	σ
Comments					
Total	3	164	6,880	225.21	210.22
Content	2	144	6,597	202.99	199.19
Summary	1	15	1,816	22.21	27.81
Summ. / Cont.	0.00	0.11	1.00	0.16	0.16
Submissions					
Total	3	296	9,973	416.40	384.72
Content	2	269	9,952	382.75	366.99
Summary	1	22	3,526	33.65	47.87
Summ. / Cont.	0.00	0.08	1.00	0.12	0.13

Table 4.4: Length statistics for the TL;DR corpus.

the length of the summary tends to increase as well (Pearson correlations of 0.40 for submissions and 0.35 for comments), while the ratio of summary to content word count increases only slightly (correlations of 0.11 and 0.07).

4.2.2 Corpus Verticals

The corpus allows for constructing verticals with regard to content type, content topic, and summary type. Content type refers to submissions vs. comments, the key difference being that submissions include an author-provided title field, which can serve as an additional source of summary ground truth. Comments may perhaps inherit the title of the submission they were posted to, but topic drift may occur. The submission of the example comment in Table 4.3 was befittingly entitled "So I found my cat after 6 hours with some power tools...", referring to a picture of a cat stuck in a wall.

Content topic refers to the subreddit a submission or comment was posted to. While subreddits cover trending topics as well as online culture very well, thus ensuring a broader range of topics than news can deliver, there is currently no ontology grouping them for ease of selection.

In our data exploration, we observed that Reddit users write TL;DRs with various intentions, such as providing a "true" summary, asking questions or for help, or forming judgments and conclusions. Although the first kind of TL;DR posts are most important for training summarization models, yet, the latter allow for various alternative summarization-related tasks. Hence, we exemplify how the corpus may be *heuristically* split according to summary type—other summary type verticals are envisioned.

To estimate the number of true summaries, we extract noun phrases from both content and summary, and retain posts where they intersect. Only

966,430 content-summary pairs—580,391 from submissions and 386,039 from comments—pass this test, but this is a lower bound: since abstractive summaries may well be semantically relevant to a post without sharing any noun phrases.

To extract question summaries, we test for the presence of one of 21 English question words,[5] as well as a question mark, in the summary. We can isolate a subset of 78,710 content-summary pairs this way (see Table 4.3 top), which allow for training tailored models yielding questions for a summary.

Many posts contain abusive words in the content, the TL;DR, or both (see Table 4.3 bottom). While retaining vulgarity in a summary may be appropriate, it seems rarely desirable if a model introduces vulgarity of its own. To separate 299,145 vulgar summaries, we use a list of more than 500 English offensive words from Google's now defunct "What Do You Love" project.[6] Come to think of it, these may still be used to train a swearing summarizer, if only for comedic effect.

4.3 Insights from the TL;DR Challenge

Based on our aforementioned corpus, we organized the TL;DR Challenge [284], the first shared task for abstractive summarization of social media posts. This section presents key details about the system submissions and our extensive evaluation of the summary quality.

Out of 16 registered participants, we received 5 submissions from 3 participants (2 from industry). In addition, we provided a `seq2seq-baseline` model with 2 layers, bi-LSTM, 256 hidden units and no attention. Participants trained models at their own premises and deployed them to a virtual machine on TIRA. Via TIRA's web interface, scripts were configured to generate summaries for a hidden test set and then remotely executed. Multiple runs were allowed for each participant.[7] Each run was fed to an automatic evaluator script to compute ROUGE scores. Each software and evaluator run on the test set was manually reviewed by organizers for errors and data leakage. After a successful review, the scores were shared on a

[5] Extension of the word list at `https://en.wikipedia.org/wiki/Interrogative_word` with "can", "should", "would", "is", "could", "does", "will" after manual analysis of the corpus.

[6] Obtained via `https://gist.github.com/jamiew/1112488`

[7] Evaluating models on TIRA using ROUGE was allowed even after the submission deadline. Thus, a participant's technical paper may have a variation of the same model with different ROUGE scores, but was not manually evaluated.

public leaderboard.[8] Two participants provided their system descriptions. We did not receive any description for the `tldr-bottom-up` model.

Gehrmann et al. [110] leveraged fine-tuned language models to generate abstractive summaries. They argue that excessive copying facilitated by the copy-attention mechanism hinders paraphrasing and information compression (abstraction). As part of the TL;DR challenge, they compared two summarization approaches (`pseudo-self-attn` and `transf-seq2seq`) demonstrating the effectiveness of transfer learning at generating abstractive summaries. Our manual evaluation confirms that these models generate concise and coherent summaries.

Tackling the same problem of excessive copying in pointer-generator models, Choi et al. [62] proposed using Variational Autoencoder (VAE) in combination with an extractive summarization model. The `unified-pgn` model uses a BERT-based extractive model that is fine-tuned to select important sentences, which are then summarized using a pointer-generator network. In order to introduce diversity, the `unified-vae-pgn` model uses a VAE for generating summaries of the extracted important sentences. This multi-stage architecture preserves a substantial amount of key information while generating acceptable summaries as revealed in our manual evaluation. We refer readers to the system description papers for further details.

4.3.1 Automatic Evaluation

We begin with a novelty analysis as per See et al. [270], calculating the fraction of n-grams in the summary that are absent from the text as its novelty (Table 4.5). The ground truth has the highest novelty, underlining the abstractive nature of author-provided summaries. Next, we used ROUGE [176] for automatic evaluation and report the F1-scores.[9] From Table 4.5 it is difficult to draw any conclusions just by looking at ROUGE scores. Furthermore, a key issue of ROUGE is that it does not provide any upper bounds for the quality of a summarization system [266], thus warranting an extensive manual evaluation of the systems.

4.3.2 Manual Evaluation

Using Amazon Mechanical Turk, we crowdsourced our manual evaluation within two tasks: preference scoring and quality scoring. One hundred

[8] `https://www.tira.io/task-overview/tldr-generation/inlg-19-tldr-generation-test-dataset-2018-11-05`

[9] `https://github.com/pltrdy/rouge`; we intentionally rounded off the scores in our evaluation script in order to show differences of at least one point on the ROUGE metric.

Model	ROUGE			Novelty (n-grams)				Len.
	1	2	L	1	2	3	4	
unified-pgn	19	4	**15**	0.80	5.15	8.67	11.42	33.5
unified-vae-pgn	19	4	**15**	0.86	5.04	8.90	11.92	32.8
transf-seq2seq	19	**5**	14	0.82	4.28	6.44	7.54	14.5
pseudo-self-attn	18	4	13	1.49	7.21	9.54	9.98	12.1
tldr-bottom-up	**20**	4	**15**	1.90	5.29	8.32	10.73	37.3
seq2seq-baseline	3	0	2	0.00	2.27	2.47	2.05	4.9
ground truth	–	–	–	**9.48**	**21.94**	**24.86**	**25.20**	26.1

Table 4.5: ROUGE-1, 2, and L scores and novelty analysis for 1 to 4-grams of the generated summaries along with their average lengths in words.

randomly selected examples from the test set were scored in both tasks, where each HIT (Human Intelligence Task) was assigned to 3 workers. We employed master workers with a minimum approval rate of 95% and at least 10,000 approved HITs.[10]

Preference scoring. The DUC guidelines for manually evaluating summaries by Dang [71] were designed for experts. Gillick and Liu [111] reported that Mechanical Turk workers were unable to provide expert-like scores and had strong disagreements. Therefore, we kept the task as simple as possible: "Given a text and its summaries from all models (and the ground truth), score each summary for how well it summarizes the given text." We employed a four-point Likert scale ((1) very bad, (2) bad, (3) good, and (4) very good), since [33] showed that presenting a middle alternative causes many people to choose it to escape uncertainty. Moreover, we asked for a written justification for each score. The scores collected reflect the summaries' overall quality, combining all aspects of summary quality relative to all other summaries, as perceived by the workers. Note that the summaries were shown in random order to prevent order effects. The score justifications required the workers to reflect about their judgments, and at the same time, they provide for an error analysis (see Table 4.3.3 for details). Moreover, the justifications allowed for double-checking whether workers actually read the summaries while scoring. Figure 4.1 shows which pairs of systems have significant differences along with effect sizes.[11]

[10] We paid $0.80 per HIT for preference scoring and $0.20 for quality scoring at an average hourly rate of $8 and $825 total.

[11] We use Mann-Whitney U for pairwise comparison using Bonferroni correction.

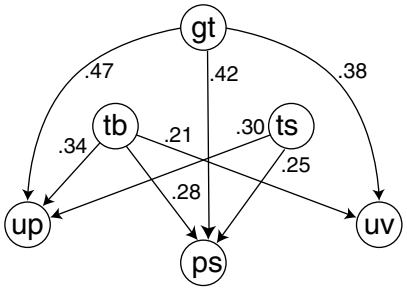


Figure 4.1: Summary of the preference scoring task: directed edges denote significantly higher scores ($p < 0.001$), and are annotated with effect sizes. The baseline model is much worse in comparison and hence not included. Key: **gt**: ground truth, **tb**: tldr-bottom-up, **ps**: pseudo-self-attn, **up**: unified-pgn, **ts**: transf-seq2seq, **uv**: unified-vae-pgn.

Model	Sufficiency				Text quality			
	1	**2**	**3**	**Avg.**	**1**	**2**	**3**	**Avg.**
unified-pgn	2	38	60	2.11	6	68	26	1.78
unified-vae-pgn	4	30	66	2.13	9	62	29	1.78
transf-seq2seq	4	27	69	2.20	0	5	95	**2.70**
pseudo-self-attn	12	35	53	1.97	2	8	90	**2.67**
tldr-bottom-up	2	25	73	2.30	1	28	71	2.29
seq2seq-baseline	79	14	7	1.11	73	21	6	1.11
ground truth	2	8	90	**2.52**	0	15	85	2.57

Table 4.6: Sufficiency and text quality score distribution in the majority category.

Quality Scoring. Our second evaluation task was to independently assess a model's summaries across two specific qualitative dimensions. We adopt the term *sufficiency* to group multiple properties of a summary, such as informativeness, relevance, and focus. Similarly, *text quality* groups properties independent of the content, such as structure, coherence, grammar, and readability. In contrast to the first task, this gives workers specific goals and helps us to better differentiate between the models. Furthermore, it may help to identify if non-expert annotators can still produce reliable judgments without a guideline. Gillick and Liu [111] cautioned that workers have difficulties distinguishing the content of a summary from its text quality. With that in mind, we devised two orthogonal three-level rating scales. With respect to *sufficiency*, workers could rate a summary as *insufficient* (incomplete and unrelated to the source text), as *barely acceptable* (missing the main point, but capturing relevant secondary information), or as *sufficient* (capturing the main point of the text). In terms of *text quality*, we distin-

Sufficiency	
Missing context (*MC*)	The summary does not provide any context, misses primary information or captures only secondary information.
Wrong sentiment (*WS*)	The overall sentiment of the post is either flipped or neutralized due to wrong negations.
Factually incorrect (*FI*)	Entities, such as names, locations, dates are wrongly reproduced, making the summary factually incorrect.
Overly simplistic (*OS*)	Summary lacks reasoning and necessary details making it too generic.
Text quality	
Bad grammar (*BG*)	A bad summary contains incorrect punctuations, wrong connectives, or formatting errors.
Incoherence (*IC*)	Improper flow of text which renders the summary meaningless.
Repetition (*RP*)	Excessive repetition of tokens.
Bad continuity (*BC*)	Summary starts off well but later culminates to gibberish text.

Table 4.7: Categories of worker criticism; the score of a summary was in many cases influenced by a combination of these aspects.

guished the levels *badly written* (incoherent or major errors), *needs improvement* (minor errors breaking the flow, but understandable), and *well written* (no errors, coherent, and understandable).

Table 4.6 shows the score distribution for both dimensions in the majority category. For text quality, multiple models perform well compared to `ground truth`. Models with longer summaries (see Table 4.5), require further improvement in terms of text quality despite having a similar number of sufficient summaries. To compute significance, we assign the score of a summary to be the average of sufficiency and quality score. Figure 4.2 shows which pairs of systems had significant differences in scores along with effect sizes.

Model	Sufficiency				Text quality				Pos.
	MC	WS	FI	OS	BG	IC	RP	BC	
unified-pgn	94	12	11	6	40	81	22	10	56
unified-vae-pgn	52	6	21	9	39	61	12	8	100
transf-seq2seq	102	5	15	23	2	23	1	–	128
pseudo-self-attn	106	15	38	29	1	28	4	–	83
tldr-bottom-up	61	1	25	6	20	43	1	7	137
seq2seq-baseline	–	–	–	–	–	221	68	–	0
ground truth	69	1	11	14	10	12	0	3	178

Table 4.8: Distribution of summary aspects obtained from error analysis. The last column (positive) is the number of judgments (out of 300) where workers found no major problems with the summary .

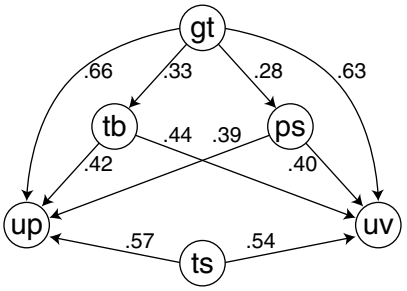


Figure 4.2: Summary of the quality scoring task: directed edges denote significantly higher scores ($p < 0.001$), and are annotated with effect sizes. The baseline model is much worse in comparison and hence not included. Key: **gt**: ground truth, **tb**: tldr-bottom-up, **ps**: pseudo-self-attn, **up**: unified-pgn, **ts**: transf-seq2seq, **uv**: unified-vae-pgn.

4.3.3 Error Analysis: Score Justifications

We manually reviewed all 2100 justifications given during the preference scoring task, and identified the summary aspects that most frequently influenced the scores. We further categorize these reasons under the two dimensions of sufficiency and text quality as shown in Table 4.7. These justifications may help the participants in improving their systems, and also aid the development of new models and evaluation methodologies. Moreover, comparing the ordering of systems in Figure Figure 4.1 and Figure 4.2, we see that master workers could differentiate the systems reasonably well without a guideline in the preference scoring task.

Table 4.8 shows the distribution of summary aspects for each model. *Missing context* (*MC*) was a key concern across all models where summaries failed to either capture enough details, or provide a proper reasoning, rendering them as *partial summaries* instead. This was prominent in

the `transf-seq2seq` and `pseudo-self-attn` models, which produce shorter summaries that either lack relevant details or are overly simplistic (*OS*). However, these models generate the most coherent and readable summaries with very few cases of incoherence (*IC*) and no repetition (*RP*), obtaining an overall positive feedback. In contrast, the `tldr-bottom-up` and `unified-vae-pgn` models with much longer summaries preserved more information, but with issues in grammar (*BG*) or continuity (*BC*), leading to higher numbers of incoherent summaries.

4.3.4 Results

Both `transf-seq2seq` and `pseudo-self-attn` generated the highest-quality text, but especially the latter often lacked information; `tldr-bottomup` generated the most informative summaries (with acceptable text quality), followed by `transf-seq2seq`. We found that, in the absence of a guideline, master workers provided reliable judgments by identifying influential summary aspects as seen in Table 4.7. All models struggled with capturing sufficient context spread throughout the posts, further aggravated by the casual writing style. Nevertheless, we observed encouraging results in terms of text quality. We envision that summarization will benefit from including formalisms of importance, argumentation, and reasoning into the models, while striking a balance between summary length and text quality.

4.4 Summary

In this chapter, we demonstrated how user-generated content can serve as a source of large-scale summarization training data, and mined a set of 4 million content-summary pairs from Reddit, which we make available to the research community as the Webis-TLDR-17 corpus.[12] Our filtering pipeline, data exploration, and vertical formation allow for fine-grained control of the data, and can be tailored to one's own needs. Other data sources should be amenable to mining TL;DRs, too: a cursory examination of the Common-Crawl and Clueweb12 web crawls unearths more than 2 million pages containing the pattern—though extracting clean content-summary pairs will likely require more effort for general web content than for self-contained social media posts.

Finally, we presented key findings from the TL;DR Challenge, a large-scale human evaluation of summarization models on Reddit data. We found that the models struggle with capturing sufficient context, but pro-

[12] `https://webis.de/data/webis-tldr-17`

duce summaries of acceptable text quality. In particular, our human-centered error analysis revealed that master workers can provide reliable judgments without a guideline, and that the preference scoring task is more difficult than the quality scoring task for assessing summary quality.

Example - CNN/DailyMail Corpus

Article

NASA will launch Space Shuttle Endeavour on February 7, which will be the first of five launches this year before the shuttle fleet is retired. Endeavour will blast off from the Kennedy Space Center in Florida on a 13-day mission to the international space station. The mission will include three spacewalks, NASA said. The shuttle will also deliver the final U.S. portion of the space station. This portion will provide more room for crew members. NASA plans to retire its space shuttles Discovery, Endeavour and Atlantis later this year. The space agency has been looking for places, such as museums, to house the shuttles after they are retired. Space Shuttle Discovery will be transferred to the Smithsonian National Air and Space Museum in Washington. The privilege of showing off a shuttle won't be cheap – about $29 million, NASA said.

Highlights

- This will be first of five launches this year before the shuttle fleet is retired
- NASA is scheduled to launch Space Shuttle Endeavour on February 7.
- Shuttle will deliver final U.S. portion of the international space station
- NASA has been looking for places to house the shuttles once they are retired

Example- Webis-TLDR-17 Corpus

Post

I'm so upset at myself. My boyfriend surprised me with an amazing, fancy dinner for our one year anniversary yesterday. I already wasn't feeling well when he told me we were going to dinner but when I saw what he planned I didn't have the heart to tell him I wasn't that hungry. In the end I pushed myself to eat the fixed menu he ordered for us and the bill was over 500, I couldn't handle it and after dessert I ended up going to the bathroom and throwing it all up.

I can't believe I wasted so much of his money and am so disappointed in myself for not speaking up and simply saying I didn't feel well. I feel like I've wasted the effort he put into planning this. I also feel like I missed out on some amazing food that we would usually never splurge for. He doesn't know I threw it up and I just told him I loved it because regardless of how I felt health wise I loved that he put in so much effort to make sure I felt special. But I can't stop stewing in my own feelings. Help.

TL;DR

my boyfriend is amazing and bought us an expensive anniversary dinner. Threw it all up, he doesn't know. Feel horrible guilt and FOMO

Table 4.9: Comparison of summary styles from the CNN/DailyMail and the Webis-TLDR-17 corpus. Emphasized text shows the extractive nature of the summary (highlights) for the news domain. The highlights are concatenated and used as the target summary for training summarization models. In contrast, the example from the Webis-TLDR-17 corpus exhibits higher abstraction, abbreviations and composition of multiple facts into single phrases.

5

Generating Conclusions for Argumentative Texts

This chapter extends the idea of identifying summary-worthy content via human signals to the domain of argumentative texts in online discussions. Compared to the well-structured news editorials discussed in Chapter 3, online persuasive discussions, such as those found on Reddit's ChangeMyView[1] or debate portals like idebate.org, present a more informal type of argumentative text. These discussions typically commence with the author succinctly expressing their viewpoint in a short statement, accompanied by a more extensive explanation or background justifying their perspective. Other users then respond to the original post, either agreeing or disagreeing with the initial viewpoint by presenting their own arguments. The brief statement can be interpreted as the *conclusion* of the corresponding argument, as it encapsulates the target of the author's stance.

In this chapter, we utilize these reasoning texts and conclusion pairs as a source of training data for the new task of automatic conclusion generation for argumentative texts. We construct a dataset and then focus on generating informative conclusions that aim to balance the trade-off between informativeness and abstractiveness. To achieve this, we incorporate external argumentative knowledge into the model, such as the discussion topic, various aspects from the provided reasoning, and the conclusion targets. We evaluate the approaches through a detailed human-centered error analysis to comprehend the strengths and weaknesses of the proposed models.

[1] `https://old.reddit.com/r/changemyview/`

5.1 Identifying Author-provided Conclusions

A conclusion of an argument is a statement that conveys a stance towards a specific target [10, 21]. Drawing conclusions is an integral part of argumentation, but often various conclusions may be drawn from a set of premises. Consider the following argumentative text on caffeine adapted from the web:[2]

"Caffeine stimulates the nervous system, signaling fat cells to break down body fat. It also increases epinephrine (adrenaline) levels, a fight-or-flight hormone preparing the body for physical exertion. With free body fat acids as fuel, on average, 12% higher performance is attainable."

Consider further these alternative conclusions:

1. *Caffeine is good.*
2. *Caffeine improves physical performance.*

The first conclusion conveys a pro stance towards the target, caffeine. The second, conveys a pro stance towards caffeine, too, but it also emphasizes a specific concept ("physical performance"). The former conclusion is generic, only *indicating* the stance, while the latter is *informative*; a distinction also made in text summarization (Section 5.1.1).[3]

Argumentative texts include short arguments, such as forum posts and reviews, as well as long-form texts, such as essays, blogs, and editorials. Most of these typically have an intended conclusion of which the authors seek to persuade their readers.[4] While the conclusion may be already implied in a given text, authors often choose not to explicitly provide one, either for rhetorical reasons [4, 120], or to encourage critical thinking [195]. However, when browsing many argumentative texts (e.g., via a search engine or on a social media timeline), having an explicit conclusion helps human readers (and by extension also machines) to quickly process the texts.

In this paper, we introduce the task of generating informative conclusions for argumentative texts, and take the first steps with four key contributions: (1) Adaptation of the notion of informativeness from text summarization as a desired property of a conclusion besides stating a target and the stance towards it. (2) Compilation of Webis-ConcluGen-21, a corpus

[2] `https://www.healthline.com/nutrition/top-13-evidence-based-health-benefits-of-coffee`

[3] Other works on argumentation use the term *specificity* to express a similar idea [87, 157].

[4] An exception is an argumentative text dedicated to deliberation, which merely surveys the argument landscape on a given topic without trying to influence the reader's opinion.

of 136,996 pairs of argumentative texts and associated conclusions, creating the first large-scale ground truth for conclusion generation. (3) Modeling conclusion generation as an end-to-end task by finetuning a pretrained sequence-to-sequence model, and augmenting the corpus with three types of argumentative knowledge: topic, target, and aspect. (4) Extensive quantitative and qualitative (crowdsourced) evaluation of both the quality of our dataset and the effectiveness of two paradigms for conclusion generation, namely extractive and abstractive approaches.

We present three key findings: (a) Finetuning pretrained language models on our dataset shows strong in-domain performance compared to the extractive approach. (b) Qualitative evaluation shows that the extractive approach generates more informative conclusions, demonstrating a trade-off between conciseness and informativeness. (c) Encoding argumentative knowledge guides the finetuning towards generating argumentative sentences; however, more sophisticated encoding techniques than just using the conventional control codes are needed to generate informative conclusions.

Related Work

Our work complements and builds on that of Alshomary et al. [10], who introduced a conceptual model for conclusion generation, outlining a three-step process: inferring the conclusion's target from the argument's premises, inferring the author's stance towards this target, and generating the conclusion based on these two pieces of information. But Alshomary et al. focused only on the first step of target inference, whereas we model conclusion generation as an end-to-end task.

Conclusion generation can be viewed as a complementary task to summarizing argumentative texts. Previous approaches to the summarization of such texts have been primarily extractive. Egan et al. [90] proposed summarizing online discussions via "point" extraction, where a point is a verb and its syntactic arguments. Similarly, Bar-Haim et al. [22] compiled the *ArgKP* corpus (which we also sample from in Section 5.1.2) comprised of arguments for a given topic mapped to *key points*, composing a summary from a large collection of relevant arguments. Wang and Ling [325] proposed a data-driven approach using sequence-to-sequence models [18, 283] for summarizing movie reviews and debate portal arguments from idebate.org. Several argument mining approaches have also been applied to identify the main claim from arguments [75, 229]. Recently, Alshomary et al. [9] proposed a graph-based model using PageRank [222] that extracts

the argument's conclusion and the main supporting reason as an extractive snippet. This model is the core of our extractive summarization approach (Section 5.2).

A key difference between conclusion generation and general text summarization is the constraint that a conclusion must have a clear stance towards a certain topic. A similar constraint applies to high-quality summaries of long-form argumentative texts such as editorials as described in Chapter 3, where the persuasiveness of the editorial should be preserved alongside its thesis. Therefore, existing summarization corpora (although large-scale) are unsuitable for studying conclusion generation. A majority of them contain only non-argumentative texts (e.g., news reports) which are more suitable to general-purpose summarization [167]. Moreover, intrinsic evaluation of summarization corpora has revealed a lower-quality and/or inconsistent ground truth, rendering them partially unfit for their intended purpose [36]. To fill this gap, we compile Webis-ConcluGen-21, a large-scale corpus of argumentative texts and their conclusions on diverse topics.

Pre-trained language models have significantly advanced the state-of-the-art in neural text summarization [140, 180, 255, 338]. However, they have been applied to the domain of argumentation only recently, specifically for argument generation. Gretz et al. [116] proposed a pipeline based on GPT-2 [244] for generating coherent claims for a given debate topic. A more controlled approach for argument generation was developed by Schiller et al. [264], which performs argument generation with fine-grained control of topic, aspect (core reasoning), and stance.

Conclusion generation can be viewed as supplementing argument generation. Ideally, given a conclusion, an argument can be generated constrained by the conclusion's target and stance. To the best of our knowledge, studies investigating pretrained language models for end-to-end conclusion generation do not exist. Besides providing a suitable corpus, we analyze the impact of encoding argumentative knowledge in pretrained language models and assess the popular method of control codes [50, 160] for encoding the knowledge in our dataset. Furthermore, our qualitative evaluation highlights three key errors (Section 5.3) arising in the generated outputs that disqualify them as conclusions.

5.1.1 On Informative Conclusions

In the literature, the conclusion of an argument is the statement that depicts a particular stance towards a certain concept, the target [10, 321]. Such a statement is also referred to as the *claim* of the argument [75, 300]. For a

long-form argumentative text with multiple claims, the conclusion is the *main claim* that conveys the overall stance towards the subject matter under discussion. The main claim is also known as *thesis*, or *central claim* in different genres [49, 227, 278, 305].

The quality of the conclusion of an argumentative text can be assessed in terms of several dimensions, including strength, clarity, and specificity [157]. Here, a strong connection between argumentation and text summarization can be observed, where the dimension corresponding to specificity is called *informativeness*. Text summarization distinguishes between indicative and informative summaries. An indicative summary only hints at the principal subject matter of a document to help decide whether to read it [136, 153]. An informative summary, on the other hand, covers the main information in the source document, ideally serving as its surrogate [197].

The conceptual connection between argumentation and summarization could be described as follows: the *informativeness* of a conclusion is closely connected to the specificity dimension, in the sense that an informative conclusion must be specific to allow for a better understanding of an argumentative text's gist. Seeing that "specificity" and "informativeness" may be used interchangeably, we opted for the latter and the term "informative conclusion" here, to underline the connection.

In contrast to *indicative* conclusions, which broadly convey (implicitly or explicitly) the stance towards a topic (e.g., "Caffeine is good."), informative conclusions also discuss specific concepts from (or implied by) the argumentative text (e.g., "Caffeine improves physical performance."). Concepts of the argumentative text exemplified in Section 6.1 may refer to the topic (e.g., "Is coffee beneficial?"), the target of the conclusion (e.g., "caffeine"), or a specific aspect (e.g.,"energy levels").

5.1.2 The Webis-ConcluGen-21 Corpus

This section details the construction of the Webis Conclusion Generation Corpus 2021 (Webis-ConcluGen-21), a corpus of 136,996 pairs of argumentative texts and conclusions covering diverse topics. The corpus is derived from two reliable sources, where the conclusions of argumentative texts are explicitly identifiable: Reddit's ChangeMyView forum and debate corpora.

Data Source: Reddit's ChangeMyView

ChangeMyView (CMV) is an online forum for persuasive discussions that start with a user who presents a view and asks others to challenge it. The

Type	Description	%
Extractive	Conclusion is present verbatim in the argumentative text.	12.8
Paraphrase	Conclusion is synonymous to, or a fusion of a part of the argumentative text.	24.1
Abstractive	Conclusion is inferred from the argumentative text.	57.8
No conclusion	Conclusion cannot be derived from the argumentative text.	5.3

Table 5.1: Different types of conclusions in 200 CMV samples, and their relative proportion.

forum's rules strictly enforce that (1) users' posts must contain sufficient reasoning, (2) posts must take a stance (and not be neutral), and (3) the title of a post must sufficiently sum up an author's view (as a statement and not a question).[5] Given these constraints, the original post of a discussion can be operationalized as an *argumentative text*, and the corresponding title as its (intended) *conclusion*. Starting from the Reddit crawls provided by Baumgartner et al. [24], we compiled 61,695 such pairs by processing all CMV discussions up until August 2019. The included posts are those whose argumentative text was longer than ten words, the conclusion longer than two words, and the title includes the "CMV" tag.[6] An average argumentative text is 312 words long and a conclusion 15 words.

To better understand the relation of the conclusions to their respective argumentative texts, and the expected difficulty of generating them, we analyzed a sample of 200 pairs manually.[7] Table 5.1 shows the proportion of extractive, paraphrased, and abstractive conclusions in our sample, where the former only need to be extracted, and the latter demand actual text synthesis. Paraphrases share aspects of both, though arguably, extracting the paraphrased part would suffice. Altogether, CMV provides for 94.7% valid pairs of argumentative texts and conclusions at sufficiently low noise (5.3%). The amount of non-trivial conclusions (abstractive + paraphrase) are sufficiently challenging, as found in our qualitative evaluation (Section 5.3).

[5] `https://old.reddit.com/r/changemyview/wiki/rules`

[6] These heuristics reflect manual inspections, and the fact that we did not wish to compile a representative sample of ChangeMyView's discussions, but a purposeful selection of high-quality pairs of argumentative texts and their conclusions: In light of this, the lower bounds are still quite inclusive with respect to extremely short samples.

[7] These examples were taken from the Dec-2019 Reddit submissions to ensure a truly-hidden sample as BART was originally trained on the OpenWebText dataset containing samples from Reddit [180, 244].

5.1.3 Data Source: Debate Corpora

Online debate portals facilitate semi-structured debates on controversial topics, where pro and con arguments or argumentative texts are collected. Conclusions are clearly stated even for individual arguments. Given their high-quality curation, debate portals constitute the majority of argument corpora. We utilized the following existing corpora:

Kialo. is a debate platform that enables "visual reasoning" in complex debates via a tree-based structure [57]. A key advantage here is the role of moderators in curating accepted arguments, rendering it a rich resource [87]. As debates progress, the arguments are reorganized into multiple hierarchies, each with a conclusion at its root.[8] We compiled this corpus from scratch in accordance with the website's terms and conditions. In 1,640 English discussions, at each level of the discussion tree, all pro arguments were matched to the corresponding root conclusion, obtaining a total of 82,728 examples.

Args.me. is a search engine [316] indexing the Args.me Corpus [2], comprised of argumentative texts, their conclusions and their stance from four debate portals: `debatewise.org`, `idebate.org`, `debatepedia.org`, and `debate.org`. We used the "cleaned" version of this corpus containing 387,606 samples and applied further post-processing. On manual inspection, we observed that a number of examples from `debate.org` contained spam, sarcasm, or ad hominem attacks, or they were not self-contained due to references to previous turns. To avoid noise, we excluded all examples from this portal. Next, we removed arguments with con stance towards a conclusion.[9] This is due to the fact that considering these examples for training would first require negating their conclusions to reflect the con stance. We leave such automatic claim negation [32] for future work. Finally, to favor informative conclusions, we excluded arguments whose conclusion was the same as the discussion topic (which is generally indicative). This heavy filtering resulted in a total of 23,448 argument-conclusion pairs.

ArgsKP. is a corpus of arguments and a set of key points written by domain experts on 28 topics [22]. For each topic, the corpus contains multiple arguments which have been mapped via crowdsourcing to their respective key points. From this corpus, we obtained 2,341 pairs; again, only pro arguments and those that have been mapped to a specific key point, the conclusion.

[8] For an example, see: `https://www.kialo.com/pro-life-vs-pro-choice-should-abortion-be-legal-5637`

[9] This does not exclude conclusions that are already negations.

Postprocessing.. The structure of debate portals allows for multiple arguments to be mapped to a single conclusion. This happens when different users independently contribute pro and con arguments, which is acceptable, since the same conclusion can be drawn from different arguments with different frames [1]. Apart from the ones filtered in preprocessing the debates corpora, we preserved duplicate conclusions across debates as their arguments are still unique. Similar to CMV, the included argumentative texts were those whose length exceeded ten words. Also, argumentative texts shorter than their conclusion were excluded. This removed many pairs from the Kialo discussions. Altogether, we retained 75,301 usable examples from all three corpora.

5.1.4 Corpus Statistics

The argumentative texts are on average longer in CMV (312 words) compared to those in debates (44.5 words). A reason is that, on debate portals, each argumentative text seems to be a self-contained argument. CMV posts, by comparison, often contain multiple arguments and/or preface the actual argument with additional background. However, the corresponding conclusions are of similar length (15 words for CMV and 18.4 words for debates on average, about the length of an average English sentence). For both data sources, we measured the percentage of words in a conclusion that do not occur in the argumentative text as a measure of "novelty" [209]. For CMV, the average novelty is 33.2%, and for debates, the novelty is 81.6%, which is due to the fact that multiple arguments have been mapped to a single conclusion, and that arguments supporting (or attacking) a conclusion during an ongoing discussion are usually not directly derived from it.

5.2 Enhancing Pretrained Models with External Knowledge

Given the mixture of conclusion types shown in Table 5.1, we approach the generation of informative conclusions according to two paradigms, one extractive approach combined with paraphrasing, and one abstractive approach combined with state-of-the-art argument mining technology.

5.2.1 Paraphrased Conclusion Generation

Paraphrased conclusions are fundamentally extractive in nature, where an extracted sentence is reformulated to improve it. To extract conclusions, we employ the graph-based approach of Alshomary et al. [9], originally designed to generate snippets for argument search results. Given an ar-

gument, a snippet is generated as follows: (1) related arguments are retrieved as context, (2) all argument's sentences and those from the retrieved ones are embedded, (3) the PageRank of the sentences is computed, and lastly (4) the argument's two top-ranked sentences are returned. Underlying this approach is the hypothesis that an extractive snippet for an argument should comprise its conclusion and its most important supporting premise. Sentences are thus scored regarding their centrality in context of other arguments and their argumentativeness.

Our goal is to generate a single conclusion statement, thus we consider only the top-ranked sentence as the conclusion from the approach of Alshomary et al. [9]. This sentence is automatically paraphrased using PEGASUS [339], finetuned on the Google PAWS dataset [343].[10] For instance, consider the top-ranked sentence from a post questioning the use of hormone blockers on transgender kids:[11]

"I don't see it as anything different, and I think it is scandalous to permanently change a child's entire life on a whim rather than treating their mental health."

After paraphrasing, it reads as follows:

"I think it's scandalous to change a child's life on a whim, rather than treating their mental health, and I don't see it as anything different."

The paraphraser primarily rearranges the sentence; and shared phrases with the original are typical in the paraphrased sentences we reviewed. This approach, called `Arg-PageRank`, represents an advanced extractive paradigm.

5.2.2 Abstractive Conclusion Generation

Abstractive conclusions can be formulated freely, provided they capture the main pieces of information required for an informative conclusion: topic, targets, stance, and aspects. In this regard, our approach is three-fold (see Figure 5.1): (1) Automatic extraction of the aforementioned pieces of information from a given argumentative text; (2) augmentation of the training examples in Webis-ConcluGen-21 using control codes, and (3) domain transfer of a pretrained abstractive news summarization model via finetuning on the augmented corpus.

Argumentative Knowledge Extraction: This step details our respective approaches at providing the prerequisite pieces of information to formulate an

[10] `https://huggingface.co/tuner007/pegasus_paraphrase`

[11] `https://old.reddit.com/r/changemyview/comments/e97sir/cmv_giving_children_puberty_blockers_to_allow/`

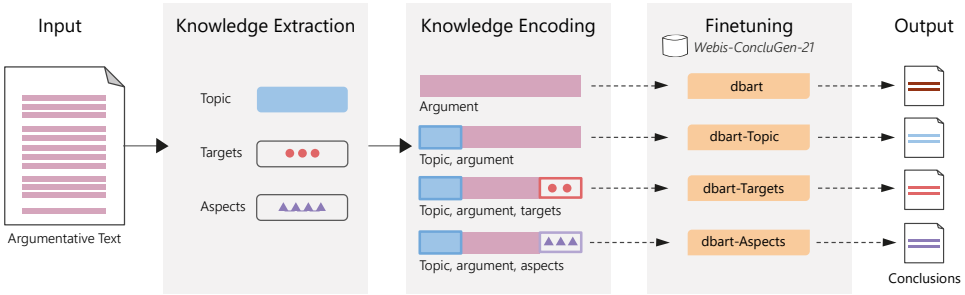


Figure 5.1: The three steps of our approach to abstractive conclusion generation: For all examples in the Webis-ConcluGen-21 corpus (1) different pieces of argument knowledge are extracted namely the discussion topic, possible conclusion targets, and covered aspects, (2) this knowledge is encoded using control codes, and (3) knowledge-specific variations are finetuned of the distilled BART model to generate informative conclusions.

informative conclusion, namely topic, targets, and aspects. Table 5.2 shows an example.

Topic: An argumentative text's topic is a description of what it is about. For argumentative texts from debates, we use the associated debate title as the topic. For CMV posts, their titles are also their conclusions; here, topic information is considered missing (denoted as 'NA' token).

Targets: The target of a conclusion is typically a controversial concept or statement [21]. For an argumentative text, though, an overlap with its topic is possible, different targets can also be found in its premises. Moreover, when not explicitly stated, the targets of a conclusion can be inferred from either the targets of premises, or external knowledge bases. A set of possible targets for every argumentative text in the corpus are automatically identified using the target identification model of Alshomary et al. [10].

Aspects: Text spans that contribute to the core reasoning of an argument are called its aspects [264]. Aspects can be viewed as subtopics related to the main topic of an argumentative text, encoding a stance. Including aspects into a conclusion can render it more specific and, thus, informative. We identify aspects for all samples in the corpus, using the model of Schiller et al. This model trains a BERT-based [83] ranker on a corpus containing 5,032 high-quality argumentative sentences that are manually labeled with aspects at the token level.

Stance is excluded as an explicit input to our models. For CMV, by design, a post supports its title. For debate portals, only argumentative texts with pro stance towards their conclusion have been considered. Nevertheless, argumentative texts and their conclusions in our corpus may, implicitly or

Argument	Feminism as a 'linguistic term' often misses clarity, universal definition and regularly incorporates opposite goals at the same time in regard to key feminist issues as gender equality, gender-neutrality, non-binary and gender-related rights. The linguistic term thereby clouds public debate and hampers the setting of clear social and political goals in society.
Conclusion	Feminism is an umbrella of ideologies first and foremost, and consequently, it muddies the discussion of gender equality with its ideological baggage.
Topic	Is Feminism a Force For Good?
Aspects	clouds, gender equality, non-binary, opposite goals, public debate, gender-related rights, clarity, gender-neutrality, social and political goals, universal definition
Targets	The linguistic term, Feminism as a ' linguistic term'
Encoded Input	<|TOPIC|>Is Feminism a Force For Good?<|ARGUMENT|>Feminism as a 'linguistic term' often misses clarity, universal definition and regularly incorporates opposite goals at the same time in regard to key feminist issues as gender equality, gender-neutrality, non-binary and gender-related rights. The linguistic term thereby clouds public debate and hampers the setting of clear social and political goals in society.<|TARGETS|> The linguistic term, Feminism as a ' linguistic term<|CONCLUSION|>

Table 5.2: Example argument-conclusion pair along with topic, targets, and aspects. The last row shows the input format for finetuning models on specific types of encoded external knowledge (here, on conclusion targets).

explicitly, express their own stance towards implicit or explicit targets. Implicit stance can be encoded via the aspects.

Argumentative Knowledge Encoding.. The extracted pieces of knowledge are encoded into a training example with control codes using special tokens Cachola et al. [50]: <|TOPIC|>, <|ARGUMENT|>, <|ASPECTS|>, <|TARGETS|>, and <|CONCLUSION|>. Table 5.2 shows a corresponding example input sequence encoding the topic and the conclusion targets. To examine the impact of individual knowledge types, we create three versions of Webis-ConcluGen-21: *topic-encoded*, *aspect-encoded*, and *target-encoded*. Presuming the availability of a topic in nearly all real-world applications, it is also encoded in the latter two versions. Since aspects and targets overlap in 38.3% of the case in the corpus, they are independently encoded.

Finetuning.. As conclusion generation is closely related to abstractive text summarization, we picked BART [173], a pretrained state-of-the-art summarization model, for finetuning on the three augmented versions of Webis-

Model	Data	#Train	#Valid
dbart-XSum	XSum	204,045	n/a
dbart-CMV	CMV	55,768	5,577
dbart-Debates	Debates	67,770	6,777
dbart	All	123,538	12,354
dbart-Topic	All+topic	123,538	12,354
dbart-Aspects	All+topic+aspects	122,040	12,192
dbart-Targets	All+topic+targets	110,867	11,068
Arg-PageRank	*none, unsupervised model*		

Table 5.3: Corpus splits for all six variants. 'All' refers to the entire Webis-ConcluGen-21 corpus. Models were automatically evaluated on a test set of 1,000 examples, and qualitatively on 300 examples (Section 5.3).

ConcluGen-21. However, BART has approximately 10% more parameters than BERT, which makes it resource-intensive for finetuning. To account for this, we used the distilled checkpoint derived using the "shrink-and-finetune" approach of Shleifer and Rush [276], where large sequence-to-sequence models are compressed by extracting "distilled student models" [260] from a teacher model (here, BART). We used distilled BART finetuned on the XSum corpus [209] (`dbart-XSum`) provided by the Transformers library [330],[12] since the average length of our ground truth conclusions is similar to the summaries in XSum. Additionally, we also added our control codes as special tokens to the BART tokenizer during finetuning in order to avoid splitting them into sub-word tokens while processing the encoded sequences.

We first applied `dbart-XSum` on the held-out test set of 200 examples analyzed for Table 5.1 to evaluate the domain transfer from news reports to argumentative texts. On manual evaluation, 79.1% of the outputs were invalid conclusions, primarily due to being non-argumentative (Section 5.3). This demonstrates that existing summarization models are ineffective when applied on argumentative texts and must be trained on task-specific data.

5.2.3 Training Details

We compiled six variations of the corpus (with and without encoded knowledge) for finetuning the The `dbart-XSum` model with 306M parameters.[12] Table 5.3 shows the training and validation splits for each model variant and the corresponding data subsets, and Table 5.4 shows the chosen hyperparameters. The standard finetuning regimen was employed

[12] `https://huggingface.co/sshleifer/distilbart-xsum-12-6`

Parameter	Value
max_target_length	100
warmup_steps	500
eval_steps	500
attention_dropout	0.1
label_smoothing	0.1
sampling	sortish_sampler
seed	5153
num_beams	6
length_penalty	0.5
gradient_accumulation_steps	1
lr_scheduler	linear

Table 5.4: Hyperparameters for finetuning BART.

from the Transformers library[13] to train each model on a V100 GPU for 6 epochs with batch size 1, dropout rate 0.1, adafactor optimizer, learning rate of 3e-5, and beam search for inference. For `dbart-<CMV|Debates|All>` the maximum source sequence length was set to 512 tokens, while for `dbart-<Topic|Aspects|Targets>` we increased it to 750 tokens to account for the appended knowledge in the input sequence. On a single V100 GPU, the runtime varies between 3 to 5 days per model, depending on their corresponding training splits.

5.3 Evaluating Informative Conclusion Generation

Our models are evaluated via both: (1) An automatic evaluation on a large test set using standard metrics, and (2) a manual evaluation on a smaller test set via crowdsourcing.

5.3.1 Automatic Evaluation

On a test set of 1,000 examples with known ground truth (500 each from CMV and from the debate corpora), we computed ROUGE [176][14] and BERTScore [341][15] for all models. Table 5.5 shows that `dbart-XSum` performs poorly on argumentative texts. Inspecting the reasons for this shortcoming, we found several outputs of the model to be either neutral sentences (despite having the right target), or hallucinations with artifacts from the XSum corpus (e.g., "*In our series of letters from African journalists* [...]" or "*This week I've been writing about* [...]"). Among the finetuned models, `dbart`,

[13] `https://github.com/huggingface/transformers/tree/master/examples/legacy/seq2seq`

[14] `https://github.com/pltrdy/rouge`

[15] `https://github.com/Tiiiger/bert_score`

Model	BERTScore (F1)	Rou.-1	Rou.-2	Rou.-L
dbart-XSum	0.21	15.28	3.10	13.31
dbart-CMV	0.32	20.35	7.11	18.80
dbart-Debates	0.23	15.38	4.85	14.22
dbart	**0.39**	**31.73**	**19.48**	**30.87**
dbart-Topic	0.34	23.74	9.56	22.14
dbart-Aspects	0.33	23.47	9.46	22.01
dbart-Targets	0.34	23.80	9.63	22.25
Arg-PageRank	0.20	15.35	3.20	13.37

Table 5.5: Automatic evaluation of models on the internal test set consisting of 1,000 pairs (500 each from CMV and Debates). BERTScore is the re-scaled F1 score; in addition, average Rouge-1, -2, and -L are reported.

trained on the entire corpus without any encoded knowledge, performs best across all metrics. The knowledge-encoded models exert a drop in effectiveness, but still outperform models trained on the sub-datasets `dbart-CMV` and `dbart-Debates`.

All finetuned models generate concise outputs of similar lengths (average 12 words), while `Arg-PageRank` extracts longer spans (25 words). Outputs of the knowledge-encoded models are somewhat similar to each other (average pairwise Jaccard similarity of 0.43), compared to those from `dbart` (0.27 with any knowledge-encoded model).

5.3.2 Manual Evaluation

Given the results of the automatic evaluation, only the models trained on the entire corpus were manually evaluated against our baseline approach `Arg-PageRank`. A test set of 300 examples was employed, 100 each from debates and CMV posts, plus 100 comments to CMV posts. The latter include only comments with at least 100 words and exclude non-argumentative ones as per automatic claim-detection [56]. This part of the test set corresponds to an unsupervised evaluation of the conclusions, since no ground truth for the comments is available.

Two expert writers, both native English speakers, were hired via `Upwork.com`.[16] For every given argumentative text in the test set, all candidate conclusions generated by the different models were shown to the annotators in random order, and without revealing the respective model's name. Assessment was cast as a series of binary decisions: first, whether a given candidate is a conclusion, and if yes, whether it is fluent, and whether it is informative. To simplify judging informativeness, we only asked if the conclusion was too generic. For each candidate judged not to be a conclusion,

[16] An hourly rate of about 30 USD was paid.

Model	Concl.	Inform.	Error Types		
			Wrong Target	*Wrong Stance*	*Non-argumentative*
CMV Posts					
`dbart`	36%	4%	56%	22%	22%
`dbart-Topic`	28%	0%	59%	23%	18%
`dbart-Aspects`	33%	6%	69%	23%	8%
`dbart-Targets`	27%	4%	69%	23%	8%
`Arg-PageRank`	11%	7%	0%	0%	100%
Debates					
`dbart`	14%	6%	65%	9%	26%
`dbart-Topic`	14%	3%	76%	12%	12%
`dbart-Aspects`	7%	2%	77%	13%	10%
`dbart-Targets`	11%	2%	71%	17%	12%
`Arg-PageRank`	10%	6%	7%	0%	93%
Comments					
`dbart`	12%	2%	52%	18%	30%
`dbart-Topic`	6%	2%	58%	24%	18%
`dbart-Aspects`	7%	3%	52%	33%	15%
`dbart-Targets`	8%	3%	55%	35%	10%
`Arg-PageRank`	17%	9%	5%	5%	90%

Table 5.6: Full agreement percentages of two annotators on 300 examples, grouped by the example type (posts, debates, comments). The first column is the % of valid conclusions, the second the % of informative conclusions, followed by the % distribution of error types (lower is better) of a model. On average, all models were judged to be fluent for 97% of the conclusions.

we asked whether it either has the (1) *wrong target*, conveys the (2) *wrong stance*, or whether it is (3) *non-argumentative*.

Table 5.6 shows the percentage of cases on which both annotators agreed. For CMV and debates, finetuning outperforms `Arg-PageRank` at generating conclusions that convince the experts: `dbart` performs best on CMV (36%), and `dbart` and `dbart-Topic` on debates (14%).

Comments appear to be a particularly difficult type of test cases. This is because comments to the first post may not be self-contained but refer back to the post, they may have a mixed stance (supporting only part of the post while opposing the rest), and they may introduce new targets and aspects (different concepts)—based on our inspection of the comments. In such cases, extracting the conclusion from the comment (and paraphrasing it) using `Arg-PageRank` performs best (17%).

Encoding knowledge slightly impacts the effectiveness. Across all example types, knowledge-encoded models perform equally well, sometimes worse, sometimes better than `dbart`. Encoding topic with aspects or targets performs better on posts and comments.

As for *informativeness*, `dbart-Aspects` generates a higher number of informative conclusions for posts, while `dbart` does best in debates, among the finetuned models. In all domains, `Arg-PageRank` performs similar to or better than all approaches due to extracting claims that are twice as long on average (24 words) compared to the finetuned models (12 words), hence capturing more information.

Inspecting the various error types, we observed that encoding argumentative knowledge increases the number of argumentative candidate conclusions, validating its positive impact. All knowledge-encoded models have fewer non-argumentative errors compared to `dbart`. However, this affects target inference; the knowledge-encoded models generate more wrong targets. The mixed stance of comments (supporting part of the original post, while opposing the rest) leads to a higher number of stance errors for `dbart-Aspects` and `dbart-Targets`. Finally, for `Arg-PageRank`, almost all errors were non-argumentative sentences.

5.3.3 Discussion

Our qualitative evaluation indicates that generating informative conclusions is challenging, and that our data is well-suited for the task, due to a mix of conclusion types (Table 5.1), and diverse data sources. Leveraging external knowledge, though a promising feature for guiding finetuning, may benefit from better encoding strategies compared to the conventional method of using control codes in text. However, given that the identified knowledge is extractive and that we encoded multiple aspects and targets per example in contrast to related controlled text generation approaches [50, 116, 160, 264], further investigations with importance sampling of argumentative knowledge are advised. Ideally, such sampling would be tailored to a specific domain or target audience.

Likewise, regarding the informativeness of the generated conclusions, a trade-off between conciseness and specificity must be decided. Our experiments suggest that long extractive conclusions capture more information compared to the more concise (and fluent) abstractive one of the finetuned models, rendering them preferable to the annotators when sufficient background is missing. Finally, for comments, modeling the argumentative context supplemented by explicit stance identification is necessary to generate valid conclusions.

5.4 Summary

In this chapter, we introduced the notion of an informative conclusion in the context of computational argumentation as well as text summarization. Informative conclusions are to argumentation what brief summaries are to text: they concisely convey its main points while expressing a stance towards a certain target. We laid the foundation for studying the conclusions of argumentative texts, compiling the Webis-ConcluGen-21 corpus, comprising 136,996 pairs of argumentative texts and corresponding conclusions.

Conclusions are diverse and typically depart significantly from the argumentative text they are derived from, paraphrasing it, and more than half the time abstracting over it. Authors typically tailor their conclusions to the occasion; and in many cases, they are not necessarily made explicit. This is where we contribute by tackling the task of generating an informative conclusion. The two main paradigms we study—paraphrased (incl. extractive) vs. abstractive conclusion generation—compete closely with each other.

6

Frame-Oriented Summarization of Argumentative Discussions

This chapter advances from summarizing individual argumentative texts to entire discussions with multiple participants and up to hundreds of arguments. However, in contrast to generating a single summary for the entire discussion, we propose a novel paradigm of *frame-oriented* summarization, where argumentation frames are employed as anchor points to group the discussions' arguments. A summary is then compiled for each frame in an extractive fashion. This enables the reader to choose from multiple summaries that best fit their information need. We first describe our approach to frame assignment, followed by methods for re-ranking arguments of a frame based on their relevance to the discussion topic and informativeness. We then describe the dataset on which our approach was evaluated, the various retrieval models with their respective parameters, and the content features that we used in our experiments. Also described is the supervised baseline for frame assignment that we implemented to assign multiple frames to each argument. Finally, we present the results of our experiments and discuss the implications of our findings.

6.1 Importance of Summaries for Argumentative Discussions

Web-based forums like Reddit facilitate discussions on all kinds of topics. Given the size and scope of some communities (known as "Subreddits"), multiple individuals regularly participate in the discussions of timely con-

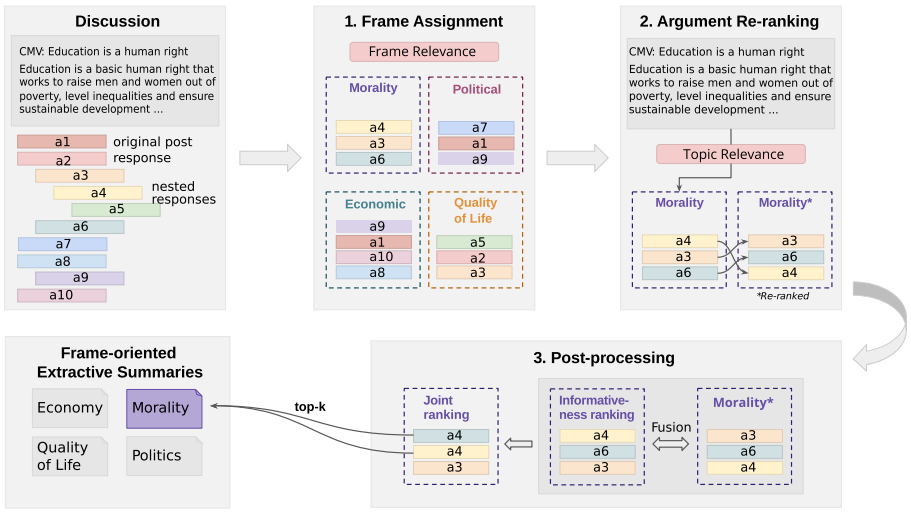


Figure 6.1: The proposed modular approach to frame-oriented discussion summarization: *1. Frame assignment* assigns arguments to frames ensuring frame relevance. *2. Argument re-ranking* ensures topic relevance of a frame's arguments (here, the *morality* frame is exemplified). *3. Post-processing* fuses the re-ranked arguments with an informativeness ranking. The top-k arguments are then taken as an extractive summary of the discussion.

troversial topics, such as on ChangeMyView.[1] Notably, the volume of arguments tends to grow substantially in a tree-like response structure wherein each branch forms a concurrent discussion thread. These threads develop in parallel as different perspectives are introduced by the participants. After a discussion subsides, the resulting collection of threads and their arguments often represents a comprehensive overview of the most pertinent perspectives (henceforth, referred to as *frames*) put forth by the participants.

Frames help shape one's understanding of the topic and deliberating one's own stance [63, 94]. However, in large discussions, prominent arguments as well as the various frames covered may be distributed in arbitrary (and often implicit) ways across the various threads. This makes it challenging for participants to easily identify and contribute arguments to the discussion. Large online forums like Reddit typically provide features that enable the reorganization of posts, for example, based on their popularity, time of creation, or in a question–answer format. A popularity-based ranking may seem beneficial, but Kano et al. [154] discovered that an argument's popularity is not well correlated with its informativeness. Furthermore, a popularity-based ranking does not cover the breadth of frames of a discussion, as we will show in this paper (Section 6.3.3).

[1] (CMV) `https://old.reddit.com/r/changemyview/`

In this paper, we cast discussion summarization as a ranking task with an emphasis on frame diversity, thereby introducing a new paradigm to discussion summarization in the form of *multiple* summaries per discussion (one per frame). Previous research has focused on creating a single summary per discussion instead. As illustrated in Figure 6.1, we first assign arguments to one or more frames. Next, we re-rank arguments in a frame according to their topic relevance. Additionally, we also rank them based on their informativeness via post-processing. Finally, we fuse these rankings to create the final ranking from which the top-k candidates can be used as an *extractive* summary of the discussion centered around a specific frame.

In our experiments, we explore various state-of-the-art methods to realize the three steps of our approach. Our results suggest that: (1) Utilizing retrieval models together with query variants is an effective method for frame assignment, reducing the reliance on large labeled datasets. Here, our approach outperforms a state-of-the-art supervised baseline. (2) Re-ranking arguments of a frame based on content overlap with the discussion topic is more effective than retrieval-based approaches for ensuring the relevance of the frame's arguments to the topic. (3) Post-processing the argument rankings based solely on content features is insufficient to signal informativeness.

In summary, our contributions include: (1) A fully unsupervised frame assignment approach that assigns one or more frame labels to every argument within a discussion (Section 6.3.1). (2) An argument retrieval approach that ranks frame-specific arguments based on their topic relevance and informativeness (Section 6.3.2). (3) A dataset consisting of 1871 arguments sourced from 100 ChangeMyView discussions, where each argument has been judged in terms of frame relevance, topic relevance, and informativeness (Section 6.3.3) which forms the basis for an extensive comparative evaluation (Section 6.4).[2]

Related Work

Previous approaches to summarizing discussions can be broadly classified into two categories: *discussion unit extraction* and *discussion unit grouping*. We survey the literature on discussion summarization according to these two categories, followed by the literature on *argument framing*.

[2]Code and data: `https://github.com/webis-de/SIGDIAL-23`

6.1.1 Discussion Unit Extraction

Extraction-based approaches use either heuristics or supervised learning to identify important units, such as key phrases, sentences, or arguments within a discussion, then presented as the summary.

Tigelaar et al. [297] identified several features for identifying key sentences from the discussion, such as the use of explicit author names to detect the response-tree structure, quoted sentences from the preceding arguments, and author-specific features such as participation and talkativity. They found that, while these features can be helpful, summarizing discussions primarily involves balancing coherence and coverage in the summaries. Ren et al. [251] developed a hierarchical Bayesian model trained on labeled data to track the various topics within a discussion and a random walk algorithm to greedily select the most representative sentences for the summary. Ranade et al. [247] extracted relevant and sentiment-rich sentences from debates, using lexical features to create indicative summaries. Bhatia et al. [30] leveraged manually annotated dialogue acts to extract key posts as a concise summary of discussions on question-answering forums (Ubuntu, TripAdvisor). This dataset was further extended with more annotations by Tarnpradab et al. [293] who proposed a hierarchical attention network for extractive summarization of forum discussions. Egan et al. [90] extracted key content from discussions via "point" extraction derived from a dependency parse graph structure, where a point is a verb together with its syntactic arguments.

Closely related to the domain we consider, Kano et al. [154, 155] studied the summarization of non-argumentative discussions on Reddit. They found that using the karma scores of posts was not correlated with their informativeness and that combining both local and global context features for comments was the most effective way to identify informative ones. Therefore, we do not rely on karma scores in our post-processing module (Section 6.3.3) and instead extract several content features for computing informativeness.

The outlined approaches all create a single summary for the entire discussion via end-to-end models. In contrast, we model the extraction of informative arguments organized by frames, thus enabling diverse summaries for a discussion. Furthermore, our experiments with unsupervised retrieval models for frame assignment (Section 6.3.3) enable us to assess the need to create labeled datasets beforehand to develop strong frame-oriented summarization models tailored to discussions.

6.1.2 Discussion Unit Grouping

Grouping-based approaches first categorize a discussion's units into explicit (or implicit) classes, such as queries, aspects, topics, dialogue acts, argument facets, or expert-labeled keypoints, and then generate individual summaries for each class. They rely on specific reference points to organize a discussion's units, providing flexibility to the readers by allowing them to choose from diverse summaries that best fit their information needs.

Qiu and Jiang [239] modeled the discovery of latent viewpoints to group arguments based on two user characteristics: *user identity*, as arguments from the same user are likely to contain the same viewpoint; and *user interaction*, as users with different viewpoints may express disagreement or attack each other, while those with similar viewpoints may support each other. Misra et al. [203] used summarization to discover repeating arguments and grouped them into facets. Reimers et al. [250] proposed agglomerative clustering via contextual embeddings to identify similar arguments on a sentence level based on their aspects.

Nguyen et al. [216] proposed an unsupervised approach to class-specific abstractive summarization of customer reviews with the goal of reducing generic and uninformative content in summaries. They model reviews in the context of topical classes of interest, which are treated as latent variables. These classes represent their reference points as latent variables to be discovered through supervised or reinforcement learning. In contrast, our frame inventory provides a more controlled—and thus more interpretable—set of reference points for discussion summarization. More recently Shapira et al. [274] proposed a query-assisted, sentence-level interactive summarization approach for news reports using reinforcement learning. Their approach consists of two subtasks of query-based sentence selection and generating query suggestions to enable an interactive setting. In our scenario, we enable this interaction via the predefined set of frames.

Summarizing public debates, Bar-Haim et al. [22, 23] investigated mapping similar arguments to expert-written key points. Bražinskas et al. [41] summarized product reviews by selecting subsets of informative reviews, treating the choice of review subset as a latent variable that is learned by a model trained on a dataset compiled from professional product review forums. Amplayo et al. [13] proposed aspect-controlled opinion summarization via employing multi-instance learning on a labeled dataset to identify aspects in reviews for grouping followed by summarization. The reference points of these approaches are defined either through manual annotations or distant supervision. Some of these reference points are highly

topic-specific, requiring them to be created manually for each topic, for instance, the key points from Bar-Haim et al. [22]. In contrast, we use a fixed and topic-independent set of reference points, namely media frames [40], grounded in framing theory [63].

6.2 Employing Argumentation Frames as Anchor Points

Framing theory was initially utilized to categorize (political) newspaper articles in order to manifest the specifically reported perspective [40, 214, 272]. It was first introduced to the field of argumentation by Naderi and Hirst [205]. Later, Ajjour et al. [1] modeled framing in argumentation more systematically, introducing automatically extracted, fine-grained, issue-specific frame labels. Heinisch and Cimiano [131] successfully combined computational argumentation with framing theory by showing a latent connection between the different frame granularities for the media frames defined by Boydstun et al. [40]. Hartmann et al. [130] also used frame-labeled data from newswire corpus to successfully train frame classifiers for political discussions via multi-task and adversarial learning. Following the literature, we use the media frames due to their wide adoption in categorizing arguments [55, 60].

6.3 Extractive Summarization of Argumentative Discussions

This section describes our ranking-based approach to the extractive summarization of online discussions, centered around argumentation frames (Figure 6.1). First we describe our novel unsupervised approach for frame assignment, followed by methods for re-ranking arguments of a frame based on their relevance to the discussion topic and informativeness. The top-k arguments from the joint ranking are taken as the frame's summary.

6.3.1 Frame Assignment

Our approach to frame assignment IRFrame is completely unsupervised in that it employs information retrieval models to rank arguments in a discussion by their *frame relevance*. Here, we consider arguments as documents and frames as queries. This offers a basic and interpretable alternative to frame assignment that does not require labeled data to train supervised models. We investigated both lexical and dense retrieval models.

We used an existing inventory of media frames to organize the arguments in a discussion. This originates from Boydstun et al. [40] and consists of

Frame Inventory	
Capacity & Resources	Fairness & Equality
Constitutionality & Jurisprudence	Health & Safety
	Morality
Crime & Punishment	Policy Prescription & Evaluation
Cultural Identity	Political
Economic	Public Opinion
External Regulation & Reputation	Quality of Life
	Security & Defense

Table 6.1: Inventory of frames proposed by Boydstun et al. [40] to track the media's framing of policy issues.

the 15 frames listed in Table 7.1. This inventory aims to support an issue-generic frame categorization of political communication. In the context of discussions on Reddit CMV, these issue-generic frames ideally cover a wide variety of controversial topics. The *other* frame is a catch-all category for frames that do not fit into any of the others. We excluded it from our experiments as it is not well-defined, and thus difficult to evaluate. For full frame descriptions see Table A.1 in the appendix.

Employing query variants—semantically related queries derived from the primary query—has been shown to improve the retrieval performance [27]. Thus, we manually created ten query variants for each frame to retrieve and rank all arguments in the discussion based on their frame relevance. Each variant is a high-quality sentence describing the various *aspects* of a frame. We manually curated these sentences from the Wikipedia pages of the frame labels as well as those of the various aspects mentioned in their descriptions (in Table A.1). For example, a query variant for the frame *cultural identity* is: "Cultural identity is defined as the identity of a group or culture or of an individual as far as one is influenced by one's belonging to a group or culture and is similar to, and overlaps, with identity politics". The complete list of query variants for all frames is provided in the supplementary material. The output of this module is a ranked list of arguments for each frame, which is then used for extractive summarization (Section 6.3.2).

We first obtained ten rankings of the arguments (one for each query variant) and then combined these via reciprocal rank fusion [70] to obtain the final list of ranked arguments for a frame. We also compare our approach with a supervised baseline, SuperFrame, a classifier finetuned on a set of labeled arguments (details in Section 6.3.3).

6.3.2 Extractive Summarization

Building upon the frame assignment component described above that ensures frame relevance, we now perform an *extractive* summarization of the discussion by re-ranking the frame-relevant arguments based on their relevance to the discussion topic and informativeness. This modular approach to summarizing discussions does not require expensive ground truth summaries, and is thus more scalable than supervised approaches. We first describe the argument re-ranking module followed by the post-processing module.

Argument Re-ranking Besides being relevant to a frame, arguments in the summary must also be relevant to the discussion topic. Thus, we re-rank the frame's arguments according to their *topic relevance*. In our scenario, a "topic" is the combination of the title and the reasoning of the original post on CMV. We propose two approaches for computing topic relevance. The first approach computes content overlap (lexical and semantic) between each argument and the topic. We used Jaccard similarity for lexical overlap, and for semantic overlap, we used the cosine similarity between the contextual sentence embeddings of an argument and the topic. Arguments within a frame are then re-ranked by their overlap scores. The second approach employs retrieval models and (re-)ranks the frame's arguments using the entire topic as the query (details in Section 6.3.3).

Post-processing Parallel to the aforementioned re-ranking by topic relevance, we derive a separate re-ranking of the frame's arguments based on their *informativeness*. Our goal is to prioritize content-rich and argumentative texts in the top-k arguments of our approach. We operationalize this through *content scoring* and *argumentativeness scoring*. For content scoring we employed a set of content-specific features such as named entities, noun phrases, the number of discourse markers, and the number of children an argument has in the discussion. Next, for argumentativeness scoring, we trained a topic-based argumentativeness scoring model (details in Section 6.3.3). The informativeness score of an argument is the sum of its content score and the argumentativeness score. We then re-rank the frame's arguments by this score.

Frame-oriented Extractive Summaries Given the list of arguments first ranked by frame relevance, then re-ranked by topic relevance, we fuse this ranking with the standalone informativeness ranking from the post-

processing module (via reciprocal rank fusion) to derive the final ranking. The top-k arguments from this ranking are taken as the *extractive* summary of the discussion. A key benefit of our ranking-based extractive summarization approach is the flexibility to determine the summary length (i.e., k) by the user according to the discussion's length and their information need. Thus we refrain from setting a specific length budget for the summary.

6.3.3 Data and Experiments

This section describes the dataset on which our approach was evaluated, the various retrieval models with their respective parameters, and the content features that we used in our experiments. Also described is the supervised baseline for frame classification SuperFrame that we implemented to assign multiple frames to each argument.

Data

We constructed a dataset of 100 long discussions from CMV, dated January 2020, using the Pushshift Reddit dataset [24]. For the purpose of this study, we defined a long discussion as a post with at least 100 comments. As preprocessing, we filtered out comments that were deleted by their authors, removed by moderators due to violating community rules, or posted by bots (e.g., DeltaBot, RemindMeBot). The average length of the posts in our dataset is 304 words, with a minimum of 83 words and a maximum of 1611 words. These posts have a total of 25,385 comments, with an average of 253 comments per discussion. The shortest discussion has 105 comments, while the longest has 1066 comments. The average length of a comment is 90 words, with a minimum of 2 words and a maximum of 1589 words excluding the quoted text from either the post or the parent comments they responded to.[3]

Popularity Ranking We investigated to what extent does ranking the arguments only by their popularity (via karma scores on Reddit) cover all the top-k arguments of the frames in the discussion (as assigned by our approach). To quantify this, we computed the mean coverage of the top 10 arguments across all frames and models by their popularity ranking. We considered discussions with at least 500 arguments and ranked them by

[3]The strict community guidelines of CMV (`https://old.reddit.com/r/changemyview/wiki/rules`) ensure that comments are primarily argumentative. Therefore, in this paper, we consider each comment to be an argument and do not perform any argument mining.

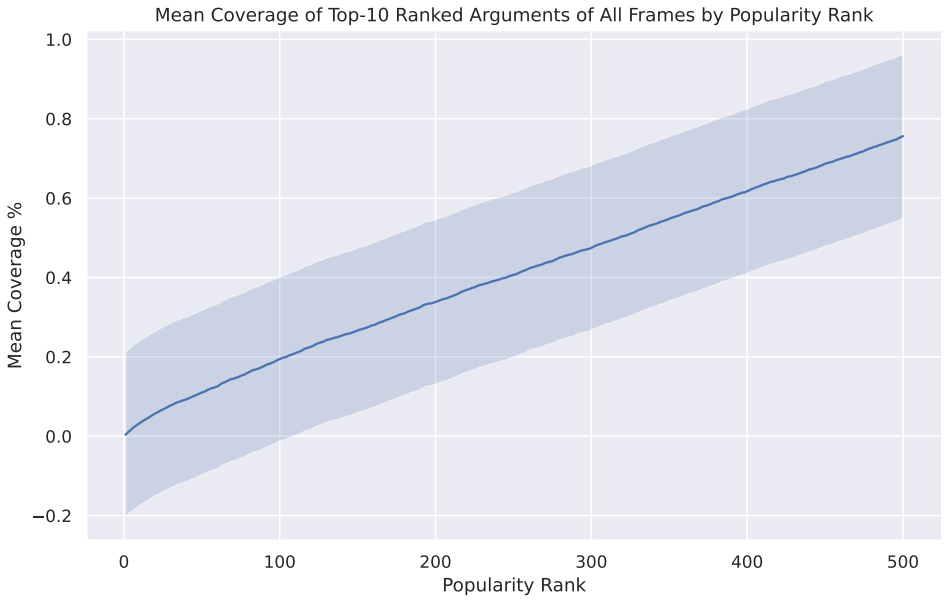


Figure 6.2: Mean coverage percentage by popularity rank of the top 10 (unique) frame arguments as assigned by our approaches.

their popularity scores provided by the Reddit API. Then, at each rank, we computed the percentage of top 10 arguments from all frames that have been covered by the popular arguments. Figure 6.2 shows that in order to completely cover the top 10 arguments from all frames, a user must read through hundreds of arguments. This encourages us to investigate novel approaches to group arguments in a discussion via frames instead of solely relying on their popularity. A similar conclusion was drawn by Kano et al. [154] who investigated the effectiveness of popularity scores as a feature for summarizing Reddit discussions.

Experiments

We first describe the models and parameters for our approaches to frame assignment and extractive summarization. We then describe the supervised baseline for frame assignment.

Frame Assignment We experimented with three retrieval models for IR-Frame to retrieve frame-relevant arguments: BM25 [253], SBERT [249], and ColBERT [161]. The latter two are dense retrieval models based on contextual embeddings to match arguments to frames, addressing the limitation of BM25 not finding arguments with exact lexical matches to our query

variants. We used the Okapi BM25 model with default settings (k=1.5, b=0.75),[4] initialized SBERT with the `all-mpnet-base-v2` model, and used ColBERT-v2 [262].[5]

Argument Re-ranking We experimented with two approaches to re-rank the arguments retrieved by IRFrame: content overlap and retrieval-based re-ranking. Content overlap considers both lexical and semantic overlap between the topic and the argument. For lexical overlap, we used Jaccard similarity and for semantic overlap, we used SBERT (`all-mpnet-base-v2` model). For the retrieval-based re-ranking, we experimented with BM25 and ColBERT, with the topic as the query, to (re-)rank the frame's arguments. We excluded SBERT as an additional retrieval model since it is already integrated in the content overlap approach.

Post-processing Informativeness is computed based on the content richness and the argumentativeness of the arguments. Content is scored as the sum of the ratios of named entities, discourse markers, and noun phrases found in the argument and the number of children for an argument in the discussion. We used spaCy [135] for text tokenization and extraction of the named entities and noun phrases.[6] For discourse markers, we used a lexicon of claim-related words constructed by Levy et al. [172] for identifying claim-containing sentences. The ratios of named entities and noun phrases were on the token level, while the ratio of discourse markers was on the word level, all normalized by the arguments' lengths. For argumentativeness, we developed *ArgDetector*,[7] a RoBERTa model [182] fine-tuned on the dataset by Schiller et al. [265], containing 150 controversial topics with 144 sentences labeled for their argumentativeness, given the topic. Implementation details are described in Appendix A.1.

SuperFrame This is the supervised baseline for frame assignment. Extending the state-of-the-art frame classification model of Heinisch and Cimiano [131], we developed a new classifier trained on an external frame-labeled dataset. The existing classifier of Heinisch and Cimiano [131], utilizes a recurrent neural network to assign a *single* frame to an argument, and combines it with a model that predicts a cluster of frame labels from

[4] We used the Rank BM25 toolkit [46]

[5] We used PyTerrier [188] for the ColBERT pipeline

[6] We used the `en_core_web_md` model.

[7] `https://huggingface.co/pheinisch/roberta-base-150T-argumentative-sentence-detector`

the inventory of Ajjour et al. [1] in a multi-task setting. Particularly longer arguments, however, often contain multiple frames. Thus, assigning a single frame to an argument may not be sufficient [250]. We therefore extend the model to predict *multiple* frames for an argument. Given the probability distribution of the classification model $P = (p_{f_1}, \cdots, p_{f_k})$ over a set of frames $\mathcal{F} = \{f_1, \ldots, f_k\}$, $k \geq 2$, we apply nucleus sampling [134] to predict multiple frames for an argument. Specifically, given a cumulative probability mass threshold τ, we assign the minimal subset of frames $F \subseteq \mathcal{F}$ such that:

$$\sum_{f \in F} p_f \geq \tau$$

When the model is very confident in predicting one frame, it is hence likely that an argument is classified to that frame. In cases where the model has lower confidence in its prediction, the argument may consist of multiple frames. This overcomes the limitation of clustering-based approaches and classifiers which strictly assign a single frame to arguments that may contain multiple ones [131, 250].

To train SuperFrame, we used the Media Frames Corpus by Card et al. [55] consisting of 14,515 news articles with text spans manually annotated for the frame classes in Table 7.1. Following Heinisch and Cimiano [131], we trained two variants of the classifier, a *single-task* and a *multi-task* classifier which additionally used the framing dataset by Ajjour et al. [1] with 12,326 labeled arguments. Both models were based on BiLSTMs, used GloVe embeddings,[8] and trained up to 12 epochs using early stopping. We truncated the input to 75 words with a batch size of 64. To choose between the *single-task* and *multi-task* variants, three of the authors first manually assigned frame(s) for 150 arguments. We then predicted the frames for these arguments using both variants.[9] We opted for higher precision as our goal is to minimize mislabeling arguments with an unrelated frame that can negatively impact the resulting frame-oriented summaries. Since frame assignment is a subjective task [55] and the boundaries of the frame classes are fuzzy [48, 250], we observed some diversity in our manual annotations. Specifically, we observed that 92% of all the annotated arguments have at least one frame, which was assigned by only a single annotator (minority), indicating different perceptions of observing specific frames in texts. On

[8] `https://nlp.stanford.edu/data/glove.840B.300d.zip`

[9] We also experimented with multiple preprocessing methods (e.g. generating a conclusion or ranking the sentences) before automatically predicting the frames. However, these methods negatively impacted the frame prediction.

average, an argument was assigned 3.8 frames (or 1.3 and 0.4 considering the majority and full agreements, respectively).

Model	Minority	Majority	Full
single-task	59.6 / 49.6	**41.7 / 34.1**	**38.8 / 28.8**
single-$\tau = .8$	55.0 / 45.5	32.6 / 27.6	34.8 / 27.6
single-$\tau = .9$	**60.5 / 55.4**	27.8 / 24.5	30.4 / 23.7
multi-task	52.4 / 50.1	27.9 / 22.7	38.4 / 29.5
multi-$\tau = .8$	56.4 / 55.0	33.0 / 26.6	27.4 / 20.1
multi-$\tau = .9$	51.0 / 46.9	26.7 / 21.7	25.4 / 17.9

Table 6.2: Precision scores (micro / macro %) of the SuperFrame model variants at different annotator agreements and thresholds τ for multi-frame prediction.

Table 6.2 presents the precision scores of both variants with cumulative probability threshold $\tau = 0.9$. Assigning only the most probable frame as predicted by the *single-task* model results in a precision of 59.6% (micro-average) and 49.6% (macro-average), respectively. The *multi-task* model is slightly better at predicting rare frame classes (+0.5% macro-average) but worse at predicting the frequent ones (-7.2% micro-average). Assigning multiple frames per argument increases the effectiveness of the *single-task* model by +0.9% (micro-average), and especially the prediction of rare frame classes, increasing the macro-average prevision by +5.8% (at $\tau = 0.9$).

Considering only the majority-labeled frame classes as ground truth restricts the set of manually assigned frame classes, and hence, reduces the precision scores. On this restricted subset of frame labels, the *single-task* model performs best in nearly all cases, by predicting only the most probable frame class due to the sparsity of the manually assigned frame classes. This variant of the *single-task* model which predicts only a single frame for an argument has a micro-averaged precision of 41.7% and 38.8% in the majority and full agreement scenarios, respectively. Despite this, we extended the *single-task* variant to predict multiple frames per argument, resulting in a high overlap with ground truth frame labels from at least one annotator as well as benefiting from a higher recall. This also avoids having sparse sets of arguments assigned under rare frames.

In conclusion, our internal evaluation supports using the *single-task* model, as opposed to the findings of Heinisch and Cimiano [131] due to our emphasis on precision while the *multi-task* variant primarily encourages the model in its recall-generalization ability. On average, SuperFrame

(*single-task* variant) assigned 2.6 frames per argument, with a minimum of 1 and a maximum of 8.2. The frequency counts of all frames in both posts and arguments are shown in Appendix Table A.2.

6.4 Evaluation of Extractive Summaries via Relevance Judgments

Given that our entire approach is based on retrieval models, we evaluated it manually via relevance judgments. We followed the evaluation style of TREC [127] as best practice. Our evaluation was comprised of judging the *frame relevance*, the *topic relevance*, and the *importance* (in the discussion's context) of arguments retrieved by our models. Following the TREC protocol, we first created 50 evaluation topics, each comprising a post's title, the post itself, and a frame of interest (see supplementary material). To obtain a sufficiently large set of arguments to pool from, we then selected only those discussions for which all models assigned at least 20 arguments to each of the five most frequent frames identified in the comments: *cultural identity*, *economic*, *quality of life*, *public opinion*, and *political* (see Table A.2 in the Appendix for the full list). We retrieved arguments for each evaluation topic and performed pooling at depth 5 using TrecTools [224], resulting in 1871 unique arguments to be judged.

6.4.1 Pilot Study

Multi-annotator relevance judgments can often result in low agreement due to the subjective nature of defining *relevance* and the varying perspectives of annotators [19, 199, 296, 314]. Additionally, judges may experience inconsistencies in their decisions as the task progresses [267]. To mitigate these issues, we conducted a pilot study with 100 arguments (not included in the main evaluation) to train three annotators and gather feedback for improving the main evaluation interface. The annotators were Computer Science graduates with backgrounds in NLP and IR.

Task Design Following McDonnell et al. [199], we used a four-point scale for assessing the frame and topic relevance, and the importance of an argument with these options: *definitely not*, *probably not*, *probably*, and *definitely* relevant/important.[10] In assessing importance, we asked annotators to indicate the relevance of an argument to a discussion by answering this question: "How important is the argument to be included in a *summary* of the

[10]We mapped these labels to numerical values ranging from 0 (*definitely not* relevant/important) to 3 (*definitely* relevant/important) for computing nDCG scores.

discussion?". We also experimented with an automatic summary [210] for long arguments to reduce the cognitive load of the annotators. They were instructed to use the summary if they found it helpful, otherwise to read the entire argument (for details, see Appendix A.2, Figure A.1).

Pilot Agreement and Feedback We measured the inter-annotator agreement (IAA) for the three evaluated criteria using Krippendorff's α, similar to Card et al. [55]. The resulting α values were 0.22 for frame relevance, 0.33 for topic relevance, and 0.22 for importance, respectively. While the agreement is thus limited, the values are consistent with the findings of Card et al. [55] in their annotation of frame-relevant text spans for the Media Frames Corpus, particularly the frame relevance α value. From feedback, we improved the task design for the main evaluation. Firstly, we removed the automatic summary for each argument since it did not provide significant help. Secondly, we rephrased the importance question to "How important is the argument to be included in the *discussion* of the given topic?" to make it more straightforward, since we did not have ground truth summaries of the discussions at hand. Annotators also reported that assessing the relevance of an argument for a *single* frame was too restrictive, since an argument may belong to multiple frames, which aligns with the observations of Card et al. [55]. Therefore, we allowed them to assign multiple frames to an argument if the currently-assigned one was not relevant. Accordingly, we proceeded with the main evaluation by assigning each annotator an independent set of arguments to judge. This allowed us to collect more relevance judgments while ensuring a certain level of *shared* understanding of the task.

6.4.2 Main Evaluation Results

The evaluated models are shown in Table 6.3.[11] We obtained relevance judgments for a total of 1871 arguments and calculated nDCG@5 [144] as the effectiveness measure (mean over all topics). Described below are the key findings for each module of our ranking-based extractive summarization framework.

Frame Relevance Our frame assignment approach (*IRFr* with BM25) outperforms other models for identifying frame-relevant arguments in a discussion with an nDCG@5 of 0.573. Among the retrieval models, BM25 per-

[11] Model names in Table 6.3 shortened for brevity. SuperFrame → *SupFr* denotes the baseline, IRFrame → *IRFr* denotes our frame assignment approach, Argument Re-ranking → *_rr* (via overlap and retrieval models), and Post-processing → *_post*

forms better than SBERT and ColBERT, also for re-ranking by topic relevance. Upon further inspection, we found that BM25 often retrieves longer arguments compared to the embedding-based SBERT and ColBERT models. This may provide annotators with more context for informed judgments compared to the shorter arguments. Given the computational costs of running dense retrieval models in real-time, it is promising that a relatively simple and explainable model performs well on our query variants. For the baseline (*SupFr*), combinations with argument re-ranking (via BM25 and topic overlap) also perform reasonably well. However, as various query variants can be easily designed, our *IRFr* approach is more flexible and can be adapted to other domains and topics without the need for labeled data.

Topic Relevance Argument re-ranking by overlap (**_rr_overlap*) outperforms retrieval models for ensuring topic relevance of a frame's arguments. This benefits both *IRFr* and *SupFr* frame assignment approaches with an nDCG@5 scores of 0.847 and 0.785 for the top two models, respectively. Among the retrieval models, BM25 slightly outperforms ColBERT. Given the intuitive nature of content overlap, we conclude that it is favorable to use for re-ranking arguments in a frame.

Importance None of the post-processed models (using informativeness) appear in the top-5 for ranking arguments by importance in the context of the discussion. Instead, argument re-ranking by topic relevance performs best, with nDCG@5 of 0.381 combined with *SupFr* for frame assignment. This contradicts our intuition of post-processing to promote important arguments in the final ranking. As future work, we plan to investigate using context features of the arguments [154], as well as pairwise judgments for importance [187, 348].

6.5 Summary

This chapter transitioned from single document summarization, specifically argumentative texts, to multi-document summarization, encompassing entire online discussions. We introduced a novel, ranking-based approach for frame-oriented (extractive) discussion summarization in web-based forums, aiming to enhance the accessibility and comprehension of large-scale online discussions for participants. Our approach involves three key steps: frame assignment, argument re-ranking, and post-processing. Specifically,

Model	nDCG@5		
	Frame	Topic	Imp.
Our Approach			
IRFr_BM25	**0.573**[1]	0.708	0.375[2]
IRFr_SBERT	0.480	0.525	0.303
IRFr_ColBERT	0.522	0.659	0.361[3]
IRFr_BM25_rr_BM25	0.516	0.781[3]	0.349
IRFr_BM25_rr_overlap	0.560[2]	**0.847**[1]	0.350[5]
IRFr_BM25_rr_ColBERT	0.540[4]	0.761	0.358[4]
IRFr_BM25_rr_BM25_post	0.489	0.735	0.297
IRFr_BM25_rr_overlap_post	0.522	0.755	0.339
IRFr_BM25_rr_ColBERT_post	0.526	0.719	0.325
Supervised Baseline			
SupFr_rr_BM25	0.545[3]	0.765[4]	**0.381**[1]
SupFr_rr_overlap	0.536[5]	0.785[2]	0.334
SupFr_rr_ColBERT	0.529	0.764[5]	0.348
SupFr_rr_BM25_post	0.493	0.714	0.322
SupFr_rr_overlap_post	0.493	0.734	0.348
SupFr_rr_ColBERT_post	0.487	0.709	0.329

Table 6.3: nDCG@5 for the manual relevance judgments for frame relevance, topic relevance, and importance. The best results for each evaluated criterion are highlighted in bold, alongside the rankings for the five best models. We evaluated our frame assignment approach (*IRFr*) against the supervised baseline (*SupFr*), combined with our argument re-ranking (*_rr*) and post-processing components (*_post*). We see that our approach to frame assignment results in the best models for frame and topic relevance and is also competitive for argument importance.

we developed unsupervised methods for both frame and topic assignment leveraging standard retrieval models. Extensive experiments on a dataset of 1871 arguments from 100 ChangeMyView discussions demonstrate the effectiveness of our approach in ensuring frame and topic relevance in the summary, outperforming a state-of-the-art supervised baseline for frame assignment. Nevertheless, further exploration is needed to enhance summary informativeness through post-processing.

7

Indicative Summarization of Long Discussions

The previous chapter illustrated the utility of using argumentation frames to categorize a discussion's arguments, thereby generating insightful summaries for each frame. In this chapter, we delve into the related aspect of creating *indicative* summaries for lengthy discussions, which can serve as a navigational guide for exploring such discussions. When discussions comprise hundreds of arguments, pinpointing the optimal juncture to introduce new arguments can be a challenging endeavor. Moreover, it could be advantageous to classify these arguments (or their conclusions) based on predefined categories, like argumentation frames. These frames encapsulate various perspectives presented by the participants, thereby enriching the understanding of the discussion.

Our method generates a table-of-contents for a discussion, providing two levels of information: the first level outlines the argumentation frames, while the second level details the subtopics (abstractive summaries of groups of arguments that talk about a frame). This summarization process is completely unsupervised and leverages large language models (LLMs) to summarize argument clusters and assign argumentation frames to each summary. We assess our approach on discussions from the ChangeMyView subreddit through an interactive user study. The results indicate that users prefer our summaries over alternative views when exploring lengthy discussions. An interactive tool (Discussion Explorer) leveraging our approach for easily exploring long discussions is available at `https://discussion-explorer.web.webis.de/`.

7.1 Table of Contents as an Indicative Summary

Online discussion forums are a popular medium for discussing a wide range of topics. As the size of a community grows, so does the length of discussions held there, especially when current controversial topics are discussed. On ChangeMyView (CMV),[1] for example, discussions often often go into the hundreds of arguments covering many perspectives on the topics in question. Initiating, participating in, or reading discussions generally has two goals: to learn more about others' views on a topic and/or to share one's own.

To help their users navigate large volumes of arguments in long discussions, many forums offer basic features to sort them, for example, by time of creation or popularity. However, these alternative views may not capture the full range of perspectives exchanged, so it is still necessary to read most of them for a comprehensive overview. In this paper, we depart from previous approaches to summarizing long discussions by using *indicative* summaries instead of *informative* summaries.[2] Figure 7.1 illustrates our three-step approach: first, the sentences of the arguments are clustered according to their latent subtopics. Then, a large language model generates a concise abstractive summary for each cluster as its label. Finally, the argumentation frame [40, 63] of each cluster label is predicted as a generalizable operationalization of perspectives on a discussion's topic. From this, a hierarchical summary is created in the style of a table of contents, where frames act as headings and cluster labels as subheadings. To our knowledge, indicative summaries of this type have not been explored before.

Our four main contributions are: (1) A fully unsupervised approach to indicative summarization of long discussions (Section 7.2). We develop robust prompts for generative cluster labeling and frame assignment based on extensive empirical evaluation and best practices (Section 7.3). (2) A comprehensive evaluation of 19 state-of-the-art, prompt-based, large language models (LLMs) for both tasks, supported by quantitative and qualitative assessments (Section 7.4). (3) A user study of the usefulness of indicative summaries for exploring long discussions (Section 7.4). (4) Discussion Explorer, an interactive visual interface for exploring the indicative summaries generated by our approach and the corresponding discus-

[1] `https://www.reddit.com/r/changemyview/`

[2] Unlike an informative summary, an indicative summary does not capture as much information as possible from a text, but only its gist. This makes them particularly suitable for long documents like books in the form of tables of contents.

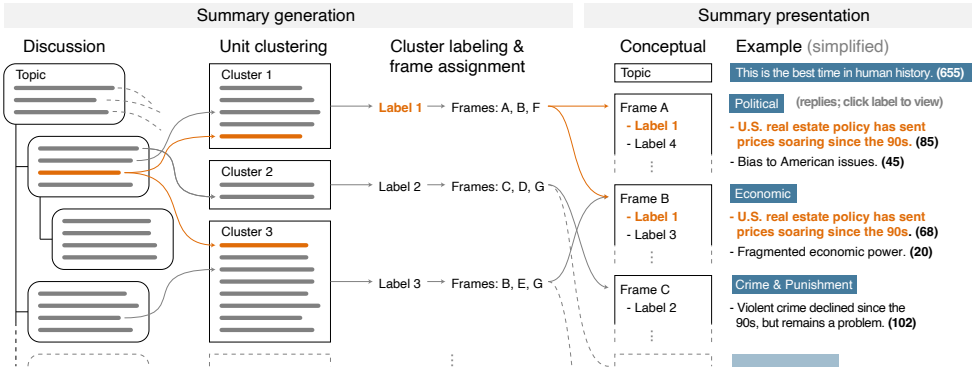


Figure 7.1: Left: Illustration of our approach to generating indicative summaries for long discussions. The main steps are (1) unit clustering, (2) generative cluster labeling, and (3) multi-label frame assignment in order of relevance. Right: Conceptual and exemplary presentation of our indicative summary in table of contents style. Frames act as headings and the corresponding cluster labels as subheadings.

sions.[3] Our results show that the GPT variants of OpenAI (GPT3.5, ChatGPT, and GPT4) outperform all other open source models at the time of writing. LLaMA and T0 perform well, but are not competitive with the GPT models. Regarding the usefulness of the summaries, users preferred our summaries to alternative views to explore long discussions with hundreds of arguments.

Related Work

Previous approaches to generating discussion summaries have mainly focused on generating extractive summaries, using two main strategies: extracting significant units (e.g., responses, paragraphs, or sentences), or grouping them into specific categories, which are then summarized. In this section, we review the relevant literature.

7.1.1 Extractive Summarization

Extractive approaches use supervised learning or domain-specific heuristics to extract important entities from discussions as extractive summaries. For example, Klaas [164] summarized UseNet newsgroup threads by considering thread structure and lexical features to measure message importance. Tigelaar et al. [297] identified key sentences based on author names and citations, focusing on coherence and coverage in summaries. Ren et al. [251] developed a hierarchical Bayesian model for tracking topics, using a

[3] `https://discussion-explorer.web.webis.de/`

random walk algorithm to select representative sentences. Ranade et al. [247] extracted topic-relevant and emotive sentences, while Bhatia et al. [30] and Tarnpradab et al. [293] used dialogue acts to summarize question-answering forum discussions. Egan et al. [90] extracted key points using dependency parse graphs, and Kano et al. [154] summarized Reddit discussions using local and global context features. These approaches generate informative summaries, substituting discussions without back-referencing to them.

7.1.2 Grouping-based Summarization

Grouping-based approaches group discussion units like posts or sentences, either implicitly or explicitly. The groups are based on queries, aspects, topics, dialogue acts, argument facets, or key points annotated by experts. Once the units are grouped, individual summaries are generated for each group by selecting representative members, respectively.

This *grouping-then-summarization* paradigm has been primarily applied to multi-document summarization of news articles [243]. Follow-up work proposed cluster link analysis [322], cluster sentence ranking [51], and density peak identification in clusters [342]. For abstractive multi-document summarization, Nayeem et al. [211] clustered sentence embeddings using a hierarchical agglomerative algorithm, identifying representative sentences from each cluster using TextRank [201] on the induced sentence graph. Similarly, Fuad et al. [101] clustered sentence embeddings and selected subsets of clusters based on importance, coverage, and variety. These subsets are then input to a transformer model trained on the CNN/DailyMail dataset [206] to generate a summary. Recently, Ernst et al. [95] used agglomerative clustering of salient statements to summarize sets of news articles, involving a supervised ranking of clusters by importance.

For Wikipedia discussions, Zhang et al. [337] proposed the creation of a dynamic summary tree to ease subtopic navigation at different levels of detail, requiring editors to manually summarize each tree node's cluster. Misra et al. [203] used summarization to identify arguments with similar aspects in dialogues from the Internet Argument Corpus [320]. Similarly, Reimers et al. [250] used agglomerative clustering of contextual embeddings and aspects to group sentence-level arguments. Bar-Haim et al. [22, 23] examined the mapping of debate arguments to key points written by experts to serve as summaries.

Our approach clusters discussion units, but instead of a supervised selection of key cluster members, we use vanilla LLMs for abstractive sum-

marization. Moreover, our summaries are hierarchical, using issue-generic frames as headings [40, 63] and generating concise abstractive summaries of corresponding clusters as subheadings. Thus our approach is unsupervised, facilitating a scalable and generalizable summarization of discussions.

7.1.3 Cluster Labeling

Cluster labeling involves assigning representative labels to document clusters to facilitate clustering exploration. Labeling approaches include comparing term distributions [193], selecting key terms closest to the cluster centroid [254], formulating key queries [112], identify keywords through hypernym relationships [233], and weak supervision to generate topic labels Popa and Rebedea [234]. These approaches often select a small set of terms as labels that do not describe a cluster's contents in closed form. Our approach overcomes this limitation by treating cluster labeling as a zero-shot abstractive summarization task.

7.1.4 Frame Assignment

Framing involves emphasizing certain aspects of a topic for various purposes, such as persuasion [63, 94]. Frame analysis for discussions provides insights into different perspectives on a topic [182, 204]. It also helps to identify biases in discussions resulting, e.g., from word choice [122, 123]. Thus, frames can serve as valuable reference points for organizing long discussions. We use a predefined inventory of media frames [40] for discussion summarization. Instead of supervised frame assignment [1, 131, 205], we use prompt-based LLMs for more flexibility.

7.2 Unsupervised Summarization with Large Language Models

Our indicative summarization approach takes the sentences of a discussion as input and generates a summary in the form of a table of contents, as shown in Figure 7.1. Its three steps consist of clustering discussion sentences, cluster labeling, and frame assignment to cluster labels.

7.2.1 Unit Clustering

Given a discussion, we extract its sentences as discussion units. The set of sentences is then clustered using the density-based hierarchical clustering algorithm HDBSCAN [52]. Each sentence is embedded using SBERT [249]

and these embeddings are then mapped to a lower dimensionality using UMAP [200].[4] Unlike previous approaches that rank and filter clusters to generate informative summaries [95, 290], our summaries incorporate all clusters. The sentences of each cluster are ranked by centrality, which is determined by the λ value of HDBSCAN. A number of central sentences per cluster are selected as input for cluster labeling by abstractive summarization.

Meta-sentence filtering Some sentences in a discussion do not contribute directly to the topic, but reflect the interaction between its participants. Examples include sentences such as "I agree with you." or "You are setting up a straw man." Pilot experiments have shown that such meta-sentences may cause our summarization approach to include them in the final summary. As these are irrelevant to our goal, we apply a corpus-specific and channel-specific meta-sentence filtering approach, respectively. Corpus-specific filtering is based on a small set of frequently used meta-sentences M in a large corpus (e.g., on Reddit). It is bootstrapped during preprocessing, and all sentences in it are omitted by default.[5]

Our pilot experiments revealed that some sentences in discussions are also channel-specific (e.g., for the ChangeMyView Subreddit). Therefore, we augment our sentence clustering approach by adding a random sample $M' \subset M$ to the set of sentences D of each individual discussion before clustering, where $|M'| = \max\{300, |D|\}$. The maximum value for the number of meta-sentences $|M'|$ is chosen empirically, to maximize the likelihood that channel-specific meta-sentences are clustered with corpus-specific ones. After clustering the joint set of meta-sentences and discussion sentences $D \cup M'$, we obtain the clustering $\mathcal{C}$. Let $m_C = |C \cap M'|$ denote the number of meta-sentences and $d_C = |C \cap D|$ the number of discussion sentences in a cluster $C \in \mathcal{C}$. The proportion of meta-sentences in a cluster is then estimated as $P(M'|C) = \frac{m_C}{m_C + d_C}$.

A cluster C is classified as a meta-sentence cluster if $P(M'|C) > \theta \cdot P(M')$, where $P(M') = \frac{|M'|}{|D|}$ assumes that meta-sentences are independent of others in a discussion. The noise threshold $\theta = \frac{2}{3}$ was chosen empirically. Sentences in a discussion that either belong to a meta-sentence cluster or whose nearest cluster is considered to be one are omitted. In our evaluation, an average of 23% of sentences are filtered from discussions. Figure 7.2 illustrates the effect of meta-sentence filtering on a discussion's set of sentence.

[4] Implementation details are given in Appendix A.4.

[5] The set is used like a typical stop word list, only for sentences.

7.2.2 Generative Cluster Labeling

Most cluster labeling approaches extract keywords or key phrases as labels, which limits their fluency. These approaches may also require training data acquisition for supervised learning. We formulate cluster labeling as an unsupervised abstractive summarization task. We experiment with prompt-based large language models in zero-shot and few-shot settings. This enables generalization across multiple domains, the elimination of supervised learning, and fluent cluster labels with higher readability in comparison to keywords or phrases.

We develop several prompt templates specifically tailored for different types of LLMs. For encoder-decoder models, we carefully develop appropriate prompts based on PromptSource [17], a toolkit that provides a comprehensive collection of natural language prompts for various tasks across 180 datasets. In particular, we analyze prompts for text summarization datasets with respect to (1) descriptive words for the generation of cluster labels using abstractive summarization, (2) commonly used separators to distinguish instructions from context, (3) the position of instructions within prompts, and (4) the granularity level of input data (full text, document title, or sentence). Since our task is about summarizing groups of sentences, we chose prompts that require the full text as input to ensure that enough contextual information is provided (within the limits of each model's input size). Section 7.3.1 provides details on the prompt engineering process.

7.2.3 Frame Assignment

Any controversial topic can be discussed from different perspectives. For example, "the dangers of social media" can be discussed from a moral or a health perspective, among others. In our indicative summaries, we use argumentation frame labels as top-level headings to operationalize different perspectives. An argumentation frame may include one or more groups of relevant arguments. We assign frame labels from the issue-generic frame inventory shown in Table 7.1 [40] to each cluster label derived in the previous step.[6]

We use prompt-based models in both zero-shot and few-shot settings for frame assignment. In our experiments with instruction-tuned models, we designed two types of instructions, shown in Figure A.8, namely direct instructions for models trained on instruction–response samples, and dialog instructions for chat models. The instructions are included along with the

[6]For detailed label descriptions see Table A.1 in the Appendix.

Frame Inventory	
Capacity & Resources	Fairness & Equality
Constitutionality & Jurisprudence	Health & Safety
Crime & Punishment	Morality
Cultural Identity	Policy Prescription & Evaluation
Economic	Political
External Regulation & Reputation	Public Opinion
	Quality of Life
	Security & Defense

Table 7.1: Inventory of frames proposed by Boydstun et al. [40] to track the media's framing of policy issues.

cluster labels in the prompts. Moreover, including the citation of the frame inventory used in our experiments has a positive effect on the effectiveness of some models (see Appendix A.6.1 for details).

7.2.4 Indicative Summary Presentation

Given the generated labels of all sentence clusters and the frame labels assigned to each cluster label, our indicative summary groups the cluster labels by their respective frame labels. The cluster label groups of each frame label are then ordered by cluster size. This results in a two-level indicative summary, as shown in Figures 7.1 and 7.4.

7.3 Comprehensive Analysis of Prompt Engineering

Using prompt-based LLMs for generative cluster labeling and frame assignment requires model-specific prompt engineering as a preliminary step. We explored the 19 model variants listed in Table 7.2. To select the most appropriate models for our task, we consulted the HELM benchmark [175], which compares the effectiveness of different LLMs for different tasks. Further, we have included various recently released open source models (with optimized instructions) as they were released. Since many of them were released during our research, we reuse prompts previously optimized prompts for the newer models.[7]

[7] See Appendices A.5 and A.6 for details.

7.3.1 Cluster Labeling

The prompts for the encoder-decoder model T0 are based on the PromptSource [17] toolkit. We have experimented with different prompt templates and tried different combinations of input types (e.g. "text", "debate", "discussion", and "dialogue") and output types (e.g. "title", "topic", "summary", "theme", and "thesis"). The position of the instruction within a prompt was also varied, taking into account prefix and suffix positions. For decoder-only models like BLOOM, GPT-NeoX, OPT-66B, and OPT, we experimented with hand-crafted prompts. For GPT3.5, we followed the best practices described in OpenAI's API and created a single prompt.

Prompts were evaluated on a manually annotated set of 300 cluster labels using BERTScore [341]. We selected the most effective prompt for each of the above models for cluster labeling. Our evaluation in Section 7.4 shows that GPT3.5 performs best in this task. Figure 7.3 (top) shows the best prompt for this model.[8]

7.3.2 Frame Assignment

For frame assignment, models were prompted to predict a maximum of three frame labels for a given cluster label, ordered by relevance. Experiments were conducted with both direct instructions and dialogue prompts in zero-shot and few-shot settings. In the zero-shot setting, we formulated three prompts containing (1) only frame labels, (2) frame labels with short descriptions, and (3) frame labels with full text descriptions (see Appendix A.6.2 for details). For the few-shot setting, we manually annotated up to two frames from the frame inventory of Table 7.1 for each of the 300 cluster labels generated by the best model GPT3.5 in the previous step. We included 42 examples (3 per frame) in the few-shot prompt containing the frame label, its full-text description, and three examples. The remaining 285 examples were used for subsequent frame assignment evaluation. Our evaluation in Section 7.4 shows that GPT4 performs best on this task. Figure 7.3 (bottom) shows its best prompt.

7.4 Purpose-driven Evaluation of Summary Usefulness

To evaluate our approach, we conducted automatic and manual evaluations focused on the cluster labeling quality and the frame assignment accuracy. We also evaluated the utility of our indicative summaries in a purpose-

[8] ChatGPT and GPT4 were released after our evaluation.

driven user study in which participants had the opportunity to explore long discussions and provide us with feedback.

7.4.1 Data and Preprocessing

We used the "Winning Arguments" corpus from Tan et al. [291] as a data source for long discussions. It contains 25,043 discussions from the ChangeMyView Subreddit that took place between 2013 and 2016. The corpus was preprocessed by first removing noise replies and then meta-sentences. Noise replies are marked in the metadata of the corpus as "deleted" by their respective authors, posted by bots, or removed by moderators. In addition, replies that referred to the Reddit guidelines or forum-specific moderation were removed using pattern matching (see Appendix A.3 for details). The remaining replies were split into a set of sentences using Spacy [135]. To enable the unit clustering (of sentences) as described in Section 7.2.1, the set of meta-sentences M is bootstrapped by first clustering the entire set of sentences from all discussions in the corpus and then manually examining the clusters to identify those that contain meta-sentences, resulting in $|M| = 955$ meta-sentences. After filtering out channel-specific noise, the (cleaned) sets of discussion sentences are clustered as described.

Evaluation Data From the preprocessed discussions, 300 sentence clusters were randomly selected. Then, we manually created a cluster label and up to three frame labels for each cluster. Due to the short length of the cluster labels, up to two frames per label were sufficient. After excluding 57 examples with ambiguous frame assignments, we obtained a reference set of 243 cluster label samples, each labeled with up to two frames.

7.4.2 Generative Cluster Labeling

The results of the automatic cluster labeling evaluation using BERTScore and ROUGE are shown in (Appendix) Tables A.4 and A.5, respectively. We find that ChatGPT performs best. To manually evaluate the quality of the cluster labels, we used a ranking-based method in which four annotators scored the generated cluster labels against the manually annotated reference labels of each of the 300 clusters. To provide additional context for the cluster content, the five most semantically similar sentences to the reference label from each cluster were included, as well as five randomly selected sentences from the cluster. To avoid possible bias due to the length of the clus-

ter labels by different models, longer labels were truncated to 15 tokens.[9] To determine an annotator's model ranking, we merged the preference rankings for all clusters using reciprocal rank fusion [70]. Annotator agreement was calculated using Kendall's W for rank correlation [159], which yielded a value of 0.66, indicating *substantial* agreement.

The average ranking of each model is shown in Table 7.3 along with the length distributions of the generated cluster labels.[10] GPT3.5 showed superior effectiveness in generating high-quality cluster labels. It ranked first in 225 out of 300 comparisons, with an average score of 1.38 by the four annotators. The cluster labels generated by GPT3.5 were longer on average (9.4 tokens) and thus more informative than those generated by the other models, which often generated disjointed or incomplete labels. In particular, T0 generated very short labels on average (3.1 tokens) that were generic/non-descriptive.

7.4.3 Frame Assignment

In the zero-/few-shot frame assignment settings described in Section 7.3.2, we prompted the models to predict three frames per cluster label in order of relevance. Using the manually annotated reference set of 243 cluster labels and their frame labels, we evaluated the accuracy of the frames predicted for each cluster label that matched the reference frames. The results for the first predicted frame are shown in Table 7.4. In most cases, GPT4 outperforms all other models in the various settings, with the exception of the zero-shot setting with a short prompt, where GPT3.5 narrowly outperforms GPT4 with 60.9% accuracy versus 60.5%. Among the top five models, the GPT* models that follow direct user instructions perform consistently well, with the LLaMA-/65B/-CoT and T0 models showing strong effectiveness among the open-source LLMs. Conversely, the OPT model performs consistently worse in all settings. The few-shot setting shows greater variance in results, suggesting that the models are more sensitive to the labeled examples provided in the prompts. Including a citation to the frame inventory paper in the instructions (see Figure A.8) significantly improved the effectiveness of Falcon-40B (12%) and LLaMA-65B (9%) in the zero-shot setting (see Appendix A.6.1 for details).

[9] Figure A.3 in the Appendix shows the annotation interfaces.

[10] As newer models were published after our manual evaluation, we show an automatic evaluation of all models using human and GPT3.5-based reference labels in the Appendix in Tables A.4 and A.5.

7.4.4 Usefulness Evaluation

In addition to assessing each step of our approach, we conducted a user study to evaluate the effectiveness of the resulting indicative summaries. In this study, we considered two key tasks: *exploration* and *participation*. With respect to *exploration*, our goal was to evaluate the extent to which the summaries help users explore the discussion and discover new perspectives. With respect to *participation*, we wanted to assess how effectively the summaries enabled users to contribute new arguments by identifying the appropriate context and location for a response.

We asked five annotators to explore five randomly selected discussions from our dataset, for which we generated indicative summaries using our approach with GPT3.5. To facilitate intuitive exploration, we developed Discussion Explorer (see Section 7.4.5), an interactive visual interface for the evaluated discussions and their indicative summaries. In addition to our summaries, two baselines were provided to annotators for comparison: (1) the original web page of the discussion on ChangeMyView, and (2) a search engine interface powered by Spacerini [3]. The search engine indexed the sentences within a discussion using the BM25 retrieval model. This allowed users to explore interesting perspectives by selecting self-selected keywords as queries, as opposed to the frame and cluster labels that our summaries provide. Annotators selected the best of these interfaces for exploration and participation.

Results With respect to the *exploration* task, the five annotators agreed that our summaries outperformed the two baselines in terms of discovering arguments from different perspectives presented by participants. The inclusion of argumentation frames proved to be a valuable tool for the annotators, facilitating the rapid identification of different perspectives and the accompanying cluster labels showing the relevant subtopics in the discussion. For the *participation* task, three annotators preferred the original web page, while our summaries and the search engine were preferred by the remaining two annotators (one each) when it came to identifying the appropriate place in the discussion to put their arguments. In a post-study questionnaire, the annotators revealed that the original web page was preferred because of its better display of the various response threads, a feature not comparably reimplemented in Discussion Explorer. The original web page felt "more familiar." However, we anticipate that this limitation can be addressed by seamlessly integrating our indicative summaries into a given

discussion forum's web page, creating a consistent experience and a comprehensive and effective user interface for discussion participation.

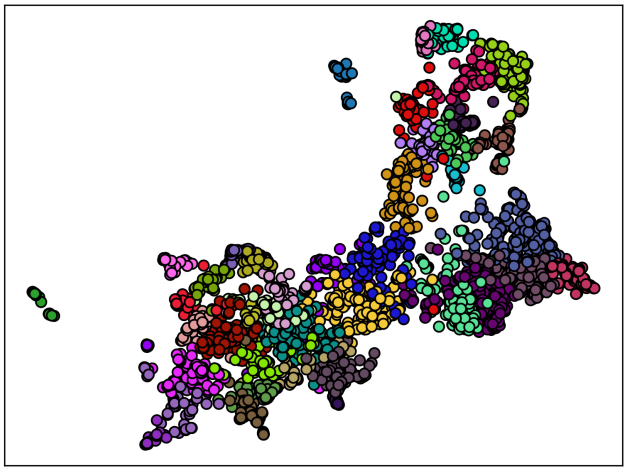
7.4.5 Discussion Explorer

Our approach places emphasis on summary presentation by structuring indicative summaries into a table of contents for discussions (see Section 7.2). To demonstrate the effectiveness of this presentation style in exploring long discussions, we have developed an interactive tool called Discussion Explorer.[11] This tool illustrates how such summaries can be practically applied. Users can participate in discussions by selecting argumentation frames or cluster labels. Figure 7.4 presents indicative summaries generated by different models, providing a quick overview of the different perspectives. This two-level table of contents-like summary provides effortless navigation, allowing users to switch between viewing all arguments in a frame and understanding the context of sentences in a cluster of the discussion (see Figure A.4).

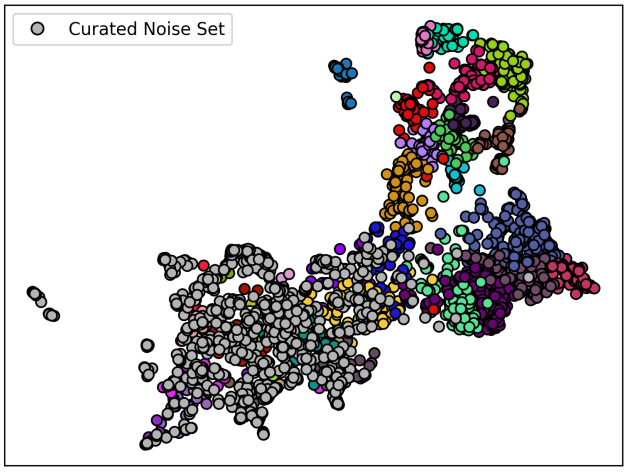
7.5 Summary

This chapter presented an unsupervised approach for generating indicative summaries of long discussions to facilitate their exploration and navigation. Our summaries resemble tables of contents, which list argumentation frames and concise abstractive summaries of the latent subtopics for a comprehensive overview of a discussion. By analyzing 19 prompt-based LLMs, we found that GPT3.5 and GPT4 perform impressively, with LLaMA fine-tuned using chain-of-thought being the second best. A user study of long discussions showed that our summaries were valuable for exploring and uncovering new perspectives in long discussions, an otherwise tedious task when relying solely on the original web pages. Finally, we presented Discussion Explorer, an interactive visual tool designed to navigate through long discussions using the generated indicative summaries. This demonstrates how indicative summaries can be used effectively in practical scenarios.

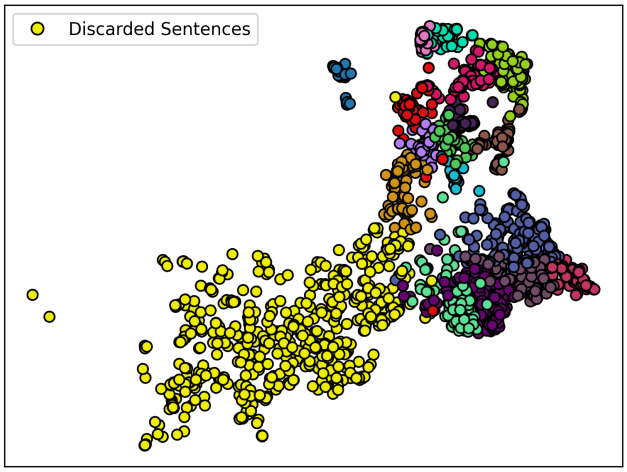
[11] `https://discussion-explorer.web.webis.de/`



(a) Joint clustering of a discussion and meta-sentences $D \cup M'$.



(b) The sampled meta-sentences $M' \subset M$ highlighted gray.



(c) Classification of meta-sentence clusters to be omitted.

Figure 7.2: Effect of meta-sentence filtering: (a and b) A discussion's sentences D are jointly clustered with a sample of meta-sentences $M' \subset M$. (c) Then each cluster is classified as a meta-sentence cluster based on its proportion of meta-sentences and its neighboring clusters. Meta-sentence clusters are omitted.

Model Variants	Description
	Pre-InstructGPT
T0 vanilla	Encoder-decoder model trained on datasets transformed as task-specific prompts.
BLOOM vanilla	A multlingual autoregressive model with 176B parameters for prompt-based text completion.
GPT-NeoX 20B	Open source alternative to GPT-3.
OPT 66B	Autoregressive model with similar effectiveness to GPT-3, but more efficient data collection and training.
	Direct Instruction
LLaMA-CoT vanilla	LLaMA-30B fine-tuned on chain-of-thought and reasoning samples [238].
Alpaca 7B	LLaMA-7B fine-tuned based on 52k self-instruct responses [326].
OASST vanilla	LLaMA-30B fine-tuned on the OpenAssistant Conversations dataset [165] using reinforcement learning.
Pythia 12B	Suite of LLMs trained on public data to investigate the effects of training and scaling on various model properties.
GPT* 3.5, Chat, 4	OpenAI models GPT3.5 (*text-davinci-003*), ChatGPT (*gpt-3.5-turbo*), and GPT4.
	Dialogue Instruction
LLaMA 30B, 65B	Suite of open-source LLMs from Meta AI trained on public datasets.
Vicuna 7B, 13B	LLaMA models fine-tuned using conversations collected by ShareGPT (`https://sharegpt.com`)
Baize 7B, 13B	Open source chat model trained on 100k dialogues generated by letting ChatGPT (GPT 3.5-turbo) talk to itself.
Falcon 40B, 40B-Instructing	Trained on the RefinedWeb corpus [228], which was obtained by filtering and deduplication of public web data.

Table 7.2: LLMs studied for cluster labeling and frame assignment. Older models are listed by Pre-InstructGPT (prior to GPT3.5) and newer models are listed by their respective prompt types investigated (Direct / Dialogue). See Appendices A.5 and A.6 for details.

GPT3.5 for Generative Cluster Labeling

Generate a single descriptive phrase that describes the following debate in very simple language, without talking about the debate or the author.
Debate: """{*text*}"""

GPT4 for Frame Assignment

The following {*input_type*}[a] contains all available media frames as defined in the work from {*authors*}: {*frames*} For every input, you answer with three of these media frames corresponding to that input, in order of importance.

[a] A list of frame labels or a JSON with frame labels and their descriptions.

Figure 7.3: Best performing instructions for cluster labeling and frame assignment. For frame assignment, providing the citation for the frame inventory via the placeholder {*authors*} positively affects the effectiveness of some models (Appendix A.6.1).

Model	Mean Rank	# First	Length		
			Min	Max	Mean
GPT3.5	**1.38**	**225**	3	27	9.44
BLOOM	2.95	33	1	37	8.13
GPT-NeoX	3.20	20	1	34	7.42
OPT	3.36	12	1	30	8.27
T0	3.72	28	1	18	3.10

Table 7.3: Results of the qualitative evaluation of generative cluster labeling. Shown are (1) the mean rank of a model from four annotators and (2) the number of times a model was ranked first by an annotator. GPT3.5 (*text-davinci-003*) performed better than other models and generated longer labels on average.

Model	Zero-Shot			Few-Shot
	–	*short*	*full*	
Alpaca-7B	39.1	39.5	28.4	20.6
Baize-7B	34.2	34.6	39.1	30.9
Baize-13B	42.4	48.1	42.0	39.5
BLOOM	26.7	31.7	25.5	–
ChatGPT	60.9[2]	58.0[3]	58.8[2]	63.4[2]
Falcon-40B	46.5	46.5	46.1	38.3
Falcon-40B-Inst.	51.4	44.4	32.9	28.4
GPT3.5	53.5[3]	**60.9**[1]	58.0[3]	53.9[4]
GPT4	**63.4**[1]	60.5[2]	**65.4**[1]	**67.1**[1]
GPT-NeoX	19.3	25.1	31.3	31.3
LLaMA-30B	45.7	41.2	39.1	40.7
LLaMA-CoT	46.9	54.3[4]	49.8	57.2[3]
LLaMA-65B	53.1[4]	50.6[5]	39.5	–
OASST	48.6[5]	48.1	53.5[5]	47.7
OPT	16.0	13.2	14.8	–
Pythia	31.7	33.3	30.5	29.6
T0	48.6[5]	54.3[4]	55.6[4]	49.8[5]
Vicuna-7B	28.4	36.2	35.4	20.2
Vicuna-13B	44.0	40.7	42.0	38.3

Table 7.4: Results of an automatic evaluation of 19 LLMs (sorted alphabetically) for frame assignment, indicating the five best models in each setting. Shown are the percentages of samples where the first frame predicted by a model is one of the reference frames. The three zero-shot columns denote the prompt type: frame label only, label with *short* description, and label with *full* description. Model types are also indicated: Pre-InstructGPT, Direct / Dialogue. Missing values are model inferences that exceed our computational resources.

Figure 7.4: Discussion Explorer provides a concise overview of indicative summaries from various models for a given discussion. The summary is organized hierarchically: The argument frames act as heading, while the associated cluster labels act as subheadings, similar to a table of contents. Cluster sizes are also indicated. Clicking on a frame lists all argument sentences in a discussion that assigned to that frame, while clicking on a cluster label shows the associated argument sentences that discuss a subtopic in the context of the discussion (see Figure A.4).

8

Summary Explorer: Visual Analytics for the Qualitative Assessment of the State of the Art in Text Summarization

Through the previous chapters, this thesis has introduced a variety of novel contributions, including datasets and models, aimed at summarizing diverse forms of user-generated discourse such as news editorials, social media posts, argumentative texts, and long discussions. In the following two chapters, we shift our focus to the evaluation and comparison of various state-of-the-art text summarization models developed by the research community over the years. The standard practice of evaluating text summarization is to compare the generated summary with one or more reference summaries to compute content overlap. While automatic metrics offer a repeatable and computationally efficient evaluation method, they often fail to provide a comprehensive assessment of summary quality (as described in Chapter 2). Specifically, they struggle to capture the relationship between the summary and the source document, track the provenance of the summary, address the position bias in supervised models that favor certain parts of a document, identify hallucinations, and evaluate the faithfulness of the summary to the source document. To address these shortcomings, this chapter introduces Summary Explorer a tool that provides novel visualizations for comparing source document and all its summaries, as well as multiple summaries against each other, designed to facilitate an easy, qualitative comparison of several state-of-the-art models.

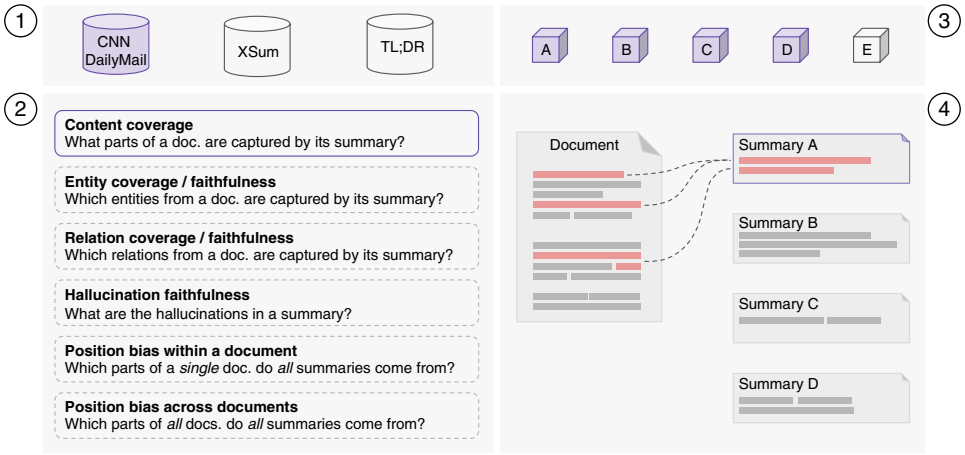


Figure 8.1: Overview of Summary Explorer. Its guided assessment process works in four steps: (1) corpus selection, (2) quality aspect selection, (3) model selection, and (4) quality aspect assessment. Exemplified is the assessment of the content coverage of the summaries of four models for a source document from the CNN/DM corpus. For each summary sentence, its two most related source document sentences are highlighted on demand.

8.1 Limitations of Automatic Evaluation Metrics

Currently, the progress in text summarization is tracked primarily using *automatic evaluation* with ROUGE [176] as the de facto standard for quantitative evaluation. ROUGE has proven effective for evaluating extractive systems, measuring the overlap of word n-grams between a generated summary and a reference summary (ground truth). Still, it only provides an approximation of a model's capability to generate summaries that are lexically similar to the ground truth. Moreover, ROUGE is unsuitable for evaluating abstractive summarization systems, mainly due to its inadequacy in capturing all semantically equivalent variants of the reference [97, 167, 215]. Besides, a reliable automatic evaluation of a summary is challenging [184] and strongly dependent on its purpose[149].

A robust method to analyze the effectiveness of summarization models is to manually inspect their outputs from individual perspectives such as coverage of key concepts and linguistic quality. However, manual inspection requires obtaining the outputs of certain models, delineating a guideline that comprises particular assessment criteria, and ideally utilizing proper visualization techniques to examine the outputs efficiently.

To this end, we present Summary Explorer (Figure 8.1), an online interactive visualization tool that assists humans (researchers, experts, and crowds) to inspect the outputs of text summarization models in a guided

fashion. Specifically, we compile and host the outputs of several state-of-the-art models (currently 55) dedicated to English single-document summarization. These outputs cover three benchmark summarization datasets comprising semi-extractive to highly abstractive ground truth summaries. The tool facilitates a *guided* visual analysis of three important summary quality criteria: *coverage*, *faithfulness*, and *position bias*, where tailored visualizations for each criterion streamline both absolute and relative manual evaluation of summaries. Overall, our use cases (see Section 8.3.3) demonstrate the ability of Summary Explorer to provide a comparative exploration of the state-of-the-art text summarization models, and to discover interesting cases that cannot likely be captured by automatic evaluation.

Related Work

Leaderboards such as Paperswithcode,[1] ExplainaBoard[2] and NLPProgress[3] provide an overview of state of the art in text summarization mainly according to ROUGE. These leaderboards simply aggregate the scores as reported by the models' developers, where the reported scores can be obtained using different implementations. Hence, a fair comparison become less feasible. For instance, the Bottom-Up model [109] uses a different implementation of ROUGE,[4] compared to the BanditSum model [85].[5] Besides, for a qualitative comparison of the models, one needs to manually inspect the generated summaries, which are missing from such leaderboards.

To address these shortcomings, VisSeq [324] aids developers to locally compare their model's outputs with the ground truth, providing lexical and semantic comparisons along with statistics such as most frequent n-grams and sentence score distributions. LIT [294] provides similar functionality for a broader range of NLP tasks, implementing a work-bench-style debugging of model behavior, including visualization of model attention, confusion matrices, and probability distributions. Closely related to our work is SummVis [311], the recently published tool that provides a visual text comparison of summaries with a reference summary as well as a source document, facilitating local debugging of hallucinations in the summaries.

Summary Explorer draws from these developments and adds three missing features: (1) Quality-criteria-driven design. Based on a careful literature review of qualitative evaluation of summaries, we derive three key

[1] https://paperswithcode.com/task/text-summarization
[2] http://explainaboard.nlpedia.ai/leaderboard/task-summ/
[3] https://nlpprogress.com/english/summarization.html
[4] https://github.com/sebastianGehrmann/rouge-baselines
[5] https://github.com/pltrdy/rouge

quality criteria and encode them explicitly in the interface of our tool. Other existing tools render these criteria implicit in their underlying design. (2) A step-by-step process for guided analysis. From the chosen quality criteria, we formulate concise and specific questions needed for a qualitative evaluation, and provide a tailored visualization for each question. While previous tools utilize visualization and enable users to (de)activate certain features, they oblige the users to figure out the process themselves, which can be overwhelming to non-experts. (3) Compilation of the state of the art. We collect the outputs of more than 50 models on three benchmark datasets providing a comprehensive overview of the progress in text summarization. Summary Explorer complements prior tools and also provides direct access to the state of the art in text summarization, encouraging rigorous analysis to support the development of novel models.

8.2 Designing Interfaces for Visual Exploration of Summaries

The design of Summary Explorer derives from first principles, namely the three quality criteria *coverage*, *faithfulness*, and *position bias* of a summary in relation to its source document. These high-level criteria are frequently manually assessed throughout the literature. Since their definitions vary, however, we derive from each criterion a total of six specific aspects that are more straightforwardly operationalized in a visual exploration (see Figure 8.1, Step 2). To render the aspects more directly accessible to users, each is "clarified" by a guiding question that can be answered by a tailored visualization. Below, the three quality criteria are discussed, followed by the visual design.

8.2.1 Summary Quality Criteria

Coverage A primary goal of a summary is to capture the important information from its source document. Accordingly, a standard practice in summary evaluation is to assess its coverage of the key content [149, 190, 223]. In many cases, a comparison to the ground truth (reference) summary can be seen as a proxy for coverage, which is essentially the core idea of ROUGE. However, since it is hard to establish an ideal reference summary [192], a comparison against the source document is more meaningful. Although an automatic comparison against it is feasible [185, 273], deciding what is *important* content is highly subjective [231]. Therefore, authors resort to a manual comparison instead [125]. We operationalize coverage assessment by visualizing a document's overlap in terms of content, entities, and entity

relations with its summary. Content coverage refers to whether a summary condenses information from all important parts of a document, measured by common similarity measures; entity coverage contrasts the sets of named entities identified in both summary and document; and relation coverage does the same, but for extracted entity relations.

Faithfulness A more recent criterion that gained prominence especially in relation to neural summarization is the faithfulness of a summary to its source document [54, 198]. Whereas coverage asks if the document is sufficiently reflected in the summary, faithfulness asks the reverse, namely if the summary adds something new, questioning its appropriateness. Due to their autoregressive nature, neural summarization models have the unique property to "hallucinate" new content [168, 345]. This is what enables abstractive summarization, but also bears the risk of generating content in a summary that is unrelated to the source document. The only acceptable hallucinated content in a summary must be textually entailed by its source document, which renders an automatic assessment challenging [88, 98]. We operationalize faithfulness assessment by visualizing previously unseen words in a summary in context, aligned with the best-matching sentences of its source document.

Position bias Data-driven approaches, such as neural summarization models, can be biased by the domain of their training data and learn to exploit common patterns. For example, news articles are typically structured according to an "inverted pyramid," where the most important information is given in the first few sentences [237], and which models learn to exploit [158, 327]. Non-news texts, such as social media posts, however, do not adopt this structure and thus require an unbiased consideration to obtain proper summaries [284]. We operationalize position bias assessment by visualizing the parts of a document that are the source of its summary's sentences, as well as the ones that are common among a set of summaries.

8.2.2 Visual Design

Guided Assessment Summary Explorer implements a streamlined process to guide summary quality assessment, consisting of four steps (see Figure 8.1). (1) A benchmark dataset is selected. (2) A list of available summary quality aspects is offered each with a preview of its tailored visualization and its interactive use. (3) Applying Shneiderman's (1996) well-known Visual Information-seeking Mantra ("overview first, zoom and filter, then

details-on-demand"), an overview of all models as a heatmap over averages of several quantitative metrics is shown (Figure 8.2a), which enables a targeted filtering of the models based on their quantitative performance. The heatmap of average values paints only a rough picture; upon model selection, histograms of each model's score distribution for each metric are available. (4) After models have been selected, the user is forwarded to the corresponding quality aspect's view.

The visualizations for the individual aspects of the three quality criteria share the property that two texts need to be visually aligned with one another.[6] Despite this commonality, we abstain from creating a single-view visualization "stuffed" with alternative options. We rather adopt a minimalistic design for the assessment of individual quality aspects.

Coverage View (Figure 8.2b,c,d) Content coverage is visualized as alignment of summary sentences and document sentences at the semantic and lexical level in a full-text side-by-side view. Colorization indicates different types of alignments. For entity coverage (relation coverage), a corresponding side-by-side view lists named entities (relations) in a summary and aligns them with named entities (relations) in its source document. For unaligned relations, corresponding document sentences can be retrieved.

Faithfulness View (Figure 8.3, Case A) Hallucinations are visualized by highlighting novel words in a summary. For each summary sentence with a hallucination, semantically and lexically similar document sentences are highlighted on demand. Since named entities and thus also entity relations form a subset of hallucinated words, the above coverage views do the same. Also, in an aggregated view, hallucinations found in multiple summaries are ordered by frequency, allowing to inspect a particular model with respect to types of hallucinations.

Position Bias View (Figure 8.2e,f) Position bias is visualized for all models given a source document, and for a specific model with respect to all its summaries in a corpus. The former is visualized as a text heatmap, where a gradient color indicates for every sentence in a source document how many different summaries contain a semantically or lexically corresponding sentence. The latter is visualized by a different kind of heatmap for 50 randomly selected model summaries, where each summary is projected on a single horizontal bar representing the source document. Bar length reflects

[6] A visualization paradigm recently surveyed by Yousef and Jänicke [335].

document length in sentences and aligned sentences are colored to reflect lexical or semantic alignment.

Aggregation Options Most of the above visualizations show individual pairs of source documents and a summary. This enables the close inspection of a given summary, and thus the manual assessment of a model by sequentially inspecting a number of summaries for different source documents generated by the same model. For these views, the visualizations also support displaying a number of summaries from different models for a relative assessment of their summaries.

8.3 Corpora, Models, and Case Studies

We collected the outputs of 55 summarization approaches on the test sets of three benchmark datasets for the task of single document summarization: CNN/DM, XSum and Webis-TLDR-17. Each dataset has a different style of ground truth summaries, ranging from semi-extractive to highly abstractive, providing a diverse selection of models. Outputs were obtained from NLPProgress, meta-evaluations such as SummEval [97], REALSumm [29], and in correspondence with the model's developers.[7]

8.3.1 Summarization Corpora

The most popular dataset, CNN/DM [132, 206], contains news articles with multi-sentence summaries that are mostly extractive in nature [36, 167]. We obtained the outputs from 45 models. While the original test split of the dataset contained 11,493 articles, we discarded ones that were not summarized by all models, resulting in 11,448 articles total. This minor discrepancy is due to inconsistent usage by authors, such as reshuffling the order of examples, de-duplication of articles in the test set, choice of tokenization, text capitalization, and truncation.

For the XSum dataset [209], the outputs of six models for its test split (10,360 articles) were obtained. XSum contains news articles with more abstractive single-sentence summaries compared to CNN/DM. The Webis-TLDR-17 dataset [313] contains highly abstractive, author-provided (single to multi-sentence) summaries of Reddit posts, although slightly noisier than the other datasets [36]. We obtained the outputs from the four submissions of the TL;DR challenge [284] for 250 posts.

[7] We sincerely thank all the developers for their efforts to reproduce and share their models' outputs with us.

8.3.2 Text Preprocessing

In a preprocessing pipeline, the input of a collection of documents, their ground truth summaries, and the generated summaries from a given model were normalized. First, basic normalization, such as de-tokenization, unifying model-specific sentence delimiters, and sentence segmentation were carried out. Second, additional information, such as named entities and relations were extracted using Spacy[8] and Stanford OpenIE [14], respectively. The latter extracts redundant relations where partial components such as either the subject or the object are already captured by longer counterparts. Such "contained" relations are merged into unique representative relations for each subject.

Alignment Every output summary is aligned with its source document, identifying the top two lexically and semantically related document sentences for each summary sentence. Lexical alignment relies on averaged ROUGE-{1,2,L} scores among the document and summary sentences. The highest scoring document sentence is taken as the first match. The second match is identified by removing all content words from the summary sentence already captured by the first match, and repeating the process as per Lebanoff et al. [171]. For semantic alignment, the rescaled BERTScore [341] is computed between a summary sentence and all source document sentences, with the top-scoring two sentences as candidates.

Summary Evaluation Measures Several standard evaluation measures enable quantitative comparisons and filtering of models for detailed analysis: (1) *compression* as the word ratio between a document and its summary [118], (2) *n-gram abstractiveness* as per Gehrmann et al. [110] calculates a normalized score for novelty by tracking parts of a summary that are already among the n-grams it has in common with its document, (3) *summary length* as word count (not tokens), (4) *entity-level factuality* as per [207] as percentage of named entities in a summary found in its source document, and (5) *relation-level factuality* as percentage of relations in a summary found in its source document. Finally, for consistency, we recompute ROUGE-{1,2,L}[9] for all the models.

8.3.3 Assessment Case Studies

We showcase the use and effectiveness of Summary Explorer by investigating two models (IMPROVE-ABS-NOVELTY, and IMPROVE-ABS-NOVELTY-LM)

[8] https://spacy.io

[9] https://github.com/google-research/google-research/tree/master/rouge

from Kryscinski et al. [166] that improve the abstraction in summaries by including more novel phrases. We investigate the correctness of their hallucinations (novel words in the summary), and identify hidden errors introduced by the sentence fusion of the abstractive models.

Hallucinations via Sentence Alignment Hallucinations are novel words or phrases in a summary that warrant further inspection. Accordingly, our tool highlights them (Figure 8.3, Case A), directing the user to the respective candidate summary sentences whose related document sentences can be seen on demand. For IMPROVE-ABS-NOVELTY, we see that the first candidate improves abstraction via paraphrasing, is concisely written, and correctly substitutes the term *"offenses"* with the novel word *"charges"*. The second candidate also improves abstraction via sentence fusion, where two pieces of information are combined: *"bennett allegedly drove her daughter"*, and *"victim advised she thought she was going to die"*. The novel word *"told"* also fits. However, the sentence fusion creates a wrong relation between the different actors (*"bennett allegedly told her daughter that she was going to die"*), which can be easily identified via the visual sentence alignment provided.

Hidden Errors via Relation Alignment The above showcase does not capture all hallucinations. Summary Explorer also aligns relations extracted from a summary and its source document to identify novel relations. For IMPROVE-ABS-NOVELTY-LM, we see that the relation *"she was arrested"* is unaligned to any relation in the source document (Figure 8.3, Case B). Aligning the summary sentence to the document, we note that it is unfaithful to the source despite avoiding hallucinations (*"Bennett was released on $10,500 bail"*, and not *"arrested on $10,500 bail"*). The word *"arrested"* was simply extracted from the document sentence (Figure 8.3, Case A). Without the visual support, identifying this small but important mistake would have been more cognitively demanding for an assessor.

8.4 Summary

In this chapter, we presented Summary Explorer, an online interactive visualization tool to assess the state of the art in text summarization in a guided fashion. It enables analysis akin to close and distant reading in particular facilitating the challenging inspection of hallucinations by abstractive summarization models. The tool is available open source[10] enabling local use.

[10] `https://github.com/webis-de/summary-explorer`

We also welcome submissions of summaries from newer models trained on the existing datasets as part of our collaboration with the summarization community. We aim to expand the tool's features in future work, exploring novel visual comparisons of documents to their summaries for more reliable qualitative assessments of summary quality. Finally, it is important to note that the accuracy of some of the views is influenced by the intrinsic drawbacks of the toolkits used for named entity recognition and information extraction.

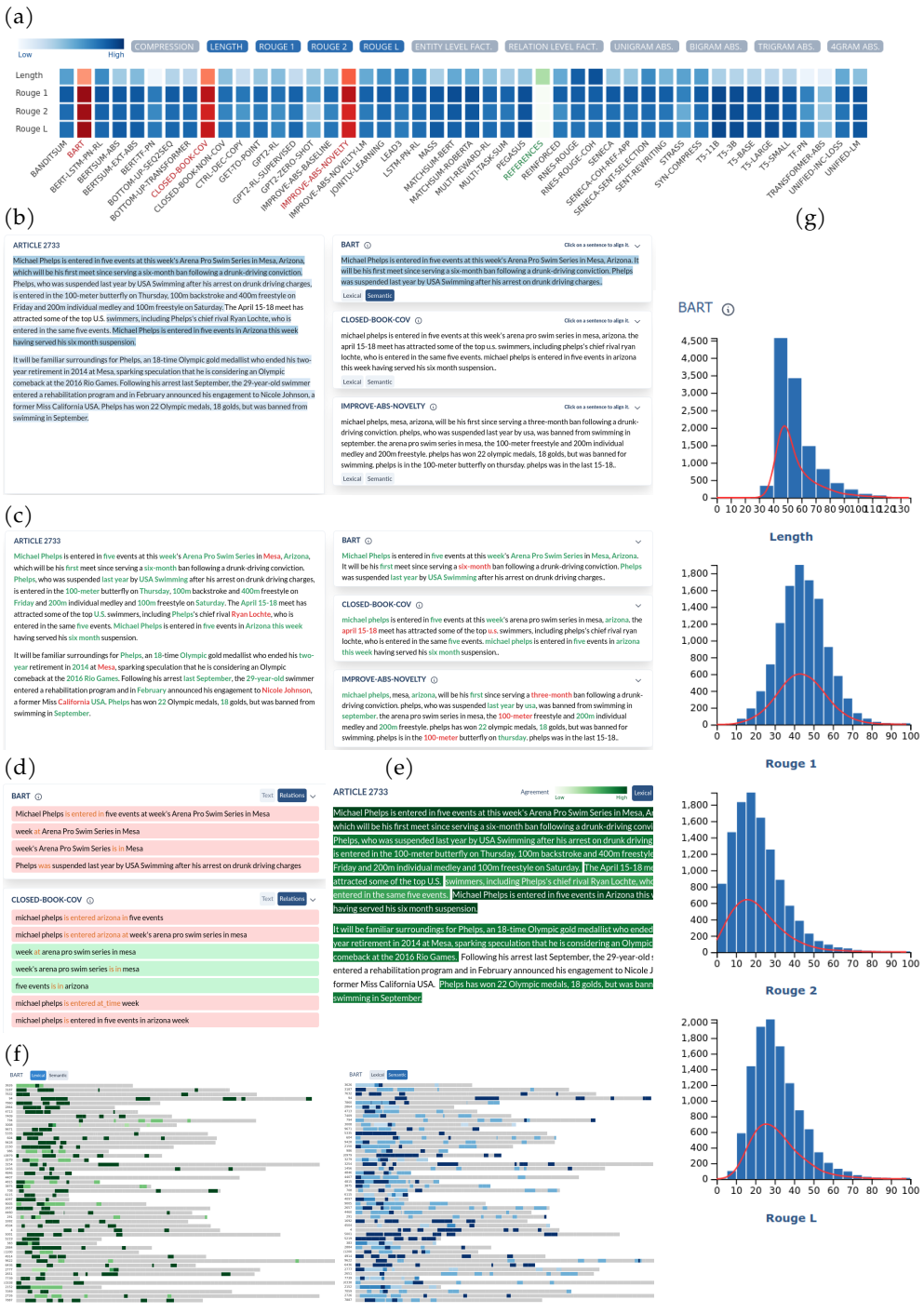


Figure 8.2: (a) Heatmap overview of 45 models for the CNN/DM corpus; ones selected for analysis are highlighted red. Views for (b) the content coverage, (c) the entity coverage, (d) the relation coverage, (e) the position bias across models for a single document, (f) the position bias of a model across all documents as per lexical and semantic alignment, (g) the distribution of quantitative metric scores for a model.

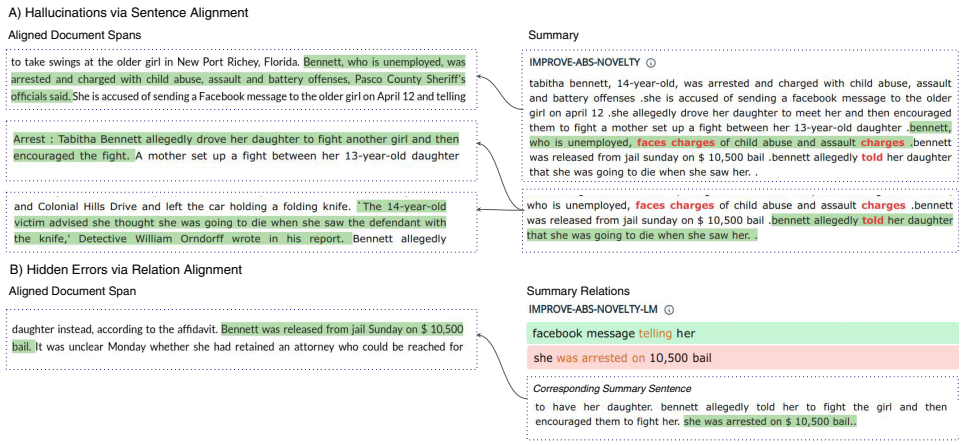


Figure 8.3: Two showcases for identifying inconsistencies in abstractive summaries using Summary Explorer. Case A depicts the verification of the correctness of hallucinations by aligning document sentences. Case B depicts uncovering more subtle hallucination errors by comparing unaligned relations.

9

Summary Workbench: Reproducible Models and Metrics for Text Summarization

The previous chapter introduced Summary Explorer, a visual analytics tool developed to evaluate the summary quality of various models across a static corpus. However, researchers often need to generate summaries from these models on their own unique corpora in the process of developing their own summarization models. Additionally, given the abundance of automatic evaluation metrics available, there is a need for a tool that allows researchers to effortlessly apply these metrics to their own corpora. To address these needs, this chapter presents Summary Workbench, a tool that can be deployed locally and enables researchers to employ multiple state-of-the-art summarization models on their own corpora. This tool also facilitates the evaluation of these summaries using various automatic metrics, all integrated within a unified user interface. Moreover, Summary Workbench allows for a visual comparison of summaries from different models for a specific corpus, with other summaries as well as their source document. The tool emphasizes the importance of reproducibility by enabling researchers to share their models and metrics, along with their dependencies, as self-contained plugins with the summarization research community.

9.1 Addressing the Requirements for Summarization Researchers

Automatic text summarization reduces a long text to its most important parts and generates a summary. Usually, a learning-based summarization

model is developed in two basic steps: *model development* and *model evaluation*. Given a collection of documents accompanied by one or more human-written (reference) summaries, first a set of features representing the documents is manually created or automatically extracted through supervised learning. The resulting model is then used to generate one or more (candidate) summaries, which are analyzed manually and/or with evaluation measures for their similarity to the reference summaries. These steps are iterated, optimizing the model and its parameters using a validation set. The models that perform best in the validation are selected for evaluation on the test set. With standardized test sets for each document collection, comparisons with models created earlier are reported.

However, these steps are associated with comparatively tedious tasks: During model development, summaries of individual documents are often generated and immediately evaluated to identify deficiencies and improve the model, including comparisons to other models. The latter requires third-party models to be operational despite their heterogeneous software stacks. Such "on-the-fly evaluation" during development entails that candidate and reference summaries as well as source documents are analyzed manually or by automatic measures. This multi-text comparison is often not supported by visualization, although this leads to a better understanding of the content coverage and possible selection biases of a model [286, 311]. The analysis of evaluation results for model selection also benefits from visual support [294]. Previous research in the field of automatic summarization has not yet resulted in a unified set of tools for these purposes which is the main goal of this paper.

With Summary Workbench, we introduce the first unified combination of application and visual evaluation environments for text summarization models. Currently, it integrates 15 well-known summarization models (26 variants in total) and 10 standard evaluation measures from the literature. With **FeatureSum**, it also includes a new feature-based extractive summarization model that implements features from the literature predating the deep learning era. Underlying all of the above is a specification and interface that allows easy integration of new models and measures to facilitate large-scale experiments and their reproducibility.

In what follows, we first review existing tools to assist summarization research and development. Section 9.2 overviews the key design principles of the Summary Workbench, and Section 9.3 provides a complete overview of all the models and measures hosted to date. Included are general-purpose models, guided models that accept user prompts to guide summary generation, and models tailored to argumentative language and to news articles.

A wide range of commonly employed evaluation measures are included, covering both lexical as well as semantic overlap measures.[1]

Related Work

The development of tools for summarization research has gained momentum recently, and several tools have been presented for (sub)tasks of the two steps above: Tools such as HuggingFace [330], FairSeq [218], SummerTime [217], TorchMetrics [77], SacreROUGE[78], PyTorch Hub,[2] and TensorFlow Hub[3] focus on hosting several state-of-the-art text summarization models and automatic evaluation measures. These tools have significantly improved accessibility to working models. However, only some provide a very minimal interface for inference of summaries and their online/offline comparative analyses. Many authors also choose to share their models independently, be they on GitHub or elsewhere, as standalone repositories instead of integrating with any tools. To lower the bar of (latter) integration as much as possible, Summary Workbench simplifies model and measure integration as plugins (using Docker). In this way, models under development or private ones can be locally compared to others and can be archived together with all their dependencies for reproducibility. Similar efforts have been made in the information retrieval community via the Docker-based toolkit such as Anserini [332].

Tools such as LIT [294], SummVis [311], and Summary Explorer [286] focus on qualitative model evaluation by providing static visual analyses of the relation between the summary and its source document. Summary Workbench adapts some of their visualizations next to new ones, and complements them with interactive visual analytics for quantitative evaluation according to multiple measures. Users can explore the distribution of scores, select data points of interest and inspect them in relation to the source document to better understand the dataset.

The success of past summarization research and development has relied a lot on in-depth manual error analyses. This being one of the most laborious tasks in every natural language generation evaluation, we believe that visually comparing summaries from multiple models for many different texts, and contextualizing manual review with multiple measures is crucial to both scale up error analysis, and to better understand the capabilities and limitations of the technology. As this still requires juggling many dif-

[1] Source code is available at `https://github.com/webis-de/summary-workbench`.
[2] `https://pytorch.org/hub/`
[3] `https://www.tensorflow.org/hub/`

ferent, incompatible tools, the unified approach of Summary Workbench aims at lowering the bar for scaling up interactive experimentation.

9.2 A Unified Interface for Applying and Evaluating State-of-the-Art Models and Metrics

Summary Workbench implements two interactive views corresponding to the two basic summarization model development steps: a *summarization view* and an *evaluation view*.

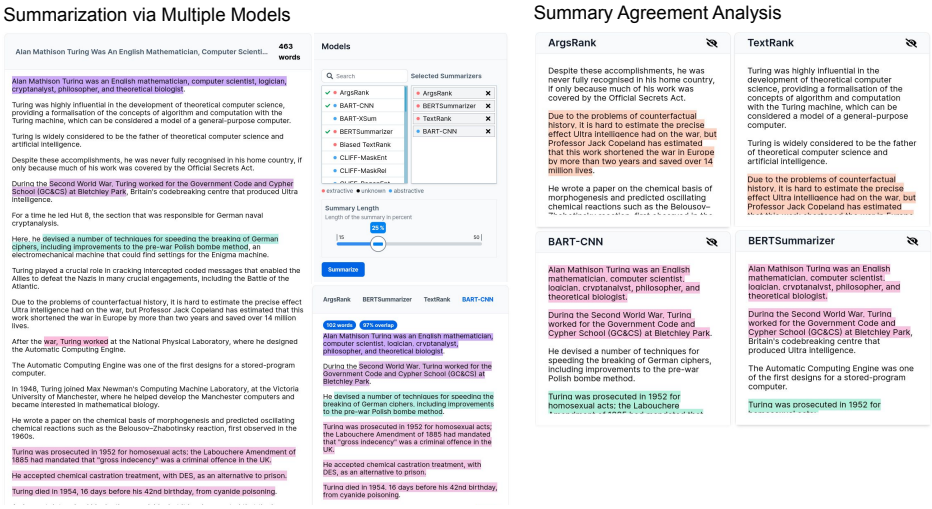


Figure 9.1: Two key components of the summarization view: On the left, an input text can be summarized via multiple extractive and abstractive summarization models; lexical overlap is highlighted on demand for each candidate summary and can be adjusted for varying n-gram lengths. On the right, content agreement among summaries from different models; any summary can be selected as the reference against which the others can be visually compared.

9.2.1 Summarization View

Figure 9.1 depicts the summarization view that allows using multiple extractive/abstractive summarization models to summarize texts, web pages, or scientific documents on demand, controlling for summary length. For scientific documents, relevant sections from a given paper to be summarized can be chosen. Explicit guidance signals for focused summarization can be provided as input to corresponding models (reviewed in Section 9.3).

Generated summaries (candidates) can be visually inspected for their lexical overlaps (highlighted on demand) with their source document or

with another summary. This provides a quick overview of the models' effectiveness at capturing important content as well as any factual errors prevalent in abstractive summarization [198]. Additional functionalities include uploading multiple documents to be summarized via a single file, and command line access to all models.

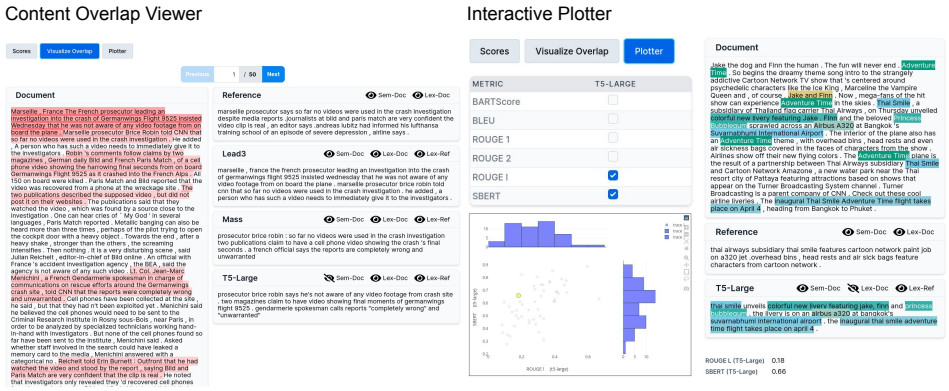


Figure 9.2: Two key components of the evaluation view: On the left, a text overlap viewer displays content coverage of the summaries in relation to the source document via lexical and semantic overlap (via Spacy embeddings). On the right, an interactive plotter allows selecting examples with specific scores for a combination of evaluation measures. Additionally, the distribution of scores is also shown.

9.2.2 Evaluation View

Figure 9.2 shows the evaluation view, where candidate summaries are compared with reference summaries using multiple lexical/semantic content overlap measures. Candidate summaries from multiple models, either generated using the summarization view or uploaded as a file where each example is encoded as $< \texttt{doc}, \texttt{ref}, \texttt{c}_1, \texttt{c}_2, ..., \texttt{c}_n >$ can be evaluated. Lexical/semantic overlap of candidate/reference summaries $\texttt{c}_i/\texttt{ref}$ with the source document `doc` can also be visualized. Computed scores can be neatly exported as CSV or LaTeX tables.

Scores from the evaluation measures can be further explored through an *interactive plotter*. Among other things, the plotter allows users to visually correlate different evaluation measures, identify outliers/challenging source documents or strongly abstractive summaries among the candidates. This facilitates a deeper understanding of the quantitative performance of the models as well as an understanding of the evaluation datasets. Two more use cases of the interactive plotter are explained in Section 9.5.

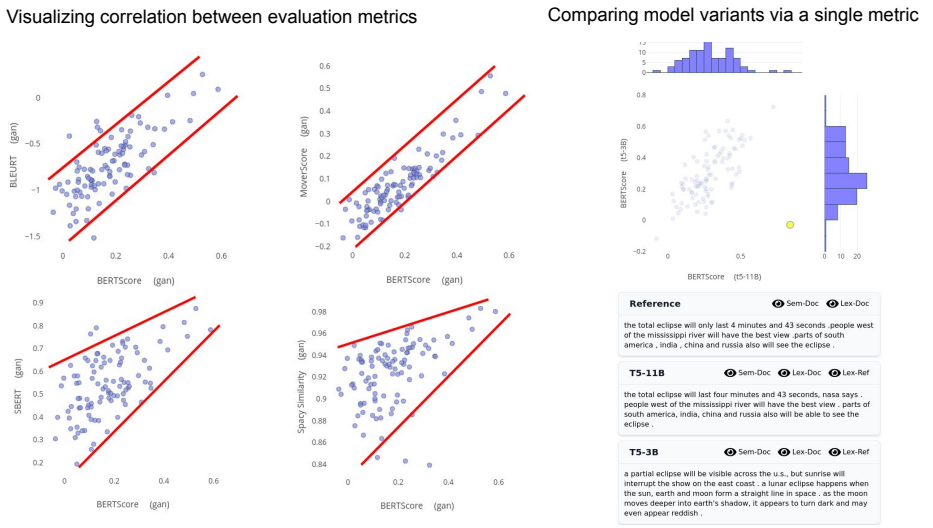


Figure 9.3: Two example use cases of interactive plotter of the evaluation view: On the left, correlations between pairs of evaluation measures are analyzed. On the right, abstractive summaries from two variants of the T5 model for the chosen data point (highlighted yellow) are shown.

9.2.3 Plugin Server

New summarization models and evaluation measures are integrated as container-based plugins. A model/measure plugin can either be a local directory or a remote Git repository containing specification of dependent software and data (checkpoints, embeddings, lexicons), and implementations of the interfaces SummarizerPlugin and MeasurePlugin. Model metadata such as name, type, version, source, citation, and other custom arguments are provided as YAML configurations. Each plugin runs inside a Docker container with its own server that handles API calls following the OpenAPI specification.[4] This setup allows users to safely self-host the entire application. Developed plugins can be easily shared with the community via DockerHub images or Git repositories. Examples are found in our tool's technical documentation.[5]

9.3 Models and Measures

Summary Workbench hosts 15 extractive/abstractive summarization models and 10 lexical/semantic evaluation measures for English text. Each of these

[4] `https://www.openapis.org`
[5] `https://webis.de/summary-workbench/`

Summarizer	Model
BERTSummarizer	distilbert-base-uncased
LoBART	podcast_4K_ORC
Longformer2Roberta	patrickvonplaten/longformer2roberta-cnn_dailymail-fp16
ConcluGen	dbart
CLIFF	pegasus_cnndm
COOP	megagonlabs/bimeanvae
BART	facebook/bart-large
Pegasus	google/pegasus
T5-Base	huggingface.co/t5-base
Evaluator	**Model**
BARTScore	facebook/bart-large-cnn
Spacy Similarity	en_core_web_lg
SBERT	roberta-large-nli-stsb-mean-tokens
BLEURT	bleurt-base-128
BERTScore	roberta-large-mnli
Greedy Matching	glove.6B.300d
MoverScore	MoverScoreV1

Table 9.1: Summarizers and evaluators currently available in Summary Workbench.

is configured as a Docker-based plugin that can be customized and instantiated accordingly. For details on model checkpoints, see Table 9.1.

9.4 Curated Artifacts and Interaction Scenarios

In this section, we initially introduce the assortment of summarization models and evaluation metrics incorporated in the tool. Subsequently, we discuss the interaction scenarios that the tool supports, which allow for the comparison and analysis of scores from multiple metrics in relation to each other. This is particularly useful for a given source document and its corresponding summaries.

9.4.1 Summarization Models

We provide a diverse set of models applicable to multiple text domains such as news, argumentative texts, web pages, and product reviews. Model types

include extractive, abstractive, supervised, unsupervised, and guided summarization, the latter requiring additional user input.

General-purpose Summarization

Models that work in an unsupervised fashion or leverage external knowledge via contextual embeddings are supposed to be capable of summarizing any kind of text. We provide the following models suitable for general-purpose text summarization.

FeatureSum. is our new extractive summarization model which scores a sentence in the text based on a combination of standard features to identify key sentences [186, 212]: TF-IDF, content units (named entities, noun phrases, numbers), position in text, mean lexical connectivity (number of tokens shared with the remaining sentences), ratio of words that are not stop words, length (relative to the longest sentence in the text), and word overlap with the title. The final score of a sentence is the product of the individual feature values. Sentences are then ranked based on these scores to produce the final summary. Different combinations of these features can be chosen by simply toggling them in the interface. This also allows for dynamically reproducing existing models from the literature provided their specific feature sets are available.

TextRank. [201] is a graph-based model which employs PageRank [42] on the document graph consisting of sentences as nodes to compute the strength of their connections. Top-ranked sentences within a length budget are taken as the extractive summary. We also provide the two variants **PositionRank** and **TopicRank**, which consider the sentence position and its overlap with topic sentence (document's title or its first sentence) to compute the ranking via PyTextRank [210].

BERTSum. [202] employs contextual embeddings from BERT [83] to extract key sentences in an unsupervised fashion by first clustering all sentence embeddings using k-means [129] and then retrieving those closest to the centroids as the summary.

PMISum. [221] is an unsupervised extractive model that includes measures to score the relevance and redundancy of the sentences of the source document. These measures are based on pointwise mutual information (PMI) computed by pre-trained language models. Summary sentences are selected via a greedy algorithm to maximize relevance and minimize redundancy.

LoBART. [189] addresses the input length limitations of transformers [309] that restrict capturing long-span dependencies in long document summa-

rization. Local self-attention and explicit content selection modules are introduced to effectively summarize long documents such as podcast transcripts and scientific documents.
Longformer2Roberta. effectively combines Longformer [26], developed for processing long documents, and RoBERTa [182], a robustly trained BERT model as the decoder, based on leveraging pre-trained checkpoints of large language models [255].

Guided Summarization

The following models accept explicit inputs provided by users to guide the summarization process towards generating user-specific summaries.
Biased TextRank. [156] is an extension of the TextRank model which takes an explicit user input as the "focus", represented via contextual embeddings to guide the ranking of the document sentences. Summary extraction is based on the semantic alignment between the document sentences and the provided focus signal.
GSum. [86] is a guidance-based abstractive model that takes different types of external guidance signals: text inputs, highlighted sentences, keywords, or extractive oracle summaries derived from the training data. These signals along with the source text are used to generate focused and faithful abstractive summaries.

Argument Summarization

Summarizing argumentative texts (opinions, product reviews) requires that the model be able to identify high-quality, informative, and argumentative sentences from the text. We provide three models specifically developed for this task.
ArgsRank. [10] is an extractive model for creating argument snippets. It augments TextRank with two new criteria: *centrality in context* and *argumentativeness* to help the model retrieve important and argumentative sentences.
ConcluGen. [287] is a transformer model for generating informative conclusions of argumentative texts by balancing the trade-off between abstractiveness and informativeness of the output. It was finetuned on the Webis-ConcluGen-21 corpus comprised of pairs of argumentative text and a human-written conclusion.
COOP. [143] is an unsupervised opinion summarization model that employs latent vector aggregation by searching for optimal input combinations of sentence embeddings to address the summary vector degeneration problem caused by simple averaging. Specifically, it finds convex combinations

that maximize the word overlap between the source document and its summary.

News Summarization

A majority of the existing summarization models are trained on news datasets, since news have been and are readily available. These models have shown strong performance in creating fluent abstractive summaries [140]. We provide the following models for summarizing news.
BART. [173] is a transformer-based denoising autoencoder for pre-training sequence-to-sequence models. Its main objective is to reconstruct the source text corrupted by employing arbitrary noising functions (masking text spans, randomly shuffling sentences) which helps the model learn better representations of the source texts for text summarization [140].
T5. [245] is a unified text-to-text transformer-based model that exploits the strengths of transfer learning on a variety of problems that can be modeled as text generation tasks. A task-specific prefix is added to each input sequence (e.g., "summarize:<document>") that teaches the model to summarize accordingly.
Pegasus. [339] is a transformer model pre-trained with a self-supervised summarization-specific training objective called "gap-sentences generation": important sentences are removed/masked from the source text and must be jointly generated as output from the remaining sentences, similar to an extractive summary.
CLIFF. [53] leverages contrastive learning for generating abstractive summaries that are faithfully and factually consistent with the source texts. Reference summaries are used as positive examples while automatically generated erroneous summaries are used as the negative examples for training the model.
Newspaper3k. is an open-source library for extracting news articles from the web which provides a module for extractive summarization that ranks sentences based on keywords and title words.[6]

9.4.2 Evaluation Measures

Automatic evaluation measures for summarization typically quantify the lexical/semantic overlap of a candidate summary with a reference summary. We provide the following measures covering both.

[6] `https://newspaper.readthedocs.io/en/latest/`

Lexical Measures

Measures based on lexical overlap return precision, recall, or F1 scores on varying granularities of text between the candidate summary and one or more reference summaries.
BLEU. [225] is a standard measure for machine translation adapted for summarization. It includes a brevity penalty to account for length differences while computing n-gram overlap.
ROUGE. [176] is the most common measure for summarization which computes precision, recall, and F1 scores based on n-gram overlap, where n-grams include unigrams, bigrams, and the longest common subsequence.
METEOR. [20] aligns a candidate with a set of references by mapping each unigram of a candidate to 0/1 unigrams of the reference based on exact, stem, synonym, and paraphrase matches. It then computes precision, recall, and F9 scores (i.e., weighted harmonic mean, strongly emphasizing recall) based on that.
CIDEr. [310] is a consensus-based measure (originally for evaluating image captioning) which measures the similarity of a candidate against a set of references by counting the frequency of the common n-grams of a candidate.

Semantic Measures

Measures based on semantic overlap compute the semantic alignment between candidates and references at the token/word/sentence level based on their static/contextual embeddings.
Greedy Matching. [256] aligns a candidate and a reference by greedily matching each candidate word to a reference word based on their embeddings' cosine similarity. Average similarity over all candidate words aligned to reference words and vice versa are computed whose average is the final score.
MoverScore. [344] combines contextual embeddings from BERT using the word mover's distance [170] to compare a candidate against a set of references by considering both the amount of shared content as well as the extent of deviation between them.
BERTScore. [341] computes a similarity score for each candidate token with each reference token using contextual embeddings from BERT. The measure is also robust to adversarial modifications of the generated text.
BLEURT. [271] is a learned measure based on BERT that models human judgments with a few thousand biased training examples. The model is pre-trained using millions of synthetic examples created via scores from

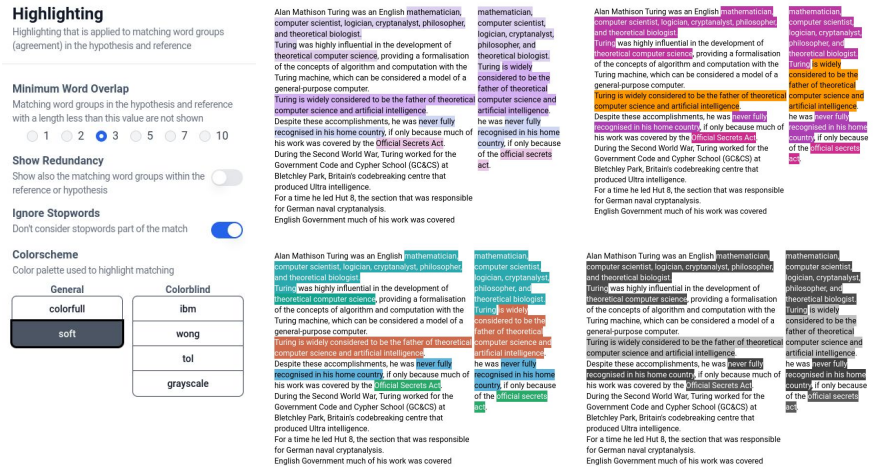


Figure 9.4: Customization options available for visualization of document and summary overlap. Users can select the minimum word overlap, preserve duplicate words, and ignore stop words to be visualized. Also, they can instantly preview each color scheme and set it as their default. The tool provides colorful, soft gradient-based, and grayscale schemes to account for color blindness.

existing measures (BLEU, ROUGE, BERTScore), and textual entailment, for better generalization.

BARTScore. [336] uses the weighted log probability of generating one text given another to compute faithfulness (source → candidate), precision (reference → candidate), recall (candidate → reference), and the F1 score.

CosineSim. includes two embedding-based cosine similarity measures using Spacy word vectors [135] and Sentence-BERT [249].

9.5 Interaction Use Cases

Figure 9.3 shows two use cases of the interactive plotter. First, users can analyze any correlation between two measures of choice for a summarization model. Here, we find that MoverScore and BERTScore have strong correlation as they both employ contextual embeddings from BERT to compute the overlap between candidate and reference summaries. Likewise, we find that the static token embeddings from Spacy have a broader distribution of scores in comparison.

As a second use case, the interactive plotter allows comparing two variants of the same model architecture using any measure. Here, we inspect the T5 model (its 3B and 11B variants) using BERTScore to find that the larger variant generates a summary very similar to the reference while the

smaller variant creates a summary that is topically related but not accurate in comparison to the reference.

9.6 Summary

In this chapter, we presented Summary Workbench, a tool that unifies the application and evaluation of text summarization models. The tool supports integrating summarization models and evaluation measures of all kinds via a Docker-based plugin system that can also be locally deployed. This allows safe inspection and comparison of models on existing benchmarks and easy sharing with the research community in a software stack-agnostic manner. We have curated an initial set of 15 models (26 including all variants) and 10 evaluation measures and welcome contributions from the text summarization community. An extension of the tool's features to related text generation tasks such as paraphrasing and question answering is foreseen.

10 Conclusion

This thesis presented data, methods, and evaluation tools to develop summarization technology for user-generated discourse. In contrast to the majority of the studies on summarization that primarily target news articles, my thesis focuses on textual sources like social media posts, argumentative texts, and forum discussions in particular. The following sections summarize the key contributions and discuss future research directions.

10.1 Key Contributions of the Thesis

Contributions are grouped into three categories: (1) data, (2) methodologies, and (3) evaluation strategies.

Data We developed three new corpora for the domains of news editorials, social media posts, and argumentative texts. The former corpus demonstrates a systematic way to define and operationalize the annotation of high-quality summaries. The remaining corpora are large-scale datasets constructed using author-provided signals of summaries.

1. `Webis-EditorialSum-20 Corpus`: Abstractive summarization of argumentative texts has hardly been explored. To this end, we targeted news editorials, i.e., opinionated articles with a well-defined argumentation structure that do not follow the inverted pyramid style of writing. With Webis-EditorialSum-20, we present a corpus of 1330 carefully curated summaries for 266 news editorials. We evaluate these summaries based on a tailored annotation scheme, where a high-quality summary is expected to be thesis-indicative, persuasive,

reasonable, concise, and self-contained. Our corpus contains at least three high-quality summaries for about 90% of the editorials, rendering it a valuable resource for the development and evaluation of summarization technology for long argumentative texts.

2. `Webis-TLDR-17 Corpus`: Annotating high-quality summaries at scale is infeasible. Therefore, we exploit the common practice of authors providing a `TL;DR` with their posts and created a unique corpus for abstractive summarization of social media posts. This corpus contains almost 4 million posts from Reddit (2006-2016) and their highly abstractive, author-provided summaries. We also provide a large-scale evaluation of state-of-the-art summarization models on this corpus via a shared task for abstractive summarization.

3. `Webis-ConcluGen-21 Corpus`: The purpose of an argumentative text is to support a certain conclusion. Yet, conclusions are often omitted, expecting readers to infer them rather. This rhetorical device limits accessibility when browsing many texts (e.g., on a search engine or on social media). In these scenarios, an explicit informative conclusion focused on specific concepts makes for a good candidate summary of an argumentative text. Introducing the task of generating informative conclusions, we compiled a large-scale corpus of 136,996 samples of argumentative texts and their conclusions. Additional pieces of external knowledge such as topic, target, and aspects of the argumentative text are also provided to facilitate the generation of informative conclusions.

Methodologies We devised both supervised and unsupervised approaches including zero-shot settings for abstractive summarization:

1. *Supervised Learning*: We employed a couple of techniques to generate informative conclusions. Firstly, we utilized control codes to encode external information, which was then used for fine-tuning our pretrained language models. Secondly, to categorize arguments in extensive discussions, we enhanced a state-of-the-art classification model, enabling it to predict multiple frames per argument.

2. *Unsupervised Learning*: We devised two distinct approaches for summarizing long discussions. For creating *informative* summaries, we adopted an entirely unsupervised method that relies on information retrieval models. This method generates extractive summaries of argument groups, each of which focuses on a specific argumentation

frame. For facilitating exploration of detailed discussions through *indicative* summaries, we crafted an end-to-end unsupervised approach. This approach first clusters the discussion into coherent subtopics, then generates a summary for each subtopic, and finally assigns argumentation frames to each subtopic.

We leveraged state-of-the-art large language models in this pipeline and designed effective prompts following best practices and thorough experimentation. This strategy enabled us to generate two-level indicative summaries, similar to a book's table of contents, for extensive discussions on any topic, eliminating the need for any labeled data.

Evaluation Strategies We introduced innovative visual analytics techniques to enhance the evaluation of the quality of the generated summaries. Through the Summary Explorer, we developed intuitive side-by-side comparisons of the summary and the source document, enabling a more effective assessment of the quality of the generated summaries. Additionally, we offer a corpus-level visual analysis of position bias across various summarization models. Likewise, the Summary Workbench consolidates model development and evaluation within a single interface. Each model and evaluation metric can be deployed in a reproducible fashion, promoting transparency and replicability in research.

We have made all data, methods (models), and evaluation tools publicly accessible to the broader research community, fostering open collaboration and knowledge sharing.

10.2 Open Problems and Future Work

The rapid and remarkable progress of large language models has ushered in an era of unprecedented opportunities in the realm of summarization research. This progress has, in turn, led to substantial enhancements in the overall linguistic quality of generated summaries. Demonstrations conducted through crowdsourced evaluations have convincingly showcased that GPT-3 has surged past a multitude of state-of-the-art supervised summarization models, particularly in the field of news summarization [113]. Notably, human preferences have aligned significantly with this advancement, establishing GPT-3's widespread acceptance.

The emergence of prompt-based instruction-following generative models offers higher flexibility for controlling the style, length, and structural dimensions of the summaries. However, this flexibility also introduces novel challenges in evaluating summaries of equivalently high quality of sum-

maries resulting from diverse models. Here, Summary Workbench may help the developers to better compare equivalent model variants. In order to address this issue, we propose focusing on two specific evaluation aspects: faithfulness and purpose, which we perceive to be the benchmark summary quality dimensions for future research.

Firstly, evaluating faithfulness and factuality is an open challenge. Large language models often generate information with undue confidence, making its verification difficult by simply reading the summary and the corresponding source document. While the generated summary is still based on the contents of the source document, assessing its faithfulness requires additional efforts beyond automatic metrics. To address this, we suggest extending the Summary Explorer tool by incorporating web search capabilities into visualizations. This enhancement allows users to verify information against various sources. Also, LLMs may be integrated that automatically assess certain quality dimensions of the summary by chain-of-thought prompting [181].

Additionally, the purpose of summarization is crucial but currently overlooked. Effective evaluation requires understanding why a summary was generated. For instance, a book blurb should avoid spoilers. This aspect should be considered alongside the existing evaluation criteria. Therefore, we anticipate the need for a new evaluation paradigm that incorporates the purpose of the summary as a first class citizen into the evaluation pipeline.

In conclusion, large language models offer promising opportunities for summarization research such as personalization, interactive (or assisted) summarization, and leveraging document structure to better organize the summaries. Challenges include evaluating faithfulness and considering the summary's purpose. By addressing these, we can enhance the overall summarization quality and its usefulness for the downstream tasks.

A Appendix

A.1 Argumentativeness Scoring for Frame Assignment

The dataset from Schiller et al. [265] consists of topics formulated as phrases as opposed to the topic titles in CMV, which are often formulated as claims. To unify this, we manually transformed their topics by appending them with stance-indicative phrases (e.g., "Abortion" $\rightarrow$ "Abortion should be banned"). We trained the RoBERTa model for the binary classification task with default training parameters: a learning rate of 5e-5, 5% of the training data for warmup, early stopping, and a batch size of 32. On the test split provided by Schiller et al. [265], our fine-tuned model performs with a macro-F1 of 67%, which is comparable with the results from the best model reported in Schiller et al. [265]. A text is labeled as argumentative if the output probability from the finetuned classifier is higher than 50%. Given an input text and the discussion topic we take the mean scores of its constituent sentences as the text's argumentativeness score.

A.2 Collecting Relevance Judgments for Frame Assignment

Annotation interfaces for the pilot study and the main evaluation are shown in Figures A.1 and A.2, respectively. We improved the interface for our main evaluation based on annotator feedback from the pilot study with the following changes: (1) We substituted "probably" with "rather" in our scales to indicate a clearer relevance judgment. (2) For non-argumentative texts or meta-arguments (e.g. "I agree.", "I don't understand what you mean." etc.), we allowed annotators to mark the text as *noisy* and skip it. (3) We

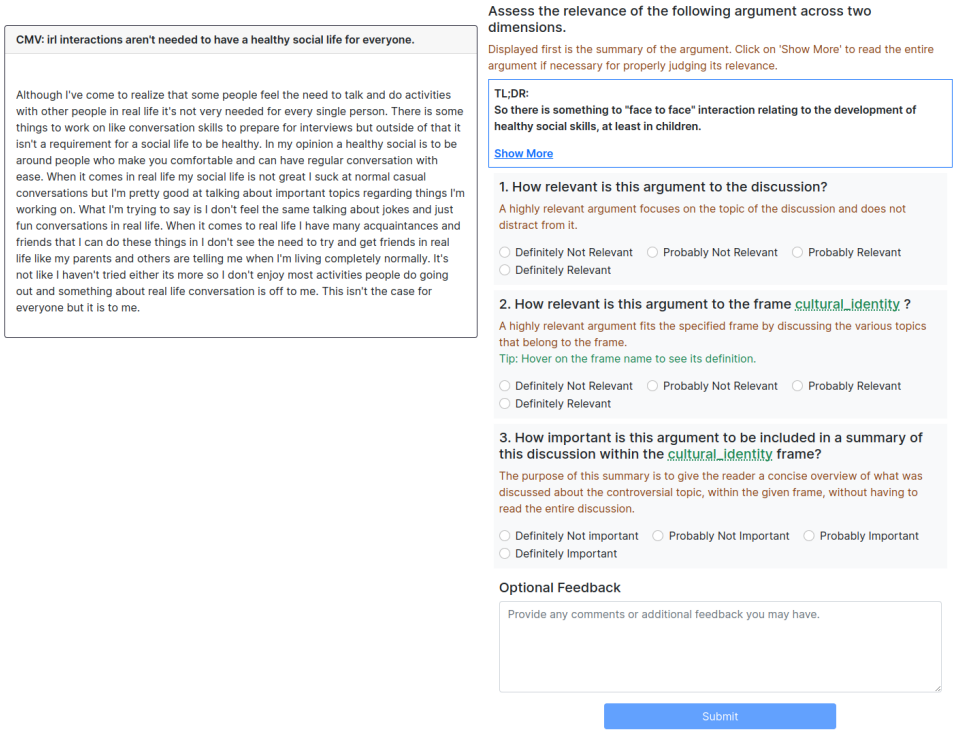


Figure A.1: Annotation interface for the **pilot study**. Annotators were provided a summary of the argument alongside the entire argument. There was no option to mark a text as noisy/non-argumentative. Furthermore, the importance of an argument was assessed based on how likely it was to be included in a frame-oriented *summary* of the discussion.

asked annotators to select at least one relevant frame if the current frame was (definitely/rather) not relevant, with the possibility of selecting multiple frames if required.

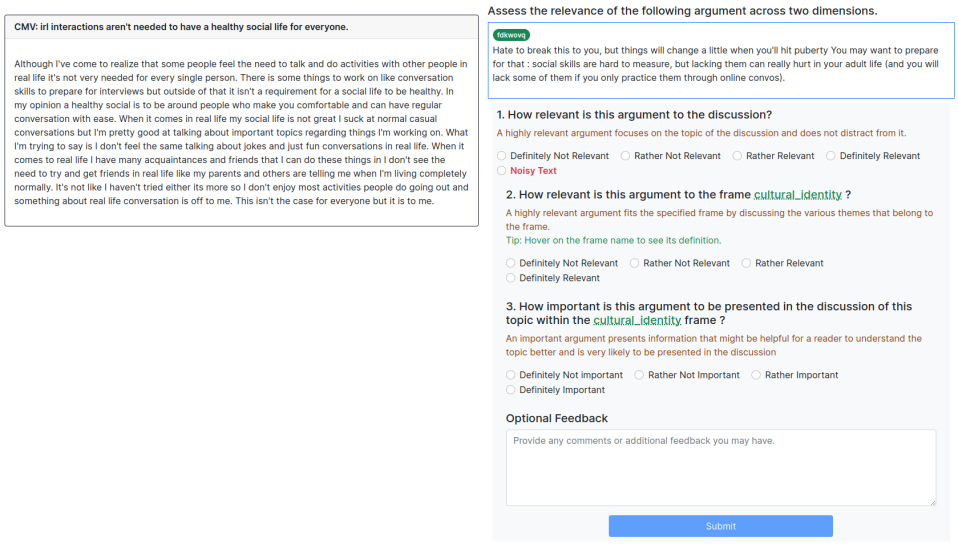


Figure A.2: Annotation interface for the **main evaluation**. First, we removed the summary of the argument and always showed the complete argument. Next, we allowed marking a text as "noisy" and skip answering the remaining questions. Finally, as it was difficult to decide if an argument was important enough to be included in a summary of the discussion before reading the entire discussion, we rephrased the important question as the likelihood of including an argument in the *discussion* of the topic.

Frame	Description
Capacity & Resources	The lack of or availability of physical, geographical, spatial, human, and financial resources implement or carry out policy goals.
Constitutionality & Jurisprudence	The constraints imposed on or freedoms granted to individuals, government, and corporations.
Crime & Punishment	Specific policies about enforcement and interpretation of laws by individuals and law enforcement, breaking laws, loopholes, fines, sentencing and punishment.
Cultural Identity	The social norms, trends, values and customs constituting culture(s), as they relate to a specific policy issue.
Economic	The costs, benefits, or monetary/financial implications of the issue.
External Regulation & Reputation	A country's external relations with another nation; the external relations of one state with another; or relations between groups.
Fairness & Equality	Equality or inequality with which laws, punishment, rewards, and resources are applied or distributed among individuals or groups.
Health & Safety	Healthcare access and effectiveness, illness, disease, sanitation, obesity, mental health effects, prevention of or perpetuation of gun violence, infrastructure and building safety.
Morality	Any perspective—or policy objective or action— that is compelled by religious doctrine or interpretation, duty, honor, righteousness or any other sense of ethics or social responsibility.
Policy Prescription & Evaluation	Particular policies proposed for addressing an identified problem, and figuring out if certain policies will work, or if existing policies are effective.
Political	Issue actions or efforts or stances that are political, such as lobbyist involvement, bipartisan efforts, deal-making and vote trading.
Public Opinion	References to general social attitudes, polling and demographic information.
Quality of Life	The effects of a policy, an individual's actions or decisions, on individuals' wealth, mobility, access to resources, happiness, social structures, quality of community life, etc.
Security & Defense	Security, threats to security, and protection of one's person, family, in-group, nation, etc.
Other	Any frames that do not fit into the above categories.

Table A.1: Descriptions of frames as per Boydstun et al. [40]. Descriptions have been adapted for clarity and conciseness.

Posts		Comments	
Frame	**Count**	**Frame**	**Count**
Cultural Identity	53	Cultural Identity	13,540
Quality of Life	37	Economic	8931
Economic	33	Quality of Life	8559
Public Opinion	26	Public Opinion	7257
Health & Safety	22	Political	5177
Political	19	Health & Safety	4927
Morality	12	Morality	4237
Policy Prescription & Evaluation	10	Policy Prescription & Evaluation	4108
Fairness And Equality	10	Constitutionality & Jurisprudence	3226
Constitutionality & Jurisprudence	9	Fairness & Equality	2457
Security & Defense	1	Crime & Punishment	898
Crime & Punishment	1	Security & Defense	515
		External Regulation & Reputation	216
		Capacity & Resources	169

Table A.2: Counts of frames in posts and comments in our dataset of 100 discussions as predicted by `SuperFrame`. Since each text can be assigned multiple frames, the counts include duplicates. Here, we observe that there are two additional frames found in the comments: *External Reputation & Regulation*, *Capacity & Resources* that are not found in the posts.

Table A.3: Automatic metrics for evaluating summarization.

Metric	Description	Overlap	Unit	Reference-based
BLEU [225]	A corpus-level precision-focused metric primarily used to evaluate automatic machine translation that calculates n-gram overlap between candidates and references while including a brevity penalty.	Lexical	n-gram	✓
ROUGE [176]	"Recall-Oriented Understudy for Gisting Evaluation" scores a candidate summary by counting the number of its n-gram overlaps (unigram, bi-gram, skip-bigram, and longest common subsequences) with the reference summaries.	Lexical	n-gram	✓
METEOR [20]	Computes alignment between candidate and reference sentences by mapping unigrams in the candidate summary to 0 or 1 unigrams in the reference, based on stemming, paraphrastic matches, and synonyms. It reports the harmonic mean of precision and recall.	Lexical	unigram	✓

Table A.3: Automatic metrics for evaluating summarization (continued).

Metric	Description	Overlap	Unit	Reference-based
`Basic Elements` [137]	Computes the overlap between candidate the reference summary via a set of minimal semantic units known as basic elements such as syntactic constituent heads (noun-/verb-/adjective-/adverbial- phrases) and head-modifier-relation triples. Lexical parsers are used to extract such basic elements given a text. Overlap is computed in several forms such a lexically, lemma matching, synonyms, and approximate phrasal paraphrases. Each identified basic element is given a score of 1 for each reference summary it participates in which is then used to score candidate summaries based on how many high-scoring basic elements are contained in them.	Lexical	phrase	✓
`Pyramid` [213]	Analyzes multiple human-made summaries into "Summary Content Units" and assigns importance weights to each SCU. Different candidate summaries are scored by assessing the extent to which they cover SCUs according to their respective weights.	Lexical	phrase	✓

Table A.3: Automatic metrics for evaluating summarization (continued).

Metric	Description	Overlap	Unit	Reference-based
`Greedy Matching` [256]	Greedily aligns each word in the candidate summary to a word in the reference summary based on the cosine similarity of their word embeddings, averaging these similarities over the number of words in the candidate summary. The same score is computed by reversing the roles of the candidate and the reference sentences and the average of the two scores gives the final similarity score.	Semantic	word	✓
`ROUGE-WE` [215]	Extends ROUGE by including word embeddings (Word2Vec) for computing soft lexical matching based on the cosine similarity.	Semantic	n-gram	✓
`CIDEr` [310]	Computes {1-4} gram co-occurrences between candidate and reference summaries, down-weighting common n-grams and calculating the cosine similarity between the n-grams of the candidate and reference texts	Lexical	n-gram	✓
`CHRF` [235]	Calculates F-score for character n-gram overlap between the candidate and reference. This was originally introduced for evaluating machine translation.	Lexical	n-gram	✓

Table A.3: Automatic metrics for evaluating summarization (continued).

Metric	Description	Overlap	Unit	Reference-based
`MoverScore` [344]	Measures the semantic distance between a candidate summary and a reference via Word Mover's Distance [170] operating over n-gram (pooled) BERT embeddings.	Semantic	n-gram	✓
`Sentence Mover's Similarity` [65]	Extends Word Mover's Distance viewing documents as a bag-of-sentence embeddings as well as a variation including just a bag-of-sentences and a bag-of-words.	Semantic	sentence	✓
`BERTScore` [341]	Computes similarity scores by greedily aligning the candidate and reference summaries on a token-level to maximize the cosine similarity between the contextualized token embeddings from BERT [83]	Semantic	token	✓
`BARTScore` [336]	Models the evaluation as a text generation problem via sequence-to-sequence models. The intuition is that models trained to convert the generated text to/from a reference or the source document will achieve higher scores when the candidate (generated text) is better. It uses the weighted log probability of generating one text given another to compute faithfulness (source $\rightarrow$ candidate), precision (reference $\rightarrow$ candidate), recall (candidate $\rightarrow$ reference), and the F1 score.	Semantic	–	Optional

Table A.3: Automatic metrics for evaluating summarization (continued).

Metric	Description	Overlap	Unit	Reference-based
`QA-based Evaluation` [58]	Generate a set of questions for named entities in a text via predefined templates, which need to be answered by the summaries from different systems for a given set of documents (MDS task). Questions are used as queries in combination with an open source QA framework to rank the summaries by the number of correct answers.	Semantic	text	✓
`APES` [96]	A question-answering based metric that receives a set of documents, question-answer pairs, and an automatic QA system to determine the total number of questions answered correctly according to the candidate summaries. Questions are generated from reference summaries by masking its constituent named entities.	Semantic	–	✓
`SummaQA` [268]	Conceptually similar to APES that computes both F1 score and the confidence of the QA system for answering questions generated by masking entities from the document instead of the references. It uses a BERT model pretrained on the SQuAD dataset [246] as the QA system.	Semantic	–	✗

Table A.3: Automatic metrics for evaluating summarization (continued).

Metric	Description	Overlap	Unit	Reference-based
FEQA [88]	A QA-based reference-less metric that evaluates faithfulness of the candidate summary given its source document. Given question-answer pairs generated from the summary, a QA model extracts answers from the document; non-matched answers indicate unfaithful information in the candidate summary. Questions are generated from the candidate summary by masking both noun phrases and named entities which are taken as the gold answer. A BART model [173] finetuned on QA2D dataset [76] is used for generating questions (QG model), while BERT pretrained on the SQuAD dataset is used as the QA system.	Semantic	–	✗

Table A.3: Automatic metrics for evaluating summarization (continued).

Metric	Description	Overlap	Unit	Reference-based
QAGS [323]	A QA-based reference-less metric that identifies factual inconsistencies in a candidate summary given its source document. Given a QG model that generates questions based on a candidate summary, a QA model to answer those questions based on both the candidate summary as well as the source document, a quality score is computed based on the similarity of the corresponding answers, computed via F1 score on a token level. It uses a BART model finetuned on the NewsQA dataset [302] as the QG model and BERT finetuned on SQuAD 2.0 as the QA model.	Semantic	–	✗
QAEval [80]	Conceptually similar to the APES metric but asks questions based on noun phrases instead of named entities to estimate the content quality of the candidate summary based on the proportion of the correct answers to the questions derived from corresponding reference summary. It uses the pre-trained ELECTRA model [66] finetuned on the SQuAD 2.0 dataset as the QA system.	Semantic	–	✓

Table A.3: Automatic metrics for evaluating summarization (continued).

Metric	Description	Overlap	Unit	Reference-based
`QuestEval` [269]	A QA-based reference-less metric that accounts for both factual consistency and relevance of the candidate summary. It computes precision, recall, F1 score and reports the harmonic mean of precision and recall of the correct answers as the final score. It also employs query weighting to distinguish important questions from anecdotal ones via a classifier trained on a supervised dataset of document-summary pairs that gives a higher weight to those questions from the document whose answers are contained in the human summary.	Semantic	–	✗
`BLEURT` [271]	A learned evaluation metric based on BERT that can model human judgments with a few thousand biased training examples. It is based on a BERT model pretrained on a large number of synthetic reference-candidate pairs generated via lexical and semantic modifications such as mask-filling of random tokens, backtranslation, and dropping random words. This pretrained model is then finetuned on a few thousand task-specific evaluation judgments.	Semantic	–	✓

Table A.3: Automatic metrics for evaluating summarization (continued).

Metric	Description	Overlap	Unit	Reference-based
BLANC [308]	A reference-less metric that measures the performance gains of a pretrained language model given access to a document's summary while carrying out natural language understanding (NLU) tasks on the source document.	Semantic	–	✗
SUPERT [106]	A reference-less metric that rates the quality of a candidate summary by measuring its semantic similarity with a pseudo reference summary created by selecting salient sentences from the source documents, using contextual embeddings and soft token alignments.	Semantic	token	✗
GPTScore [100]	The core idea is that a generative pretraining model will assign a higher probability of high-quality generated text following a given instruction and context. Instructions can be provided in natural language consisting of the task description, specific aspect definitions on which the task must be evaluated, and a source (or reference) text. The score is defined as the conditional probability (sum of token weights at each step) of generating a candidate text (summary) given the source document (or reference summary) in the context of the prompt.	Semantic	–	Optional

A.3 Preprocessing Discussions

Deleted posts were matched using: "[deleted]", "[removed]", "[Wiki][Code][/r/DeltaBot]", "[History]". To remove posts from moderators, we used:

- "hello, users of cmv! this is a footnote from your moderators"
- "comment has been remove"
- "comment has been automatically removed"
- "if you would like to appeal, please message the moderators by clicking this link."
- "this comment has been overwritten by an open source script to protect"
- "then simply click on your username on reddit, go to the comments tab, scroll down as far as possibe (hint:use res), and hit the new overwrite button at the top."
- "reply to their comment with the delta symbol"

A.4 Soft Clustering Implementation

We employed HDBSCAN, a soft clustering algorithm [52] to cluster the contextual sentence embeddings from SBERT [249]. As these embeddings are high dimensional, we follow Grootendorst [117] and apply dimensionality reduction on these embeddings via UMAP [200] and cluster them based on their euclidean distance. Most parameters were selected according to official recommendations for UMAP,[1] and HDBSCAN.[2]

UMAP Parameters

metric We set this to "`cosine`" because this is the natural metric for SBERT embeddings.

n_neighbors We set this to 30 instead of the default value of 15 because this makes the reduction focus more on the global structure. This is important since the local structure is more sensitive to noise.

[1] `https://umap-learn.readthedocs.io/en/latest/clustering.html`
[2] `https://hdbscan.readthedocs.io/en/latest/parameter_selection.html`

n_components We set this value to 10.

min_dist We set this value to 0 because this allows the points to be packed closer together which makes separating the clusters easier.

HDBSCAN Parameters

metric We set this to "`euclidean`" because this the target metric that UMAP uses for reducing the points.

cluster_selection_method We set this value to "`leaf`". An alternative choice for this options is "`eom`". This option has the tendency to create unreasonably large clusters. There are instances where it creates only two or three clusters even for very large discussions. The "`leaf`" method does not suffer from this problem but it is more dependent on the "`min_cluster_size`" parameter.

min_cluster_size This parameter is the most important one for this approach. It is also not straight forward to find a value for this since the sizes of the main subtopics of a discussion depend on the size of the discussion. To find a good value, we sampled 50 discussion randomly and 50 discussion stratified by discussion length from all discussions. We compute the clustering for all 100 discussion for different values for min_cluster_size and manually determine a lower and upper bound for min_cluster_size that give a good clustering. We computed a regression model using the following function family as a basis: $f(x|a,b) = a \cdot x^b$ The input variable x is the number of sentences in the discussion and the output variable is the average of the upper and lower bound. This yields the following function for computing min_cluster_size: $f(x) = 0.421 \cdot x^{0.559}$. Figure A.5 visualizes upper and lower bounds as well as the found model.

A.5 Generative Cluster Labeling

Model Descriptions Best prompts for the manually evaluated models (Section 7.4.2) are shown in Figure A.6. For completion, we also generated cluster labels using instruction-following models. Direct and dialogue instructions for these models are shown in Figure A.7.

1. **`T0`** [261] is a prompt-based encoder-decoder model, fine-tuned on multiple tasks including summarization, and surpasses GPT-3 in some tasks

despite being much smaller. It was trained on prompted datasets where supervised datasets were transformed into prompts.

2. **BLOOM** [263] is an autoregressive LLM with 176B parameters, which specializes in prompt-based text completion for multiple languages. It also supports instruction-based task completions for previously unseen tasks.

3. **GPT-NeoX** [34] is an open-source, general-purpose alternative to the GPT-3 model [219] containing 20B parameters.

4. **OPT** [340] is an autoregressive LLM with 66B parameters from the suite of decoder-only pre-trained transformers. These models offer similar performance and sizes as GPT-3 while employing more efficient practices for data collection and model training.

5. **GPT3.5** [46, 219] is an instruction-following LLM with 175B parameters that outperforms the GPT-3 model across several tasks by consistently adhering to user-provided instructions and generating high-quality, longer outputs. We used the *text-davinci-003* variant. In contrast to the other open-source models, it is accessible exclusively through the OpenAI API.[3]

Prompt Descriptions We investigated several prompt templates for each model and selected the best performing one. All the prompts investigated for **T0** are shown in Table A.9. Prompt templates for the autoregressive models (**BLOOM**, **OPT**, **GPT-NeoX**) are listed in Table A.10. Prompt templates for the instruction-following LLMs are listed in Table A.11.

Automatic Evaluation For the sake of completion, we automatically evaluated the recently released (at the time of writing) instruction-following models. To adapt them to generative cluster labeling, we devised two instructions (Figure A.7) similar to the direct and dialogue style instructions used for frame assignment (Section 7.2.3). Next, we computed BERTScore and ROUGE against two sets of references: (1) manually annotated ground truth labels for 300 clusters, and (2) cluster labels from `GPT3.5` which was the best model as per our manual evaluation (Section 7.4.2, Table 7.3). Complete results for BERTScore along with length distributions for the generated cluster labels are shown in Table A.4, while results for ROUGE are shown in Table A.5.

[3] `https://platform.openai.com/docs/models/gpt-3-5`

Model	Reference			GPT3.5			Length		
	P	R	F1	P	R	F1	Min	Max	Mean
Alpaca-7B	0.20	0.15	0.17	0.31	0.28	0.29	3	21	7.92
Baize-13B	0.17	0.15	0.16	0.33	0.32	0.32	1	39	8.47
Baize-7B	0.22[3]	0.19	0.20	0.38[3]	0.38	0.38[3]	2	46	10.73
BLOOM	0.15	0.09	0.11	0.22	0.19	0.20	1	54	8.13
Falcon-40B	0.12	0.09	0.10	0.17	0.17	0.17	1	57	9.57
Falcon-40B-Inst.	0.22[3]	0.18	0.20	0.34	0.32	0.33	2	33	9.34
ChatGPT	0.23[2]	**0.24**[1]	**0.23**[1]	0.39[2]	**0.43**[1]	**0.41**[1]	3	34	11.10
GPT4	0.21	0.19	0.20	0.37	0.36	0.37	4	18	7.50
GPT-NeoX	0.19	0.07	0.12	0.24	0.17	0.20	1	34	7.42
LLaMA-30B	0.12	0.06	0.08	0.19	0.17	0.17	1	46	9.58
LLaMA-CoT	**0.24**[1]	0.21[2]	0.22[2]	**0.41**[1]	0.39[3]	0.40[2]	3	29	8.45
LLaMA-65B	0.08	0.02	0.05	0.14	0.14	0.14	1	46	10.27
OASST	0.22[3]	0.21[2]	0.21[3]	0.39[2]	0.40[2]	0.40[2]	3	31	10.15
OPT	0.16	0.09	0.12	0.22	0.19	0.20	1	30	8.27
Pythia	0.19	0.13	0.16	0.31	0.27	0.29	2	34	7.69
T0	0.15	0.00	0.06	0.15	0.03	0.09	1	18	3.10
GPT3.5	0.23[2]	0.20[3]	0.21[3]	–	–	–	3	27	9.44
Vicuna-13B	0.21	0.21[2]	0.21[3]	0.36	0.39[3]	0.37	3	39	11.87
Vicuna-7B	0.20	0.19	0.19	0.34	0.37	0.35	2	42	11.47

Table A.4: Complete results of automatic evaluation via BERTScore for the cluster labeling task of all 19 LLMs. We compared them against the manually annotated reference and **GPT3.5**, the best model from our manual evaluation. The top three models are indicated for each metric. Similar to the ROUGE evaluation, we see a strong performance by **ChatGPT** and **LLaMA-CoT**. Also shown are the statistics of the length of the generated cluster labels (in number of tokens).

Manual Evaluation Table A.7 shows the guideline provided to the annotators. Figure A.3 shows the annotation interface used to collect the rankings for cluster label quality.

Model	Reference			GPT3.5		
	R-1	R-2	R-LCS	R-1	R-2	R-LCS
Alpaca-7B	13.89	3.10	12.65	19.98	6.08	18.05
Baize-13B	14.44	2.28	13.02	24.59	8.23	22.53
Baize-7B	17.40	2.88	14.95	26.35	9.43	23.89
BLOOM	12.52	2.52	11.34	13.10	3.74	12.20
Falcon-40B	14.30	3.06	13.26	13.08	3.49	11.92
Falcon-40B-Inst.	17.59	3.97[3]	15.48[3]	21.72	7.66	19.57
ChatGPT	**20.15**[1]	**4.88**[1]	**17.42**[1]	**29.52**[1]	10.99[2]	25.75[2]
GPT4	16.43	2.84	14.42	27.76[3]	9.57[3]	24.80[3]
GPT-NeoX	12.93	2.37	11.72	11.67	2.41	10.72
LLaMA-30B	12.30	2.60	11.19	12.07	2.70	11.14
LLaMA-CoT	18.91[2]	4.50[2]	16.83[2]	28.94[2]	**11.69**[1]	**26.38**[1]
LLaMA-65B	10.25	1.93	9.40	10.81	2.49	9.95
OASST	18.28[3]	3.58	16.15	27.13	9.09	23.88
OPT	11.67	2.68	10.87	10.56	2.17	9.55
Pythia	14.78	2.99	13.23	21.64	6.44	19.67
T0	9.80	2.01	9.61	7.64	1.70	7.52
GPT3.5	16.82	2.96	14.61	–	–	–
Vicuna-13B	16.90	3.02	14.81	25.32	8.66	22.66
Vicuna-7B	17.04	2.62	14.81	23.88	7.42	20.87

Table A.5: Complete results of automatic evaluation via ROUGE for the cluster labeling task of all 19 LLMs. We compared them against the manually annotated reference and **GPT3.5**, the best model from our manual evaluation. The top three models are indicated for each metric. We see that **ChatGPT** and **LLaMA-CoT** perform strongly across the board.

Figure A.3: Annotation interface for ranking-based qualitative evaluation of cluster labels.

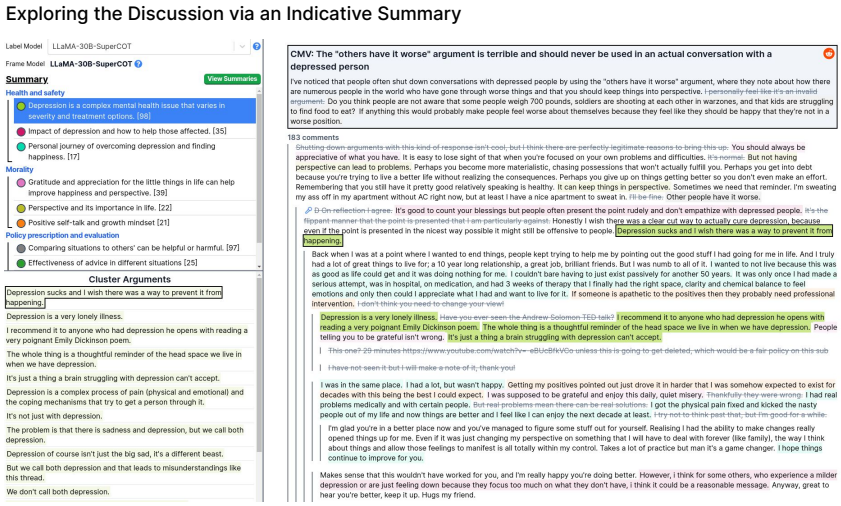


Figure A.4: An exploratory view provided by Discussion Explorer to quickly navigate a long discussion via an indicative summary. On the left, clicking on a cluster label lists all its constituent sentences. On the right, a specific sentence from the chosen cluster is presented in the context of the discussion. Softly highlighted are the sentences from other clusters that surround the selected sentence. Users can thus easily skim a discussion with several arguments for relevant information using the indicative summary in this exploratory view.

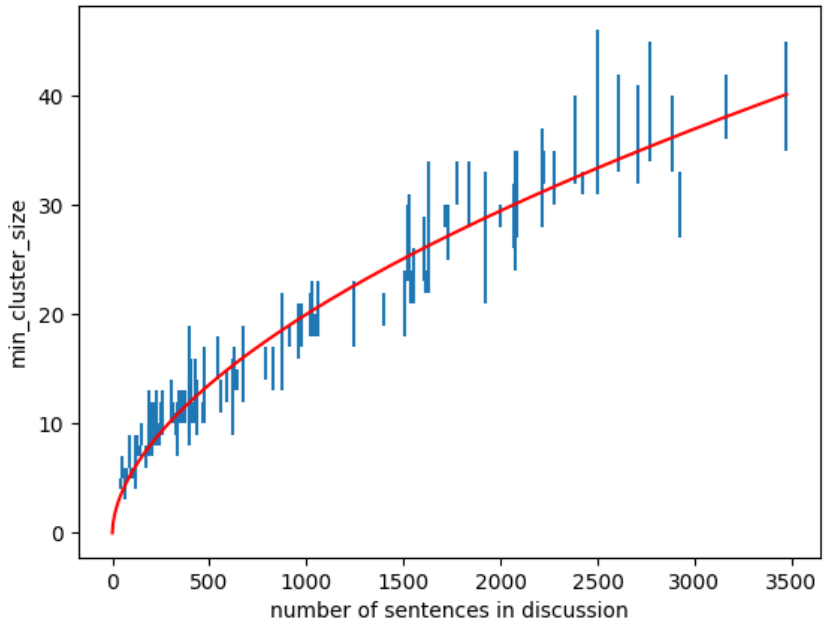


Figure A.5: Blue vertical bars show the upper and lower bound for min_cluster_size that yield a good clustering for the corresponding discussion. The red curve shows the optimal fit for the regression.

Best Prompts for Generative Cluster Labeling

T0
Read the following context and answer the question. Context: {*text*}
Question: What is the title of the discussion?
Answer:

BLOOM
AI assistant: I am an expert AI assistant and I am very good in identifying titles of debates.
DEBATE START {*text*} DEBATE END
AI assistant: The title of the debate between the two participants is "

GPT-NeoX
DISCUSSION START {*text*} DISCUSSION END
Q: What is the topic of the discussion?
A: The topic of the discussion is "

OPT-66B
DEBATE START {*text*} DEBATE END
Q: What is the topic of the debate?
A: The topic of the debate is "

GPT3.5 (text-davinci-003)
Generate a single descriptive phrase that describes the following debate in very simple language, without talking about the debate or the author.
Debate: """{*text*}"""

Figure A.6: Best prompts for generative cluster labeling for each model. These prompts were chosen based on the automatic evaluation of several prompts for each model against 300 manually annotated cluster labels.

Direct Instruction for Cluster Labeling

Every input is the content of a debate. For every input, you generate a single descriptive phrase that describes that input in very simple language, without talking about the debate or the author.

Dialogue Instruction for Cluster Labeling

A chat between a curious user and an artificial intelligence assistant. The user presents a debate and the assistant generates a single descriptive phrase that describes the debate in very simple language, without talking about the debate or the author.

Figure A.7: Direct and dialogue-style instructions for generative cluster labeling prompts. The best prompts for each model are shown in Figure A.6.

Direct Instruction for Frame Assignment

The following {*input_type*}[a] contains all available media frames as defined in the work from {*authors*}: {*frames*} For every input, you answer with three of these media frames corresponding to that input, in order of importance.

[a] A list of frame labels or a JSON with frame labels and their descriptions.

Dialogue Instruction for Frame Assignment

A chat between a curious user and an artificial intelligence assistant. The assistant knows all media frames as defined by ... : {*frames*}. The assistant answers with three of these media frames corresponding to the user's text, in order of importance.

Figure A.8: Best performing instructions for frame assignment. Providing the citation for the frame inventory via the placeholder {*authors*} positively affects the performance of some models (Appendix A.6.1).

A.6 Assigning Frames to Cluster Labels

Model Descriptions We categorize the models according to the instruction style followed for finetuning and generation. Instructions for each type are shown in Figure A.8. The best prompts for each model are listed in Figure A.9.

Direct Instruction Models

1. **LLaMA-COT**[4] is a finetuned model on datasets inducing chain-of-thought and logical deductions [238].
2. **Alpaca** [292] is finetuned from the LLaMA 7B model [301] using 52K self-instructed instruction-following examples [326].
3. **OASST** [5] is finetuned from LLaMA 30B on the OpenAssistant Conversations dataset [165] using reinforcement learning.
4. **Pythia** [31] is a suite of LLMs trained on public data to study the impact of training and scaling on various model properties. We used the 12B variant finetuned on the OpenAssistant Conversations dataset [165].
5. **GPT*** includes models such as *text-davinci-003*, *gpt-3.5-turbo* (ChatGPT), and GPT-4 [47] from the OpenAI API. These models are not open-source but have demonstrated state-of-the-art performance across various tasks.

Dialogue Instruction Models

1. **LLaMA** [301] is a suite of open-source LLMs trained on public datasets. We utilized the 30B and 65B variants.
2. **Vicuna** [61] is finetuned from LLaMA using user-shared conversations collected from ShareGPT.[6] It has shown competitive performance when evaluated using GPT-4 as a judge. We used the 7B and 13B variants of this model.
3. **Baize** [331] is an open-source chat model trained on 100k dialogues generated by allowing ChatGPT (GPT 3.5-turbo) to converse with itself. We used the 7B and 13B variants of this model.

[4] https://huggingface.co/ausboss/llama-30b-supercot
[5] https://huggingface.co/OpenAssistant/oasst-rlhf-2-llama-30b-7k-steps-xor
[6] https://sharegpt.com/

Prompt	Falcon-40B		ChatGPT		LLaMA-65B	
	Cite.	–	Cite.	–	Cite.	–
Zero-Shot	46.5	34.2	60.9	60.1	53.1	44.4
Zero-Shot (short)	46.5	42.8	58.0	57.2	50.6	42.4
Zero-Shot (full)	46.1	46.5	58.8	60.9	39.5	39.1
Few-Shot	38.3	39.1	63.4	64.6	–	–

Table A.6: Analysis of the impact of providing citation of the media frames corpus paper as additional information in the instructions for the frame assignment. Providing citation information (**Cite.**) shows up to 12% improvement for **Falcon-40B** and 9% for **LLaMA-65B** under zero-shot setting (with only frame labels in the prompt).

4. **Falcon**[7] is trained on the RefinedWeb dataset [228], which is derived through extensive filtering and deduplication of publicly available web data. It is currently the state-of-the-art (at the time of writing) on the open-llm-leaderboard.[8] We utilized the 40B and 40B-Instruct variants of this model.

A.6.1 Citation Impact on Frame Assignment

We conducted additional experiments to evaluate the impact of providing the citation of the media frames corpus paper by Boydstun et al. [40] as additional information in the instructions shown in Section 7.2.3. This piece of information was provided after the substring "defined by" in the prompt template. Table A.6 shows the results. We note that providing the citation information has a positive impact on the performance of the models. The improvement is up to 12% for **Falcon-40B** and 9% for **LLaMA-65B** under zero-shot setting (only frame labels without descriptions in the prompt). This improvement can be attributed to the models being trained on a large text corpus, with the citation serving as a strong signal for generating more accurate labels. However, **ChatGPT** is only slightly affected.

[7] https://falconllm.tii.ae/

[8] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Best Prompts for Frame Assignment

→ **Alpaca-7B** (Direct Instruction)
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that completes the request.
```
### Instruction:
{instruction}
### Input:
{input}
### Response:
```
→ **Vicuna-7B, 13B** (Dialogue Instruction)
```
{instruction}
USER: {input}
ASSISTANT:
```
→ **Pythia, OASST** (Direct Instruction)
```
<|system|>{instruction}<|endoftext|>
<|prompter|>{input}<|endoftext|><|assistant|>
```
→ **LLaMA-30B, 65B** (Dialogue Instruction)
```
{instruction}
USER: {input}
ASSISTANT: ["
```
→ **LLaMA-CoT** (Direct Instruction)
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that completes the request.
```
### Instruction:
{instruction}
### Input:
{input}
### Response:
```
→ **Falcon-40B, Instruct** (Dialogue Instruction)
```
{instruction}
USER: {input}
ASSISTANT: ["
```
→ **Baize-7B, 13B** (Dialogue Instruction)
```
{instruction}
$[|Human|]${input}
$[|AI|]$
```
→ **GPT3.5** (Direct Instruction)
```
{instruction}
Input: """
{input}
"""
Answer:
```
→ **ChatGPT, GPT4** (Direct Instruction)
```
{
    "role": "system",
    "content": "{instruction}"
},
{
    "role": "user",
    "content": "{input}"
}
```

Figure A.9: Best prompts for frame assignment for each model. The direct and dialogue instruction to be used with each prompt is shown in Figure A.8.

Guideline for judging the quality of the clustering

Task: Given a reference text and a set of hypotheses, rank the hypotheses based on how similar they are to the reference text.

How similar are the small texts to the reference text?

Drag and drop the boxes with the texts on the left and bring them in your preferred order on the right. The most preferred text is on the top and the less you prefer a text, the lower it should be in the ranking.

Similarity is less in a sense of exact meaning but much rather in a meaning of is there some relation between the reference and hypotheses.

To get a better understanding of the meaning of the reference, the title of the original discussion and some central sentences from the cluster are provided (click the "show cluster" button next to the reference). The central sentences are selected based on how central they are in the original cluster and their mean similarity to the reference and hypotheses. So these are not perfectly representative to the cluster, but they can help you to get a better understanding of some hard to understand meanings.

Recommended Strategy for judging:

The relation between the reference and hypotheses is understandable:

→ only read the reference and the hypotheses

The reference is a bit weird:

→ read the title to get a better idea in what context the reference is used

The hypotheses are hard to understand:

→ read the central sentences from the cluster for more context

The relation between the reference and hypotheses are not clear:

→ read the central and random sentences from the cluster

Note: We are looking for a label that sufficiently describes the content of a cluster of sentences. It is important to understand that the reference is not the perfect label but rather strongly representative of the cluster.

When a lot of hypotheses talk about something that is not in the reference, it is sensible to include this information in the reference (implicitly) to make it "complete" while ranking the hypotheses.

Example:

Reference: responsibilities between employee and employer

Majority of the given hypotheses mention: "the service industry"

Updated Reference: responsibilities between employee and employer in the service industry

In the end we are looking for the central meaning of the cluster and it is very likely that at least one model got the central meaning right and the task is to guess what model got the central meaning best based on what the reference suggests the best central meaning is.

Table A.7: Guideline for judging the quality of the clustering.

Model	ZS			ZS(*short*)			ZS(*full*)			Few-Shot		
	1	2	3	1	2	3	1	2	3	1	2	3
Alpaca-7B	39.1	53.9	64.2	39.5	51.0	64.6	28.4	37.4	57.2	20.6	26.7	49.4
BLOOM	26.7	46.5	53.5	31.7	52.7	57.6	25.5	51.9	60.1	–	–	–
Baize-13B	42.4	53.5	58.4	48.1	59.3	63.4	42.0	53.5	60.5	39.5	46.5	49.4
Baize-7B	34.2	44.4	52.7	34.6	46.9	56.8	39.1	46.5	53.9	30.9	38.3	45.7
Falcon-40B	46.5	68.3	72.0	46.5	67.5	75.7	46.1	56.8	64.2	38.3	53.5	68.3
Falcon-40B-Inst.	51.4	64.6	72.8	44.4	56.4	68.3	32.9	44.9	57.6	28.4	49.4	63.8
ChatGPT	60.9	76.1	86.4	58.0	78.6	88.5	58.8	76.1	84.8	63.4	80.2	90.1
GPT-4	63.4	82.3	91.8	60.5	84.4	90.1	65.4	83.1	90.5	67.1	84.8	88.5
GPT-NeoX	19.3	28.4	50.6	25.1	31.3	51.9	31.3	36.6	50.2	31.3	39.5	49.0
LLaMA-30B	45.7	63.0	70.8	41.2	57.2	65.4	39.1	58.0	66.3	40.7	70.0	77.8
LLaMA-CoT	46.9	73.3	84.0	54.3	75.7	85.6	49.8	71.2	82.3	57.2	70.0	77.0
LLaMA-65B	53.1	65.4	81.9	50.6	70.8	82.3	39.5	64.6	78.6	–	–	–
OASST	48.6	72.8	82.3	48.1	66.3	76.5	53.5	73.7	82.7	47.7	65.0	79.8
OPT-66B	16.0	18.9	43.2	13.2	16.5	45.3	14.8	18.1	45.7	–	–	–
Pythia	31.7	44.0	52.3	33.3	43.6	49.4	30.5	39.1	44.9	29.6	34.2	38.7
T0++	48.6	58.4	64.2	54.3	60.1	65.4	55.6	59.7	63.8	49.8	52.3	53.5
GPT3.5	53.5	74.1	81.9	60.9	65.4	66.7	58.0	58.8	59.7	53.9	57.6	58.0
Vicuna-13B	44.0	52.7	62.1	40.7	55.1	67.1	42.0	53.1	64.6	38.3	50.2	60.1
Vicuna-7B	28.4	34.6	50.2	36.2	48.1	61.3	35.4	42.8	55.1	20.2	24.3	46.1

Table A.8: Complete results of automatic evaluation for the frame assignment task. Shown are the **% of examples** where the first, second, and third predicted frames by a model are one of the reference frames. For the zero-shot setting (**ZS**), values are shown for each of the prompt type: only frame label (ZS), label with ***short*** description (ZS(short)), and label with ***full*** description (ZS(full)). Missing values are model inferences that exceeded our computational resources.

Prompt Templates for T0

prefix

```
What {output_type} would you choose for the
↪  {input_type} below?
{text}
```

postfix

```
{text}
What {output_type} would you choose for the
↪  {input_type} above?
```

prefix-postfix

```
What {output_type} would you choose for the
↪  {input_type} below?
{text}
What {output_type} would you choose for the
↪  {input_type} above?
```

short

```
{input_type}:
{text}
{output_type}:
```

explicit

```
{input_type} START
{text}
{input_type} END
{output_type} OF THE {input_type}:
```

question answering

```
Read the following context and answer the question.
Context:
{text}
Question: What is the {output_type} of the
↪  {input_type}?
Answer:
```

Table A.9: Prompt templates for T0 model for generative cluster labeling.

Prompt Templates for BLOOM, GPT-NeoX, OPT, GPT3.5

dialogue

```
AI assistant: I am an expert AI assistant. How can I
↪  help you?
Human: Can you tell me what the {output_type} of the
↪  following {input_type} is?
{input_type} START
{text}
{input_type} END
AI assistant: The {output_type} of the {input_type} is
↪  "
```

explicit

```
{input_type} START
{text}
{input_type} END
{output_type} of the {input_type}: "
```

assistant solo

```
AI assistant: I am an expert AI assistant and I am
↪  very good in identifying {output_type} of debates.
{input_type} START
{text}
{input_type} END
AI assistant: The {output_type} of the {input_type}
↪  between the two participants is "
```

question answering

```
{input_type} START
{text}
{input_type} END
Q: What is the {output_type} of the {input_type}?
A: The {output_type} of the {input_type} is "
```

GPT3.5

```
Generate a single descriptive phrase that describes
↪  the following debate in very simple language,
↪  without talking about the debate or the author.
Debate: """{text}"""
```

Table A.10: Prompt templates investigated for generative cluster labeling with the four decoder-only models. The `input_type` is either "title" or "topic" and the `output_type` is either "debate" or "discussion".

Prompt Templates for Instruction-following LLMs

GPT3.5

```
{instruction}
Input: """{input}"""
Answer:
```

Alpaca-7B, LLaMA-CoT

```
Below is an instruction that describes a task, paired
↪  with an input that provides further context. Write
↪  a response that appropriately completes the
↪  request.
### Instruction:
{instruction}
### Input:
{input}
### Response:
```

Baize-13B, Baize-7B

```
{instruction}
[|Human|]{input}
[|AI|]
```

BLOOM, Falcon-40B, Falcon-40B-Instruct, GPT-NeoX, LLaMA-30B, LLaMA-65B, OPT-66B, Vicuna-13B, Vicuna-7B

```
{instruction}
USER: {input}
ASSISTANT:
```

OASST, Pythia

```
<|system|>{instruction}<|endoftext|>
<|prompter|>{input}<|endoftext|><|assistant|>
```

T0++

```
{instruction}
Input: {input}
```

Table A.11: Prompt templates for instruction-following model.

CMV: The "others have it worse" argument is terrible and should never be used in an actual conversation with a depressed person
Indicative Summary (LLaMA-CoT)

Health & Safety

- Depression is a complex mental health issue that varies in severity and treatment options. **[98]** (Policy Prescription & Evaluation)
- Impact of depression and how to help those affected. **[35]** (Morality)
- Personal journey of overcoming depression and finding happiness. **[17]** (Quality of Life)

Morality

- Gratitude and appreciation for the little things in life can help improve happiness and perspective. **[39]** (Quality of Life)
- Perspective and its importance in life. **[22]** (Fairness & Equality)
- Positive self-talk and growth mindset **[21]** (Fairness & Equality)

Policy Prescription & Evaluation

- Comparing situations to others' can be helpful or harmful. **[97]** (Morality)
- Effectiveness of advice in different situations **[25]** (Capacity & Resources)
- Psychology and the power of the brain to reprogram thought patterns. **[22]** (Morality)

Table A.12: Indicative Summary from LLaMA-CoT.

CMV: The "others have it worse" argument is terrible and should never be used in an actual conversation with a depressed person
Indicative Summary (GPT3.5)

Fairness & Equality

- Complexities of comparing one's own struggles to those of others. [97] (Quality of Life)
- Advice can be helpful or unhelpful depending on how it is used. [25] (Morality)
- Focusing on personal goals and eliminating negative self-talk to create a growth mindset. [21] (Quality of Life)

Health & Safety

- How to help those with depression. [35] (Quality of Life)

Morality

- Differences between sadness and depression. [98] (Quality of Life)
- Reflecting on blessings and practicing gratitude to increase happiness. [39] (Quality of Life)
- Mindful awareness and reprogramming of thought patterns to take charge of emotions. [22] (Quality of Life)
- Gaining perspective to appreciate life and understand how one's actions affect others. [22] (Fairness & Equality)
- A journey of self-discovery and growth through difficult times. [17] (Quality of Life)

Table A.13: Indicative Summary from GPT3.5.

CMV: The "others have it worse" argument is terrible and should never be used in an actual conversation with a depressed person
Indicative Summary (GPT4)

Fairness & Equality

- Acknowledging personal struggles while recognizing others' hardships [97] (Quality of Life)

Health & Safety

- Understanding and managing depression as a complex mental state [98] (Quality of Life)
- Importance of gratitude for happiness and mental health [39] (Quality of Life)
- Impact of different approaches to supporting depressed individuals. [35] (Quality of Life)
- Controlling and reprogramming thought patterns through mindful awareness and rational evaluation of emotions. [22] (Quality of Life)

Policy Prescription & Evaluation

- Effectiveness of advice depends on individual and context. [25] (Quality of Life)

Quality of Life

- Gaining perspective for personal growth and understanding. [22] (Morality)
- Focusing on positive mindset and self-growth [21] (Health & Safety)
- Overcoming challenges and finding happiness through personal growth and change. [17] (Morality)

Table A.14: Indicative Summary from GPT4.

CMV: Today is the best time period in human history to be alive for the vast majority of people.
Indicative Summary (LLaMA-CoT)

Capacity & Resources

- The importance of having a private space for studying and building projects. **[33]** (Quality of Life)

Crime & Punishment

- Crime rates have changed over time. **[60]** (Security & Defense)

Cultural Identity

- Nostalgia for the 90s **[82]** (Quality of Life)

Economic

- Housing affordability is a complex issue with many factors at play. **[203]** (Capacity & Resources)
- Global poverty has decreased significantly over the past few decades. **[135]** (Capacity & Resources)
- Global trends and perspectives **[48]** (Policy Prescription & Evaluation)

Health & Safety

- AIDS pandemic was more fatal than the current one. **[97]** (Capacity & Resources)
- Current mental health epidemic and its causes. **[32]** (Capacity & Resources)

Policy Prescription & Evaluation

- Climate change is a serious issue that needs to be addressed. **[113]** (Economic)
- Concentration of military and economic power in history. **[47]** (Economic)

Quality of Life

- Best time period in human history to be alive. **[72]** (Economic)
- Quality of Life vs Expectations: Happiness Debate **[68]** (Other)
- Progress and improvement in society and culture **[51]** (Cultural Identity)
- Impact of technology on human connection and fulfillment. **[39]** (Cultural Identity)
- Middle Ages vs. Modern Times: Quality of Life Comparison **[29]** (Cultural Identity)

Security & Defense

- Statistics and data points in a debate about safety and progress **[43]** (Health & Safety)

Table A.15: Indicative Summary from LLaMA-CoT.

CMV: Today is the best time period in human history to be alive for the vast majority of people.
Indicative Summary (GPT3.5)

Crime & Punishment

- Violent crime rate has significantly decreased since the 1990s, but still remains an issue. [60] (Fairness & Equality)

Cultural Identity

- A constant flow of information and societal changes causing a crisis of meaning. [39] (Quality of Life)

Economic

- Catastrophic climate change leading to economic and ecological collapse. [113] (Health & Safety)
- Fragmented global economic and military power. [47] (Security & Defense)

Fairness & Equality

- Housing prices have skyrocketed in the past decade, making it difficult for the average American to afford a home. [203] (Economic)
- Decrease in global poverty and hunger since the 90s, with a majority of the world population still living in poverty. [135] (Capacity & Resources)
- Differences between the 90s and the 2000s, and the effects of time periods on different generations. [82] (Quality of Life)
- Making progress towards a better world for future generations. [51] (Quality of Life)
- Throwing around statistics without meaning and misusing percentages. [43] (Policy Prescription & Evaluation)
- Room to study and compete in the job market. [33] (Capacity & Resources)
- A comparison of the lifestyles of lower-class people in the Middle Ages and modern times. [29] (Quality of Life)

Health & Safety

- Effects of pandemics on population growth and life expectancy, with a comparison to the Bubonic Plague. [97] (Quality of Life)

Morality

- Mental health crisis in the modern world and its potential causes. [32] (Quality of Life)

Political

- Strong bias towards American perspective on global issues. [48] (Cultural Identity)

Quality of Life

- Best time period in human history to be alive. [72] (Fairness & Equality)
- Balance between quality of life, expectations, and happiness, and how they relate to each other. [68] (Fairness & Equality)

Table A.16: Indicative Summary from GPT3.5.

CMV: Today is the best time period in human history to be alive for the vast majority of people.
Indicative Summary (GPT4)

Crime & Punishment

- Violent crime rates have decreased since the 90s. **[60]** (Security & Defense)

Cultural Identity

- Nostalgia for the 90s and differing opinions on the era **[82]** (Quality of Life)
- Assuming most users are American **[48]** (Public Opinion)

Economic

- Housing affordability crisis in various locations **[203]** (Quality of Life)
- Reduced global poverty and hunger rates **[135]** (Fairness & Equality)
- Concentration of military and economic power in history **[47]** (Security & Defense)

Health & Safety

- Climate change and its worsening effects on Earth and humanity. **[113]** (Quality of Life)
- Comparing pandemics and death rates throughout history **[97]** (Quality of Life)
- Mental health awareness and treatment in modern society. **[32]** (Quality of Life)

Policy Prescription & Evaluation

- Acknowledging progress while recognizing room for improvement **[51]** (Quality of Life)

Quality of Life

- Best time to be alive debate **[72]** (Economic)
- Happiness influenced by expectations and quality of life. **[68]** (Economic)
- Misunderstanding and misuse of statistics **[43]** (Policy Prescription & Evaluation)
- Crisis of meaning and disconnection in modern society **[39]** (Cultural Identity)
- Importance of personal space for productivity and success **[33]** (Economic)
- Simple life in the Middle Ages vs modern lower class life **[29]** (Economic)

Table A.17: Indicative Summary from GPT4.

CMV: There shouldn't be anything other than the metric system. Indicative Summary (LLaMA-CoT)

Capacity & Resources

- Boiling and freezing points of water [154] (Quality of Life)

Economic

- Use of different size bottles in the dairy industry. [87] (Capacity & Resources)
- The cost of switching to the metric system is too high. [59] (Capacity & Resources)

Health & Safety

- Temperature ranges and weather conditions [86] (Quality of Life)
- Temperature range and clothing suggestions [63] (Quality of Life)

Policy Prescription & Evaluation

- Merits of Celsius and Fahrenheit temperature scales [283] (Constitutionality & Jurisprudence)
- Merits of the imperial and metric systems [196] (Constitutionality & Jurisprudence)
- Use of miles and feet in measuring distances [140] (Quality of Life)
- Merits of different systems of measurement [106] (Economic)
- Base 12 is better than base 10 for certain calculations. [104] (Capacity & Resources)
- Precision of measurements in inches and millimeters [75] (Quality of Life)
- Use of feet and inches for measuring height [72] (Constitutionality & Jurisprudence)
- Importance of precision in measurements [64] (Capacity & Resources)
- Metric vs. Imperial: Which system is better? [52] (Constitutionality & Jurisprudence)
- Merits of different counting systems [49] (Fairness & Equality)
- Merits of a decimal time system [48] (Economic)
- Merits of different scales and their practicality [46] (Capacity & Resources)
- Merits of different systems [42] (Economic)

Table A.18: Indicative Summary from LLaMA-CoT.

CMV: There shouldn't be anything other than the metric system. Indicative Summary (GPT3.5)

Capacity & Resources

- Over intuitive systems and their advantages. **[106]** (Quality of Life)
- For a more efficient counting system. **[104]** (Policy Prescription & Evaluation)
- Usefulness of different measurements for everyday use. **[87]** (Quality of Life)

Economic

- Legacy system rooted in society with benefits for everyday use and practical applications, but costly to transition away from. **[196]** (Fairness & Equality)
- Counting systems and their relative merits. **[49]** (Fairness & Equality)

Fairness & Equality

- Comparing the practicality of Celsius and Fahrenheit for everyday use, with no clear advantage to either. **[283]** (Quality of Life)
- Temperature scale based on water's freezing and boiling points. **[154]** (Constitutionality & Jurisprudence)
- Advantages and disadvantages of using inches and centimeters for measurements. **[75]** (Constitutionality & Jurisprudence)
- Usefulness of feet and inches for measuring human height. **[72]** (Quality of Life)
- A wide range of temperatures from chilly to hot, requiring different levels of clothing. **[63]** (Quality of Life)
- Costly transition to international standardization with little net benefit to average American. **[59]** (Economic)
- Advantages and disadvantages of the metric system. **[52]** (Constitutionality & Jurisprudence)
- Advantages and disadvantages of different scales. **[46]** (Capacity & Resources)
- Pros and cons of different systems. **[42]** (Policy Prescription & Evaluation)

Health & Safety

- Extremely cold temperatures ranging from -50C to +50C across the globe. **[86]** (Quality of Life)

Constitutionality & Jurisprudence

- Use of miles, yards, feet, and kilometers for measuring distances. **[140]** (Policy Prescription & Evaluation)
- Precision and accuracy in measurement. **[64]** (Policy Prescription & Evaluation)
- Complexities of measuring time. **[48]** (Policy Prescription & Evaluation)

Table A.19: Indicative summary from GPT3.5.

CMV: There shouldn't be anything other than the metric system. Indicative Summary (GPT4)

Capacity & Resources

- Water freezing and boiling points discussion [**154**] (Health & Safety)
- Base 12 system advantages [**104**] (Economic)
- Measurement units and their precision in various contexts [**75**] (Quality of Life)

Cultural Identity

- Preference for miles over kilometers in everyday language and distances [**140**] (Quality of Life)
- Preference based on familiarity and upbringing [**106**] (Quality of Life)
- Preference for feet and inches in measuring height [**72**] (Quality of Life)

Economic

- Costly and challenging transition to new system. [**59**] (Capacity & Resources)
- Using different counting systems and their efficiency in various situations. [**49**] (Capacity & Resources)

Health & Safety

- Temperature range discussion and its effects on daily life [**86**] (Quality of Life)
- Temperature and clothing preferences [**63**] (Quality of Life)

Quality of Life

- Comparing Celsius and Fahrenheit for everyday use [**283**] (Capacity & Resources)
- Imperial system vs. Metric system debate [**196**] (Cultural Identity)
- Metric and imperial measurements in daily life and their usefulness. [**87**] (Capacity & Resources)
- Misunderstanding precision and accuracy in measurements [**64**] (Health & Safety)
- Metric system advantages and precision debate [**52**] (Policy Prescription & Evaluation)
- Alternative time measurement systems [**48**] (Cultural Identity)
- Usefulness and subjectivity of different scales [**46**] (Fairness & Equality)
- Old system versus new system for everyday life [**42**] (Economic)

Table A.20: Indicative summary from GPT4.

CMV: Shoe sizes should be the same for both men and women
Indicative Summary (LLaMA-CoT)

Fairness & Equality

- Men and women's feet are different in size and shape. **[73]** (Policy Prescription & Evaluation)
- Differences between men's and women's shoes and the impact of unisex shoes. **[44]** (Policy Prescription & Evaluation)
- Shoe sizes vary by sex due to differences in foot shape. **[30]** (Quality of Life)
- Women with broad but small feet struggle to find shoes that fit. **[19]** (Quality of Life)
- Differences in clothing proportions for men and women **[17]** (Policy Prescription & Evaluation)

Policy Prescription & Evaluation

- Pros and cons of standardizing shoe sizes **[64]** (Economic)
- Use of different measurement systems **[21]** (Economic)

Table A.21: Indicative summary from LLaMA-CoT.

CMV: Shoe sizes should be the same for both men and women
Indicative Summary (GPT3.5)

Fairness & Equality

- Men and women's feet are differently shaped. **[73]** (Cultural Identity)
- Multiple shoe sizing systems causing confusion. **[64]** (Quality of Life)
- Gender-specific shoe design and comfort. **[44]** (Quality of Life)
- Different shoe sizes for men and women based on width and length. **[30]** (Quality of Life)
- Different measurement standards for length. **[21]** (Constitutionality & Jurisprudence)
- Wide feet struggle to find shoes that fit properly. **[19]** (Quality of Life)
- Clothing designed differently for men and women. **[17]** (Cultural Identity)

Table A.22: Indicative summary from GPT3.5.

CMV: Shoe sizes should be the same for both men and women
Indicative Summary (GPT4)

Economic

- Shoe durability and gender differences in footwear preferences [**44**] (Quality of Life)

Fairness & Equality

- Standardizing shoe sizes for everyone [**64**] (Quality of Life)
- Differences in clothing proportions for men and women [**17**] (Cultural Identity)

Health & Safety

- Differences in men's and women's feet [**73**] (Quality of Life)

Quality of Life

- Shoe sizes differ for men and women due to width and shape differences in feet. [**30**] (Fairness & Equality)
- Different measurement systems for shoe sizes [**21**] (Cultural Identity)
- Finding shoes for wide and small feet [**19**] (Fairness & Equality)

Table A.23: Indicative summary from GPT4.

CMV: Social media is the most destructive addiction in our society
Indicative Summary (LLaMA-CoT)

Economic

- Role of money in society and its impact on humanity. **[23]** (Capacity & Resources)
- Role of capitalism in society. **[20]** (Policy Prescription & Evaluation)
- Costs of running systems and offsetting those costs. **[19]** (Capacity & Resources)

Fairness & Equality

- Discrimination lawsuit against Amazon founder **[25]** (Constitutionality & Jurisprudence)

Health & Safety

- The impact of opioid addiction on individuals and society is devastating. **[107]** (Morality)

Morality

- Social media addiction vs opioid crisis **[38]** (Capacity & Resources)

Policy Prescription & Evaluation

- Pros and cons of social media and its impact on society. **[48]** (Morality)
- Measuring impact of technology on society **[35]** (Economic)
- The importance of education and community for a better world. **[26]** (Capacity & Resources)
- The impact of social media on society **[26]** (Public Opinion)
- The impact of social media on mental health is debated. **[24]** (Health & Safety)

Table A.24: Indicative summary from LLaMA-CoT.

CMV: Social media is the most destructive addiction in our society
Indicative Summary (GPT3.5)

Economic

- Complexities of money as a social construct. [23] (Fairness & Equality)
- Costly infrastructure needed to run systems. [19] (Capacity & Resources)

Fairness & Equality

- Importance of education, societal injustices, and the consequences of comparing oneself to others. [26] (Quality of Life)
- Powerful man accused of denying bathroom access to employees. [25] (Constitutionality & Jurisprudence)
- Effects of capitalism on human behavior. [20] (Economic)

Health & Safety

- Effects of social media on mental health. [24] (Quality of Life)

Morality

- Devastating consequences of opioid addiction leading to death and destruction. [107] (Health & Safety)
- Effects of social media and opioid use on mental health. [38] (Health & Safety)
- Negative effects of social media outweigh the positives, leading to a lack of critical thinking and a moral panic. [26] (Quality of Life)

Public Opinion

- Pros and cons of social media. [48] (Cultural Identity)

Quality of Life

- Measuring societal impact through quality of life and direction of society. [35] (Cultural Identity)

Table A.25: Indicative summary from GPT3.5.

CMV: Social media is the most destructive addiction in our society
Indicative Summary (GPT4)

Economic

- Money as a social construct and tool for exchange **[23]** (Quality of Life)
- Capitalism and human nature discussion **[20]** (Fairness & Equality)
- Costs and responsibilities of using resources and services **[19]** (Capacity & Resources)

Health & Safety

- Opioid crisis and its impact on individuals and society **[107]** (Quality of Life)
- Social media and opioid addiction relationship **[38]** (Quality of Life)
- Social media's impact on mental health and potential link to suicide rates. **[24]** (Quality of Life)

Constitutionality & Jurisprudence

- Lawsuit against Bezos for denying bathroom access **[25]** (Health & Safety)

Quality of Life

- Social media as a tool for connection and learning **[48]** (Capacity & Resources)
- Measuring impact through quality of life and societal direction **[35]** (Fairness & Equality)
- Improving society through better education and empathy. **[26]** (Fairness & Equality)
- Impact of social media on society and individuals **[26]** (Cultural Identity)

Table A.26: Indicative summary from GPT4.

A.6.2 Zero-Shot and Few-Shot Prompts for Frame Assignment

Zero-Shot (short)

```
[
  "economic",
  "capacity and resources",
  "morality",
  "fairness and equality",
  "legality, constitutionality and jurisprudence",
  "policy prescription and evaluation",
  "crime and punishment",
  "security and defense",
  "health and safety",
  "quality of life",
  "cultural identity",
  "public opinion",
  "political",
  "external regulation and reputation"
]
```

Zero-Shot

```
{
  "economic": {
    "description": "costs, benefits, or other financial
    ↪  implications"
  },
  "capacity and resources": {
    "description": "availability of physical, human or financial
    ↪  resources, and capacity of current systems"
  },
  "morality": { "description": "religious or ethical implications"
  ↪  },
  "fairness and equality": {
    "description": "balance or distribution of rights,
    ↪  responsibilities, and resources"
  },
  "legality, constitutionality and jurisprudence": {
    "description": "rights, freedoms, and authority of individuals,
    ↪  corporations, and government"
  },
  "policy prescription and evaluation": {
    "description": "discussion of specific policies aimed at
    ↪  addressing problems"
  },
```

```
  "crime and punishment": {
    "description": "effectiveness and implications of laws and their
    ↪  enforcement"
  },
  "security and defense": {
    "description": "threats to welfare of the individual, community,
    ↪  or nation"
  },
  "health and safety": {
    "description": "health care, sanitation, public safety"
  },
  "quality of life": {
    "description": "threats and opportunities for the individual's
    ↪  wealth, happiness, and well-being"
  },
  "cultural identity": {
    "description": "traditions, customs, or values of a social group
    ↪  in relation to a policy issue"
  },
  "public opinion": {
    "description": "attitudes and opinions of the general public,
    ↪  including polling and demographics"
  },
  "political": {
    "description": "considerations related to politics and
    ↪  politicians, including lobbying, elections, and attempts to
    ↪  sway voters"
  },
  "external regulation and reputation": {
    "description": "international reputation or foreign policy of
    ↪  the U.S."
  }
}
```

Zero-Shot (full)

```
{
  "economic": {
    "description": "The costs, benefits, or monetary/financial
    ↪  implications of the issue (to an individual, family,
    ↪  community, or to the economy as a whole)."
  },
  "capacity and resources": {
```

```
    "description": "The lack of or availability of physical,
    ↪  geographical, spatial, human, and financial resources, or
    ↪  the capacity of existing systems and resources to implement
    ↪  or carry out policy goals."
  },
  "morality": {
    "description": "Any perspective or policy objective or action
    ↪  (including proposed action) that is compelled by religious
    ↪  doctrine or interpretation, duty, honor, righteousness or
    ↪  any other sense of ethics or social responsibility."
  },
  "fairness and equality": {
    "description": "Equality or inequality with which laws,
    ↪  punishment, rewards, and resources are applied or
    ↪  distributed among individuals or groups. Also the balance
    ↪  between the rights or interests of one individual or group
    ↪  compared to another individual or group."
  },
  "legality, constitutionality and jurisprudence": {
    "description": "The constraints imposed on or freedoms granted
    ↪  to individuals, government, and corporations via the
    ↪  Constitution, Bill of Rights and other amendments, or
    ↪  judicial interpretation. This deals specifically with the
    ↪  authority of government to regulate, and the authority of
    ↪  individuals/corporations to act independently of
    ↪  government."
  },
  "policy prescription and evaluation": {
    "description": "Particular policies proposed for addressing an
    ↪  identified problem, and figuring out if certain policies
    ↪  will work, or if existing policies are effective."
  },
  "crime and punishment": {
    "description": "Specific policies in practice and their
    ↪  enforcement, incentives, and implications. Includes stories
    ↪  about enforcement and interpretation of laws by individuals
    ↪  and law enforcement, breaking laws, loopholes, fines,
    ↪  sentencing and punishment. Increases or reductions in
    ↪  crime."
  },
  "security and defense": {
```

```
    "description": "Security, threats to security, and protection of
    ↪  one's person, family, in-group, nation, etc. Generally an
    ↪  action or a call to action that can be taken to protect the
    ↪  welfare of a person, group, nation sometimes from a not yet
    ↪  manifested threat."
  },
  "health and safety": {
    "description": "Healthcare access and effectiveness, illness,
    ↪  disease, sanitation, obesity, mental health effects,
    ↪  prevention of or perpetuation of gun violence,
    ↪  infrastructure and building safety."
  },
  "quality of life": {
    "description": "The effects of a policy on individuals' wealth,
    ↪  mobility, access to resources, happiness, social structures,
    ↪  ease of day-to-day routines, quality of community life,
    ↪  etc."
  },
  "cultural identity": {
    "description": "The social norms, trends, values and customs
    ↪  constituting culture(s), as they relate to a specific policy
    ↪  issue."
  },
  "public opinion": {
    "description": "References to general social attitudes, polling
    ↪  and demographic information, as well as implied or actual
    ↪  consequences of diverging from or \"getting ahead of\"
    ↪  public opinion or polls."
  },
  "political": {
    "description": "Any political considerations surrounding an
    ↪  issue. Issue actions or efforts or stances that are
    ↪  political, such as partisan filibusters, lobbyist
    ↪  involvement, bipartisan efforts, deal-making and vote
    ↪  trading, appealing to one's base, mentions of political
    ↪  maneuvering. Explicit statements that a policy issue is good
    ↪  or bad for a particular political party."
  },
  "external regulation and reputation": {
    "description": "The United States' external relations with
    ↪  another nation; the external relations of one state with
    ↪  another; or relations between groups. This includes trade
    ↪  agreements and outcomes, comparisons of policy outcomes or
    ↪  desired policy outcomes."
  }
```

```
}
```

Few-Shot

```
{
  "economic": {
    "description": "The costs, benefits, or monetary/financial
    ↪  implications of the issue (to an individual, family,
    ↪  community, or to the economy as a whole).",
    "examples": [
      "Necessity of minimum wage laws and their effects on the labor
      ↪  market.",
      "Consequences of unregulated capitalism and the potential of a
      ↪  libertarian society.",
      "Risk-based insurance premiums determined by complex modeling
      ↪  of probability and cost factors."
    ]
  },
  "capacity and resources": {
    "description": "The lack of or availability of physical,
    ↪  geographical, spatial, human, and financial resources, or
    ↪  the capacity of existing systems and resources to implement
    ↪  or carry out policy goals.",
    "examples": [
      "Potential of biofuels as an alternative to fossil fuels.",
      "Physical fitness tests measure upper body strength and
      ↪  running ability for military service.",
      "Physical strength and endurance needed for modern combat."
    ]
  },
  "morality": {
    "description": "Any perspective or policy objective or action
    ↪  (including proposed action) that is compelled by religious
    ↪  doctrine or interpretation, duty, honor, righteousness or
    ↪  any other sense of ethics or social responsibility.",
    "examples": [
      "Fighting for the weak and vulnerable despite the odds.",
      "Victim-blaming debate on police brutality.",
      "Potential corruption of some native canadian bands and the
      ↪  need for transparency."
    ]
  },
  "fairness and equality": {
```

```
    "description": "Equality or inequality with which laws,
    ↪  punishment, rewards, and resources are applied or
    ↪  distributed among individuals or groups. Also the balance
    ↪  between the rights or interests of one individual or group
    ↪  compared to another individual or group.",
    "examples": [
      "Differences between humanism and feminism and their
      ↪  respective goals.",
      "Disparities in scholarship opportunities for minority
      ↪  students.",
      "Violent suppression of native american populations for
      ↪  centuries leading to a lack of advocacy and rights."
    ]
  },
  "legality, constitutionality and jurisprudence": {
    "description": "The constraints imposed on or freedoms granted
    ↪  to individuals, government, and corporations via the
    ↪  Constitution, Bill of Rights and other amendments, or
    ↪  judicial interpretation. This deals specifically with the
    ↪  authority of government to regulate, and the authority of
    ↪  individuals/corporations to act independently of
    ↪  government.",
    "examples": [
      "Guns acquired through legal and illegal channels for criminal
      ↪  use.",
      "Importance of the 2nd amendment and the implications of gun
      ↪  ownership in a democracy.",
      "Relevance of sexual history in rape cases."
    ]
  },
  "policy prescription and evaluation": {
    "description": "Particular policies proposed for addressing an
    ↪  identified problem, and figuring out if certain policies
    ↪  will work, or if existing policies are effective.",
    "examples": [
      "Religious scientists making major contributions to the world
      ↪  despite majority of scientists being agnostic atheists.",
      "Pros and cons of voluntary registration.",
      "Collective ownership of production for the betterment of
      ↪  society, with workers profiting from the sale of their
      ↪  labor."
    ]
  },
  "crime and punishment": {
```

```
    "description": "Specific policies in practice and their
    ↪  enforcement, incentives, and implications. Includes stories
    ↪  about enforcement and interpretation of laws by individuals
    ↪  and law enforcement, breaking laws, loopholes, fines,
    ↪  sentencing and punishment. Increases or reductions in
    ↪  crime.",
    "examples": [
      "Complexities of police shootings and race.",
      "Men are more likely to commit violent crimes than women.",
      "Punishment as a response to crime debated, with consideration
      ↪  of morality, severity, and aims."
    ]
  },
  "security and defense": {
    "description": "Security, threats to security, and protection of
    ↪  one's person, family, in-group, nation, etc. Generally an
    ↪  action or a call to action that can be taken to protect the
    ↪  welfare of a person, group, nation sometimes from a not yet
    ↪  manifested threat.",
    "examples": [
      "Protective physical self-defense in a fight.",
      "Powerful military technology making infantry obsolete in
      ↪  war.",
      "Protection of infants and mentally disabled through social
      ↪  policy."
    ]
  },
  "health and safety": {
    "description": "Healthcare access and effectiveness, illness,
    ↪  disease, sanitation, obesity, mental health effects,
    ↪  prevention of or perpetuation of gun violence,
    ↪  infrastructure and building safety.",
    "examples": [
      "Complexities of food choices and their effects on health.",
      "Potentially fatal consequences of taking too much
      ↪  acetaminophen.",
      "Encouraging healthy habits without shaming or pressuring
      ↪  people to lose weight."
    ]
  },
  "quality of life": {
    "description": "The effects of a policy on individuals' wealth,
    ↪  mobility, access to resources, happiness, social structures,
    ↪  ease of day-to-day routines, quality of community life,
    ↪  etc.",
```

```
    "examples": [
      "Differences between adults and children in terms of
      ↪  understanding and perception.",
      "Importance of extracurriculars and academics for college
      ↪  admissions.",
      "Appropriate times to yell at customer service workers."
    ]
  },
  "cultural identity": {
    "description": "The social norms, trends, values and customs
    ↪  constituting culture(s), as they relate to a specific policy
    ↪  issue.",
    "examples": [
      "Rapid shift in acceptance of homosexuality in the u.s.",
      "Collective action necessary for social progress and change.",
      "Complexities of gender identity and expression."
    ]
  },
  "public opinion": {
    "description": "References to general social attitudes, polling
    ↪  and demographic information, as well as implied or actual
    ↪  consequences of diverging from or \"getting ahead of\"
    ↪  public opinion or polls.",
    "examples": [
      "Gender roles and expectations are socially constructed and
      ↪  changing.",
      "Pros and cons of the 40-hour work week.",
      "Potential appeal of a political candidate."
    ]
  },
  "political": {
    "description": "Any political considerations surrounding an
    ↪  issue. Issue actions or efforts or stances that are
    ↪  political, such as partisan filibusters, lobbyist
    ↪  involvement, bipartisan efforts, deal-making and vote
    ↪  trading, appealing to one's base, mentions of political
    ↪  maneuvering. Explicit statements that a policy issue is good
    ↪  or bad for a particular political party.",
    "examples": [
      "Differences between right-wing and left-wing politics.",
      "Complexities of anarchy.",
      "Power struggle between branches of government."
    ]
  },
  "external regulation and reputation": {
```

```
    "description": "The United States' external relations with
    ↪  another nation; the external relations of one state with
    ↪  another; or relations between groups. This includes trade
    ↪  agreements and outcomes, comparisons of policy outcomes or
    ↪  desired policy outcomes.",
    "examples": [
      "Implications of us involvement in nato and its allies.",
      "Potential consequences of us intervention in ukraine.",
      "Conflicting opinions on us involvement in foreign affairs."
    ]
  }
}
```

Anhang B
Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

……
(Ort, Datum)

.....
(Unterschrift)

Bibliography

[1] Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2915–2925, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1290. URL https://www.aclweb.org/anthology/D19-1290.

[2] Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. Data acquisition for argument search: The args.me corpus. In *KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings*, pages 48–59, 2019. doi: 10.1007/978-3-030-30179-8_4. URL https://doi.org/10.1007/978-3-030-30179-8_4.

[3] Christopher Akiki, Odunayo Ogundepo, Aleksandra Piktus, Xinyu Zhang, Akintunde Oladipo, Jimmy Lin, and Martin Potthast. Spacerini: Plug-and-play search engines with pyserini and hugging face. *CoRR*, abs/2302.14534, 2023. doi: 10.48550/arXiv.2302.14534. URL https://doi.org/10.48550/arXiv.2302.14534.

[4] Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1324.

[5] Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A News Editorial Corpus for Mining Argumentation Strategies. In *26th International Conference on Computational Linguistics (COLING 2016)*, pages 3433–3443. Association for Computational Linguistics, December 2016. URL http://aclweb.org/anthology/C16-1324.

[6] Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. Patterns of Argumentation Strategies across Topics. In *2017 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2017*), pages 1362–1368. Association for Computational Linguistics, September 2017. URL http://aclweb.org/anthology/D17-1142.

[7] Khalid Al-Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness. In *58th Annual Meeting of the Association for Computational Linguistics* (*ACL 2020*), pages 7067–7072. Association for Computational Linguistics, July 2020. URL https://www.aclweb.org/anthology/2020.acl-main.632.

[8] Khalid Al-Khatib, Michael Völske, Shahbaz Syed, Anh Le, Martin Potthast, and Benno Stein. A New Dataset for Causality Identification in Argumentative Texts. In *24th Meeting of the Special Interest Group on Discourse and Dialogue* (*SIGDIAL*). Association for Computational Linguistics, September 2023.

[9] Milad Alshomary, Nick Düsterhus, and Henning Wachsmuth. Extractive snippet generation for arguments. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1969–1972. ACM, 2020. doi: 10.1145/3397271.3401186. URL https://doi.org/10.1145/3397271.3401186.

[10] Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. Target inference in argument conclusion generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4345, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.399. URL https://www.aclweb.org/anthology/2020.acl-main.399.

[11] Milad Alshomary, Timon Gurke, Shahbaz Syed, Philipp Heinisch, Maximilian Spliethöver, Philipp Cimiano, Martin Potthast, and Henning Wachsmuth. Key Point Analysis via Contrastive Learning and Extractive Argument Summarization. In Khalid Al-Khatib, Yufang Hou, and Manfred Stede, editors, *8th Workshop on Argument Mining* (*ArgMining 2021*) *at EMNLP*, pages 184–189. Association for Computational Linguistics, November 2021. doi: 10.18653/v1/

2021.argmining-1.19. URL https://doi.org/10.18653/v1/2021.argmining-1.19.

[12] Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. Argument Undermining: Counter-Argument Generation by Attacking Weak Premises. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (*ACL-IJCNLP 2021*), pages 1816–1827. ACL-IJCNLP, August 2021. doi: 10.18653/v1/2021.findings-acl.159. URL https://aclanthology.org/2021.findings-acl.159.

[13] Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.528. URL https://aclanthology.org/2021.emnlp-main.528.

[14] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 344–354. The Association for Computer Linguistics, 2015. URL https://doi.org/10.3115/v1/p15-1034.

[15] Chinatsu Aone, Mary Ellen Okurowski, James Gorlinsky, and Bjornar Larsen. A scalable summarization system using robust NLP. In *Intelligent Scalable Text Summarization*, 1997. URL https://aclanthology.org/W97-0711.

[16] Aristotle. *On Rhetoric: A Theory of Civic Discourse* (George A. Kennedy, Translator). Clarendon Aristotle series. Oxford University Press, translated 2007.

[17] Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani,

Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Xiangru Tang, Mike Tian-Jian Jiang, and Alexander M. Rush. Promptsource: An integrated development environment and repository for natural language prompts, 2022.

[18] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

[19] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, pages 667–674. ACM, 2008. doi: 10.1145/1390334.1390447. URL https://doi.org/10.1145/1390334.1390447.

[20] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics, 2005. URL https://aclanthology.org/W05-0909/.

[21] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1024.

[22] Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 4029–4039. Association for Computational Linguistics, 2020. URL https://www.aclweb.org/anthology/2020.acl-main.371/.

[23] Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 39–49. Association for Computational Linguistics, 2020. URL https://www.aclweb.org/anthology/2020.emnlp-main.3/.

[24] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In Munmun De Choudhury, Rumi Chunara, Aron Culotta, and Brooke Foucault Welles, editors, *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 830–839. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/ICWSM/article/view/7347.

[25] Phyllis B. Baxendale. Machine-made index for technical literature - an experiment. *IBM J. Res. Dev.*, 2(4):354–361, 1958. doi: 10.1147/rd.24.0354. URL https://doi.org/10.1147/rd.24.0354.

[26] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020. URL https://arxiv.org/abs/2004.05150.

[27] Rodger Benham, Joel M. Mackenzie, Alistair Moffat, and J. Shane Culpepper. Boosting search performance using query variations. *ACM Trans. Inf. Syst.*, 37(4):41:1–41:25, 2019. doi: 10.1145/3345001. URL https://doi.org/10.1145/3345001.

[28] Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, and Pengfei Liu. Metrics also disagree in the low scoring range: Revisiting summarization evaluation metrics. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5702–5711. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/

2020.coling-main.501. URL https://doi.org/10.18653/v1/2020.coling-main.501.

[29] Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.751.

[30] Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. Summarizing online forum discussions – can dialog acts of individual messages help? In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2127–2131, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1226. URL https://aclanthology.org/D14-1226.

[31] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

[32] Yonatan Bilu, Daniel Hershcovich, and Noam Slonim. Automatic claim negation: Why, how and when. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 84–93, Denver, CO, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-0511. URL https://www.aclweb.org/anthology/W15-0511.

[33] George F Bishop. Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly*, 51(2):220–232, 1987.

[34] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715.

[35] Adriana Bolívar. The structure of newspaper editorials. In *Advances in written text analysis*, pages 290–308. Routledge, 2002.

[36] Rishi Bommasani and Claire Cardie. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096,

Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.649. URL https://www.aclweb.org/anthology/2020.emnlp-main.649.

[37] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

[38] Alexander Bondarenko, Pavel Braslavski, Michael Völske, Rami Aly, Maik Fröbe, Alexander Panchenko, Chris Biemann, Benno Stein, and Matthias Hagen. Comparative web search questions. In James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang, editors, *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 52–60. ACM, 2020. doi: 10.1145/3336191.3371848. URL https://doi.org/10.1145/3336191.3371848.

[39] Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. Overview of Touché 2022: Argument Retrieval. In Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022)*, volume 13390 of *Lecture Notes in Computer Science*, Berlin Heidelberg New York, September 2022. Springer. doi: 10.1007/978-3-031-13643-6_21. URL https://doi.org/10.1007/978-3-031-13643-6_21.

[40] Amber E. Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. Tracking the development of media frames within and across policy issues, 08 2014. URL https://homes.cs.washington.edu/~nasmith/papers/boydstun+card+gross+resnik+smith.apsa14.pdf.

[41] Arthur Bražinskas, Mirella Lapata, and Ivan Titov. Learning opinion summarizers by selecting informative reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9424–9442, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.743. URL https://aclanthology.org/2021.emnlp-main.743.

[42] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Networks*, 30(1-7):107–117, 1998. URL https://doi.org/10.1016/S0169-7552(98)00110-X.

[43] Ann L Brown and Jeanne D Day. Macrorules for summarizing texts: The development of expertise. *Journal of verbal learning and verbal behavior*, 22(1):1–14, 1983.

[44] Ann L Brown, Joseph C Campione, and Jeanne D Day. Learning to learn: On training students to learn from texts. *Educational researcher*, 10(2):14–21, 1981.

[45] Ann L Brown, Jeanne D Day, and Roberta S Jones. The development of plans for summarizing texts. *Child development*, pages 968–979, 1983.

[46] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12,*

2020, virtual, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[47] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[48] Katarzyna Budzynska, Chris Reed, Manfred Stede, Benno Stein, and Zhang He. Framing in communication: From theories to computation (dagstuhl seminar 22131). *Dagstuhl Reports*, 12(3):117–140, 2022. doi: 10.4230/DagRep.12.3.117. URL https://doi.org/10.4230/DagRep.12.3.117.

[49] Jill Burstein and Daniel Marcu. A machine learning approach for identification of thesis and conclusion statements in student essays. *Computers and the Humanities*, 37(4):455–467, 2003.

[50] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S. Weld. TLDR: extreme summarization of scientific documents. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4766–4777. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.428. URL https://doi.org/10.18653/v1/2020.findings-emnlp.428.

[51] Xiaoyan Cai, Wenjie Li, Ouyang You, and Hong Yan. Simultaneous ranking and clustering of sentences: A reinforcement approach to multi-document summarization. In Chu-Ren Huang and Dan Jurafsky, editors, *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 134–142. Tsinghua University Press, 2010. URL https://aclanthology.org/C10-1016/.

[52] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II*, volume 7819 of *Lecture Notes in Computer Science*, pages 160–172. Springer, 2013. doi: 10.1007/978-3-642-

37456-2_14. URL https://doi.org/10.1007/978-3-642-37456-2_14.

[53] Shuyang Cao and Lu Wang. CLIFF: contrastive learning for improving faithfulness and factuality in abstractive summarization. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6633–6649. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.532. URL https://doi.org/10.18653/v1/2021.emnlp-main.532.

[54] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16121.

[55] Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. The media frames corpus: Annotations of frames across issues. 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 438–444, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2072. URL https://www.aclweb.org/anthology/P15-2072.

[56] Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. IMHO fine-tuning improves claim detection. 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)*, pages 558–563, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1054. URL https://www.aclweb.org/anthology/N19-1054.

[57] Stephen Chaudoin, J Shapiro, and Dustin Tingley. Revolutionizing teaching and research with a structured debate platform1. *Journal of Political Science*, 58:1064–1082, 2017.

[58] Ping Chen, Fei Wu, Tong Wang, and Wei Ding. A semantic qa-based approach for text summarization evaluation. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4800–4807. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16115.

[59] Wei-Fan Chen, Shahbaz Syed, Benno Stein, Matthias Hagen, and Martin Potthast. Abstractive snippet generation. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1309–1319. ACM / IW3C2, 2020. doi: 10.1145/3366423.3380206. URL https://doi.org/10.1145/3366423.3380206.

[60] Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. Controlled neural sentence-level reframing of news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2683–2693, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-emnlp.228.

[61] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[62] Hyungtak Choi, Lohith Ravuru, Tomasz Dryjanski, Sunghan Rye, Donghyun Lee, Hojung Lee, and Inchul Hwang. Vae-pgn based abstractive model in multi-stage architecture for text summarization. In *TL;DR Challenge System Descriptions*, 2019.

[63] Dennis Chong and James N. Druckman. *Framing theory*, pages 103–126. Annual Review of Political Science. July 2007. ISBN 0824333101. doi: 10.1146/annurev.polisci.10.072805.103054.

[64] Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL*

HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, pages 93–98. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/n16-1012. URL https://doi.org/10.18653/v1/n16-1012.

[65] Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2748–2760. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1264. URL https://doi.org/10.18653/v1/p19-1264.

[66] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1xMH1BtvB.

[67] James Clarke and Mirella Lapata. Global inference for sentence compression: An integer linear programming approach. *J. Artif. Intell. Res.*, 31:399–429, 2008. doi: 10.1613/jair.2433. URL https://doi.org/10.1613/jair.2433.

[68] John M. Conroy and Hoa Trang Dang. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In Donia Scott and Hans Uszkoreit, editors, *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*, pages 145–152, 2008. URL https://aclanthology.org/C08-1019/.

[69] John M. Conroy and Dianne P. O'Leary. Text summarization via hidden markov models. In W. Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel, editors, *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 406–407. ACM, 2001. doi: 10.1145/383952.384042. URL https://doi.org/10.1145/383952.384042.

[70] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors, *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 758–759. ACM, 2009. doi: 10.1145/1571941.1572114. URL https://doi.org/10.1145/1571941.1572114.

[71] Hoa Trang Dang. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12, 2005.

[72] Hoa Trang Dang. Overview of duc 2006. In *Proceedings of DUC 2006*, 2006.

[73] Hoa Trang Dang. Overview of duc 2007. In *Proceedings of DUC 2007*, 2007.

[74] Dipanjan Das and André Martins. A survey on automatic text summarization. 12 2007.

[75] Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1218. URL https://www.aclweb.org/anthology/D17-1218.

[76] Dorottya Demszky, Kelvin Guu, and Percy Liang. Transforming question answering datasets into natural language inference datasets. *CoRR*, abs/1809.02922, 2018. URL http://arxiv.org/abs/1809.02922.

[77] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. TorchMetrics - Measuring Reproducibility in PyTorch, 2022. URL https://github.com/Lightning-AI/metrics.

[78] Daniel Deutsch and Dan Roth. SacreROUGE: An open-source library for using and developing summarization evaluation metrics. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*,

pages 120–125. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.nlposs-1.17. URL https://aclanthology.org/2020.nlposs-1.17.

[79] Daniel Deutsch and Dan Roth. Understanding the extent to which content quality metrics measure the information quality of summaries. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, pages 300–309. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.conll-1.24. URL https://doi.org/10.18653/v1/2021.conll-1.24.

[80] Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Trans. Assoc. Comput. Linguistics*, 9:774–789, 2021. doi: 10.1162/tacl_a_00397. URL https://doi.org/10.1162/tacl_a_00397.

[81] Daniel Deutsch, Rotem Dror, and Dan Roth. A statistical analysis of summarization evaluation metrics using resampling methods. *Trans. Assoc. Comput. Linguistics*, 9:1132–1146, 2021. doi: 10.1162/tacl_a_00417. URL https://doi.org/10.1162/tacl_a_00417.

[82] Daniel Deutsch, Rotem Dror, and Dan Roth. Re-examining system-level correlations of automatic summarization evaluation metrics. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 6038–6052. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.442. URL https://doi.org/10.18653/v1/2022.naacl-main.442.

[83] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

[84] Robert L Donaway, Kevin W Drummey, and Laura A Mather. A comparison of rankings produced by summarization evaluation measures. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4*, pages 69–78. Association for Computational Linguistics, 2000.

[85] Yue Dong. A survey on neural network-based summarization methods. *CoRR*, abs/1804.04589, 2018. URL http://arxiv.org/abs/1804.04589.

[86] Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. Gsum: A general framework for guided neural abstractive summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4830–4842. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.384. URL https://doi.org/10.18653/v1/2021.naacl-main.384.

[87] Esin Durmus, Faisal Ladhak, and Claire Cardie. Determining relative argument specificity and stance for complex argumentative structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4630–4641, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1456. URL https://www.aclweb.org/anthology/P19-1456.

[88] Esin Durmus, He He, and Mona Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5055–5070. Association for Computational Linguistics, 2020. URL https://doi.org/10.18653/v1/2020.acl-main.454.

[89] H. P. Edmundson and R. E. Wyllys. Automatic abstracting and indexing - survey and recommendations. *Commun. ACM*, 4(5):226–234, 1961. doi: 10.1145/366532.366545. URL https://doi.org/10.1145/366532.366545.

[90] Charlie Egan, Advaith Siddharthan, and Adam Wyner. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 134–143, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2816. URL https://www.aclweb.org/anthology/W16-2816.

[91] Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Challenge or Empower: Revisiting Argumentation Quality in a News Editorial Corpus. In *22nd Conference on Computational Natural Language Learning (CoNLL 2018)*, pages 454–464. Association for Computational Linguistics, October 2018. URL http://aclweb.org/anthology/K18-1044.

[92] Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.287.

[93] Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. Using mechanical turk to create a corpus of arabic summaries. 2010.

[94] Robert M Entman. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, 390:397, 1993.

[95] Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger, and Ido Dagan. Proposition-level clustering for multi-document summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1765–1779, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.128. URL https://aclanthology.org/2022.naacl-main.128.

[96] Matan Eyal, Tal Baumel, and Michael Elhadad. Question answering as an automatic evaluation metric for news article summarization. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*

2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (*Long and Short Papers*), pages 3938–3948. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1395. URL https://doi.org/10.18653/v1/n19-1395.

[97] Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409, 2021. URL https://transacl.org/ojs/index.php/tacl/article/view/2563.

[98] Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2214–2220. Association for Computational Linguistics, 2019. URL https://doi.org/10.18653/v1/p19-1213.

[99] Julie Firmstone. Editorial journalism and newspapers' editorial opinions, 03 2019. URL https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-803.

[100] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *CoRR*, abs/2302.04166, 2023. doi: 10.48550/arXiv.2302.04166. URL https://doi.org/10.48550/arXiv.2302.04166.

[101] Tanvir Ahmed Fuad, Mir Tafseer Nayeem, Asif Mahmud, and Yllias Chali. Neural sentence fusion for diversity driven abstractive multi-document summarization. *Comput. Speech Lang.*, 58:216–230, 2019. doi: 10.1016/j.csl.2019.04.006. URL https://doi.org/10.1016/j.csl.2019.04.006.

[102] Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. GO FIGURE: A meta evaluation of factuality in summarization. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of

Findings of ACL, pages 478–487. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.42. URL https://doi.org/10.18653/v1/2021.findings-acl.42.

[103] Slavko Gajevic. Journalism and formation of argument. *Journalism*, 17(7):865–881, 2016.

[104] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, jan 2017. ISSN 15737462. doi: 10.1007/s10462-016-9475-9. URL http://link.springer.com/10.1007/s10462-016-9475-9.

[105] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics* (*Coling 2010*), pages 340–348, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-1039.

[106] Yang Gao, Wei Zhao, and Steffen Eger. SUPERT: towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1347–1354. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.124. URL https://doi.org/10.18653/v1/2020.acl-main.124.

[107] Ruth Garner. Efficient text summarization costs and benefits. *The Journal of Educational Research*, 75(5):275–279, 1982.

[108] Ruth Garner and Joseph L McCaleb. Effects of text manipulations on quality of written summaries. *Contemporary Educational Psychology*, 10(2):139–149, 1985.

[109] Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. Bottom-up abstractive summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4098–4109. Association for Computational Linguistics, 2018. URL https://doi.org/10.18653/v1/d18-1443.

[110] Sebastian Gehrmann, Zachary M. Ziegler, and Alexander M. Rush. Generating abstractive summaries with finetuned language models. In Kees van Deemter, Chenghua Lin, and Hiroya Takamura, editors, *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 516–522. Association for Computational Linguistics, 2019. URL https://aclweb.org/anthology/papers/W/W19/W19-8665/.

[111] Dan Gillick and Yang Liu. Non-expert evaluation of summarization systems is risky. In Chris Callison-Burch and Mark Dredze, editors, *Proceedings of the 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, USA, June 6, 2010*, pages 148–151. Association for Computational Linguistics, 2010. URL https://www.aclweb.org/anthology/W10-0722/.

[112] Tim Gollub, Matthias Busse, Benno Stein, and Matthias Hagen. Key-queries for clustering and labeling. In Shaoping Ma, Ji-Rong Wen, Yiqun Liu, Zhicheng Dou, Min Zhang, Yi Chang, and Wayne Xin Zhao, editors, *Information Retrieval Technology - 12th Asia Information Retrieval Societies Conference, AIRS 2016, Beijing, China, November 30 - December 2, 2016, Proceedings*, volume 9994 of *Lecture Notes in Computer Science*, pages 42–55. Springer, 2016. doi: 10.1007/978-3-319-48051-0_4. URL https://doi.org/10.1007/978-3-319-48051-0_4.

[113] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of GPT-3. *CoRR*, abs/2209.12356, 2022. doi: 10.48550/arXiv.2209.12356. URL https://doi.org/10.48550/arXiv.2209.12356.

[114] David Graff and C Cieri. English gigaword corpus. *Linguistic Data Consortium*, 2003.

[115] Yvette Graham. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 128–137. The Association for Computational Linguistics, 2015. doi: 10.18653/v1/d15-1013. URL https://doi.org/10.18653/v1/d15-1013.

[116] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. A large-scale dataset for

argument quality ranking: Construction and analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7805–7813. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6285.

[117] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794, 2022. doi: 10.48550/arXiv.2203.05794. URL https://doi.org/10.48550/arXiv.2203.05794.

[118] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 708–719. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1065. URL https://doi.org/10.18653/v1/n18-1065.

[119] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. doi: 10.18653/v1/p16-1154. URL https://doi.org/10.18653/v1/p16-1154.

[120] Ivan Habernal and Iryna Gurevych. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D15-1255.

[121] Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Comput. Linguistics*, 43(1):125–179, 2017. doi: 10.1162/COLI_a_00276. URL https://doi.org/10.1162/COLI_a_00276.

[122] Felix Hamborg, Anastasia Zhukova, and Bela Gipp. Illegal aliens or undocumented immigrants? towards the automated identification of bias by word choice and labeling. In Natalie Greene Taylor, Caitlin Christian-Lamb, Michelle H. Martin, and Bonnie A. Nardi, editors, *Information in Contemporary Society - 14th International Conference, iConference 2019, Washington, DC, USA, March 31 - April 3, 2019, Proceedings*, volume 11420 of *Lecture Notes in Computer Science*, pages 179–187. Springer, 2019. doi: 10.1007/978-3-030-15742-5_17. URL https://doi.org/10.1007/978-3-030-15742-5_17.

[123] Felix Hamborg, Anastasia Zhukova, and Bela Gipp. Automated identification of media bias by word choice and labeling in news articles. In Maria Bonn, Dan Wu, J. Stephen Downie, and Alaine Martaus, editors, *19th ACM/IEEE Joint Conference on Digital Libraries, JCDL 2019, Champaign, IL, USA, June 2-6, 2019*, pages 196–205. IEEE, 2019. doi: 10.1109/JCDL.2019.00036. URL https://doi.org/10.1109/JCDL.2019.00036.

[124] Sanda M Harabagiu and Finley Lacatusu. Generating single and multi-document summaries with gistexter. In *Document Understanding Conferences*, pages 11–12, 2002.

[125] Hardy Hardy, Shashi Narayan, and Andreas Vlachos. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1330. URL https://aclanthology.org/P19-1330.

[126] Victoria Chou Hare and Kathleen M. Borchardt. Direct instruction of summarization skills. *Reading Research Quarterly*, 20(1):62–78, 1984. URL http://www.jstor.org/stable/747652.

[127] Donna Harman. Overview of the second text retrieval conference (TREC-2). In Donna K. Harman, editor, *Proceedings of The Second Text REtrieval Conference, TREC 1993, Gaithersburg, Maryland, USA, August 31 - September 2, 1993*, volume 500-215 of *NIST Special Publication*, pages 1–20. National Institute of Standards and Technology (NIST), 1993. URL http://trec.nist.gov/pubs/trec2/papers/ps/overview.ps.

[128] Donna Harman and Paul Over. The effects of human variation in DUC summarization evaluation. In *Text Summarization Branches Out*,

pages 10–17, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1003.

[129] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[130] Mareike Hartmann, Tallulah Jansen, Isabelle Augenstein, and Anders Søgaard. Issue framing in online discussion fora. 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)*, pages 1401–1407, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1142. URL https://aclanthology.org/N19-1142.

[131] Philipp Heinisch and Philipp Cimiano. A multi-task approach to argument frame classification at variable granularity levels. *it - Information Technology*, 63(1):59–72, 2021. doi: doi:10.1515/itit-2020-0054. URL https://doi.org/10.1515/itit-2020-0054.

[132] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701, 2015. URL http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.

[133] Suzanne Hidi and Valerie Anderson. Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of educational research*, 56(4):473–493, 1986.

[134] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

[135] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020. doi: 10.5281/zenodo.1212303.

[136] Eduard Hovy and Chin-Yew Lin. Automated text summarization and the Summarist system. In *TIPSTER TEXT PROGRAM PHASE*

III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998, pages 197–214, Baltimore, Maryland, USA, October 1998. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/X98-1026.

[137] Eduard H. Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. Automated summarization evaluation with basic elements. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006*, pages 899–902. European Language Resources Association (ELRA), 2006. URL http://www.lrec-conf.org/proceedings/lrec2006/summaries/438.html.

[138] David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In Brian Davis, Yvette Graham, John D. Kelleher, and Yaji Sripada, editors, *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 169–182. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.inlg-1.23/.

[139] Baotian Hu, Qingcai Chen, and Fangze Zhu. LCSTS: A Large Scale Chinese Short Text Summarization Dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1967–1972. Association for Computational Linguistics, 2015. URL http://www.aclweb.org/anthology/D15-1229.

[140] Dandan Huang, Leyang Cui, Sen Yang, @inproceedingsvaswani:2017, author = Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin, editor = Isabelle Guyon and Ulrike von Luxburg and Samy Bengio and Hanna M. Wallach and Rob Fergus and S. V. N. Vishwanathan and Roman Garnett, title = Attention is All you Need, booktitle = Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages = 5998–6008, year = 2017,

url = http://papers.nips.cc/paper/7181-attention-is-all-you-need Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.33. URL https://www.aclweb.org/anthology/2020.emnlp-main.33.

[141] Ernest Hynds and Erika Archibald. Improved editorial pages can help papers, communities. *Newspaper Research Journal*, 17(1-2):14–24, 1996.

[142] Neslihan Iskender, Tim Polzehl, and Sebastian Möller. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.humeval-1.10.

[143] Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. Convex aggregation for opinion summarization. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3885–3903. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.328. URL https://doi.org/10.18653/v1/2021.findings-emnlp.328.

[144] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002. doi: 10.1145/582415.582418. URL http://doi.acm.org/10.1145/582415.582418.

[145] Hongyan Jing and Kathleen R. McKeown. Cut and paste based text summarization. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000. URL https://aclanthology.org/A00-2024.

[146] 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.

[147] Nancy S Johnson. What do you do if you can't tell the whole story? the development of summarization skills. *Children's language*, 4:315–383, 1983.

[148] Ronald E. Johnson. Recall of prose as a function of the structural importance of the linguistic units. *Journal of Verbal Learning and Verbal Behavior*, 9(1):12–20, 1970. ISSN 0022-5371. doi: https://doi.org/10.1016/S0022-5371(70)80003-2. URL https://www.sciencedirect.com/science/article/pii/S0022537170800032.

[149] K Sparck Jones et al. Automatic summarizing: factors and directions. *Advances in automatic text summarization*, pages 1–12, 1999.

[150] Karen Spärck Jones. Automatic summarising: The state of the art. *Inf. Process. Manag.*, 43(6):1449–1481, 2007. doi: 10.1016/j.ipm.2007.03.009. URL https://doi.org/10.1016/j.ipm.2007.03.009.

[151] Taeda Jovičić. Authority-based argumentative strategies: a model for their evaluation. *Argumentation*, 18(1):1–24, 2004.

[152] Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard H. Hovy. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3322–3333. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1327. URL https://doi.org/10.18653/v1/D19-1327.

[153] Min-Yen Kan, Kathleen R. McKeown, and Judith L. Klavans. Applying natural language generation to indicative summarization. In *Proceedings of the ACL 2001 Eighth European Workshop on Natural Language Generation (EWNLG)*, 2001. URL https://www.aclweb.org/anthology/W01-0813.

[154] Ryuji Kano, Yasuhide Miura, Motoki Taniguchi, Yan-Ying Chen, Francine Chen, and Tomoko Ohkuma. Harnessing popularity in social media for extractive summarization of online conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1145, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1144. URL https://aclanthology.org/D18-1144.

[155] Ryuji Kano, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. Identifying implicit quotes for unsupervised extractive summarization of conversations. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 291–302, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-main.32.

[156] Ashkan Kazemi, Verónica Pérez-Rosas, and Rada Mihalcea. Biased textrank: Unsupervised graph-based content extraction. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1642–1652. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.144. URL https://doi.org/10.18653/v1/2020.coling-main.144.

[157] Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and Vincent Ng. Give me more feedback II: Annotating thesis strength and related attributes in student essays. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3994–4004, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1390. URL https://www.aclweb.org/anthology/P19-1390.

[158] Chris Kedzie, Kathleen R. McKeown, and Hal Daumé III. Content selection in deep learning models of summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1818–1828. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1208. URL https://doi.org/10.18653/v1/d18-1208.

[159] Maurice George Kendall. Rank correlation methods. 1948.

[160] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858, 2019. URL http://arxiv.org/abs/1909.05858.

[161] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

[162] Walter Kintsch and Ely Kozminsky. Summarizing stories after reading and listening. *Journal of educational psychology*, 69(5):491, 1977.

[163] Walter Kintsch and Teun A Van Dijk. Toward a model of text comprehension and production. *Psychological review*, 85(5):363, 1978.

[164] Mike Klaas. Toward indicative discussion fora summarization. *UBC CS TR-2005*, 4, 2005.

[165] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations - democratizing large language model alignment. *CoRR*, abs/2304.07327, 2023. doi: 10.48550/arXiv.2304.07327. URL https://doi.org/10.48550/arXiv.2304.07327.

[166] Wojciech Kryscinski, Romain Paulus, Caiming Xiong, and Richard Socher. Improving abstraction in text summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1808–1817. Association for Computational Linguistics, 2018. URL https://doi.org/10.18653/v1/d18-1207.

[167] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 540–551. Association for Computational Linguistics, 2019. URL https://doi.org/10.18653/v1/D19-1051.

[168] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9332–9346. Association for Computational Linguistics, 2020. URL https://doi.org/10.18653/v1/2020.emnlp-main.750.

[169] Julian Kupiec, Jan O. Pedersen, and Francine Chen. A trainable document summarizer. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995* (*Special Issue of the SIGIR Forum*), pages 68–73. ACM Press, 1995. doi: 10.1145/215206.215333. URL https://doi.org/10.1145/215206.215333.

[170] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/kusnerb15.html.

[171] Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2175–2189. Association for Computational Linguistics, 2019. URL https://doi.org/10.18653/v1/p19-1209.

[172] Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5110. URL https://aclanthology.org/W17-5110.

[173] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics, 2020. URL https://www.aclweb.org/anthology/2020.acl-main.703/.

[174] Junyi Jessy Li, Kapil Thadani, and Amanda Stent. The role of discourse units in near-extractive summarization. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*, pages 137–147. The Association for Computer Linguistics, 2016. URL https://doi.org/10.18653/v1/w16-3617.

[175] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *CoRR*, abs/2211.09110, 2022. doi: 10.48550/arXiv.2211.09110. URL https://doi.org/10.48550/arXiv.2211.09110.

[176] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

[177] Hui Lin and Vincent Ng. Abstractive summarization: A survey of the state of the art. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9815–9822. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33019815. URL https://doi.org/10.1609/aaai.v33i01.33019815.

[178] Jimmy Lin and Dina Demner-Fushman. Evaluating summaries and answers: Two sides of the same coin? In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 41–48, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0906.

[179] Fei Liu and Yang Liu. From extractive to abstractive meeting summaries: Can it be done by sentence compression? In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Short Papers*, pages 261–264. The Association for Computer Linguistics, 2009. URL https://aclanthology.org/P09-2066/.

[180] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3728–3738. Association for Computational Linguistics, 2019. URL https://doi.org/10.18653/v1/D19-1387.

[181] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. *CoRR*, abs/2303.16634, 2023. doi: 10.48550/arXiv.2303.16634. URL https://doi.org/10.48550/arXiv.2303.16634.

[182] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

[183] Elena Lloret, Laura Plaza, and Ahmet Aker. Analyzing the capabilities of crowdsourcing services for text summarization. *Language resources and evaluation*, 47:337–369, 2013.

[184] Elena Lloret, Laura Plaza, and Ahmet Aker. The challenging task of summary evaluation: an overview. *Lang. Resour. Evaluation*, 52(1): 101–148, 2018. URL https://doi.org/10.1007/s10579-017-9399-2.

[185] Annie Louis and Ani Nenkova. Automatically assessing machine summary content without a gold standard. *Comput. Linguistics*, 39 (2):267–300, 2013. URL https://doi.org/10.1162/COLI_a_00123.

[186] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165, 1958. doi: 10.1147/rd.22.0159. URL https://doi.org/10.1147/rd.22.0159.

[187] Ge Luo, Hebi Li, Youbiao He, and Forrest Sheng Bao. Prefscore: Pairwise preference learning for reference-free summarization quality assessment. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5896–5903. International Committee on Computational Linguistics, 2022. URL https://aclanthology.org/2022.coling-1.515.

[188] Craig Macdonald and Nicola Tonellotto. Declarative experimentation in information retrieval using pyterrier. In Krisztian Balog, Vinay Setty, Christina Lioma, Yiqun Liu, Min Zhang, and Klaus Berberich, editors, *ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020*, pages 161–168. ACM, 2020. doi: 10.1145/3409256.3409829. URL https://doi.org/10.1145/3409256.3409829.

[189] Potsawee Manakul and Mark J. F. Gales. Long-span summarization via local attention and content selection. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6026–6041. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.470. URL https://doi.org/10.18653/v1/2021.acl-long.470.

[190] Inderjeet Mani. Recent developments in text summarization. In *Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, November 5-10,*

2001, pages 529–531. ACM, 2001. doi: 10.1145/502585.502677. URL https://doi.org/10.1145/502585.502677.

[191] Inderjeet Mani. Summarization evaluation: An overview. In *Proceedings of the Third Second Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, NTCIR-2, Tokyo, Japan, March 7-9, 2001*. National Institute of Informatics (NII), 2001. URL http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/sum-mani.pdf.

[192] Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim. The tipster summac text summarization evaluation. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*, pages 77–85. The Association for Computer Linguistics, 1999. URL https://www.aclweb.org/anthology/E99-1011/.

[193] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008. ISBN 978-0-521-86571-5. doi: 10.1017/CBO9780511809071. URL https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf.

[194] Daniel Marcu. Improving summarization through rhetorical parsing tuning. In Eugene Charniak, editor, *Sixth Workshop on Very Large Corpora, VLC@COLING/ACL 1998, Montreal, Quebec, Canada, August 15-16, 1998*, 1998. URL https://aclanthology.org/W98-1124/.

[195] Brett AS Martin, Bodo Lang, Stephanie Wong, and Brett AS Martin. Conclusion explicitness in advertising: The moderating role of need for cognition (nfc) and argument quality (aq) on persuasion. *Journal of Advertising*, 32(4):57–66, 2003.

[196] Woodward Matthew. User-generated content statistics. https://www.searchlogistics.com/learn/statistics/user-generated-content-statistics/. Accessed: 11.08.2023.

[197] Mani Maybury. *Advances in automatic text summarization*. MIT press, 1999.

[198] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault,

editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics, 2020. URL https://doi.org/10.18653/v1/2020.acl-main.173.

[199] Tyler McDonnell, Matthew Lease, Mücahid Kutlu, and Tamer Elsayed. Why is that relevant? collecting annotator rationales for relevance judgments. In Arpita Ghosh and Matthew Lease, editors, *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, 30 October - 3 November, 2016, Austin, Texas, USA*, pages 139–148. AAAI Press, 2016. URL http://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/view/14043.

[200] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. open-source Softw.*, 2(11):205, 2017. doi: 10.21105/joss.00205. URL https://doi.org/10.21105/joss.00205.

[201] 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, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3252.

[202] Derek Miller. Leveraging BERT for extractive text summarization on lectures. *CoRR*, abs/1906.04165, 2019. URL http://arxiv.org/abs/1906.04165.

[203] Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1046. URL https://aclanthology.org/N15-1046.

[204] Fred Morstatter, Liang Wu, Uraz Yavanoglu, Steven R. Corman, and Huan Liu. Identifying framing bias in online news. *ACM Trans. Soc. Comput.*, 1(2):5:1–5:18, 2018. doi: 10.1145/3204948. URL https://doi.org/10.1145/3204948.

[205] Nona Naderi and Graeme Hirst. Classifying frames at the sentence level in news articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages

536–542, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_070. URL https://doi.org/10.26615/978-954-452-049-6_070.

[206] Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In Yoav Goldberg and Stefan Riezler, editors, *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290. ACL, 2016. doi: 10.18653/v1/k16-1028. URL https://doi.org/10.18653/v1/k16-1028.

[207] Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cícero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathy McKeown, and Bing Xiang. Entity-level factual consistency of abstractive text summarization. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2727–2733. Association for Computational Linguistics, 2021. URL https://www.aclweb.org/anthology/2021.eacl-main.235/.

[208] Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. Annotated english gigaword. *Linguistic Data Consortium, Philadelphia*, 2012.

[209] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807. Association for Computational Linguistics, 2018. URL https://doi.org/10.18653/v1/d18-1206.

[210] Paco Nathan. PyTextRank, a Python implementation of TextRank for phrase extraction and summarization of text documents, 2016. URL https://github.com/DerwenAI/pytextrank.

[211] Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In Emily M. Bender, Leon Derczynski, and Pierre

Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1191–1204. Association for Computational Linguistics, 2018. URL https://aclanthology.org/C18-1102/.

[212] Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. *Mining text data*, pages 43–76, 2012.

[213] Ani Nenkova, Rebecca J. Passonneau, and Kathleen R. McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2):4, 2007. doi: 10.1145/1233912.1233913. URL https://doi.org/10.1145/1233912.1233913.

[214] W Russell Neuman, Russell W Neuman, Marion R Just, and Ann N Crigler. *Common knowledge: News and the construction of political meaning*. University of Chicago Press, 1992.

[215] Jun-Ping Ng and Viktoria Abrecht. Better summarization evaluation with word embeddings for ROUGE. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1925–1930. The Association for Computational Linguistics, 2015. URL https://doi.org/10.18653/v1/d15-1222.

[216] Thi Nhat Anh Nguyen, Mingwei Shen, and Karen Hovsepian. Unsupervised class-specific abstractive summarization of customer reviews. In *Proceedings of The 4th Workshop on e-Commerce and NLP*, pages 88–100, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.ecnlp-1.11. URL https://aclanthology.org/2021.ecnlp-1.11.

[217] Ansong Ni, Zhangir Azerbayev, Mutethia Mutuma, Troy Feng, Yusen Zhang, Tao Yu, Ahmed Hassan Awadallah, and Dragomir R. Radev. Summertime: Text summarization toolkit for non-experts. In Heike Adel and Shuming Shi, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 329–338. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-demo.37. URL https://doi.org/10.18653/v1/2021.emnlp-demo.37.

[218] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-4009. URL https://doi.org/10.18653/v1/n19-4009.

[219] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

[220] Paul Over, Hoa Dang, and Donna Harman. Duc in context. *Information Processing & Management*, 43(6):1506–1520, 2007.

[221] Vishakh Padmakumar and He He. Unsupervised extractive summarization using pointwise mutual information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2505–2512, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.213. URL https://aclanthology.org/2021.eacl-main.213.

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

[223] Chris D Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 26(1): 171–186, 1990.

[224] Joao Palotti, Harrisen Scells, and Guido Zuccon. Trectools: an open-source python library for information retrieval practitioners involved in trec-like campaigns. SIGIR'19. ACM, 2019.

[225] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL, 2002. URL https://www.aclweb.org/anthology/P02-1040/.

[226] Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *IJCINI*, 7(1):1–31, 2013. doi: 10.4018/jcini.2013010101.

[227] Andreas Peldszus and Manfred Stede. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948. Association for Computational Linguistics, 2015. doi: 10.18653/v1/D15-1110. URL http://aclweb.org/anthology/D15-1110.

[228] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. *CoRR*, abs/2306.01116, 2023. doi: 10.48550/arXiv.2306.01116. URL https://doi.org/10.48550/arXiv.2306.01116.

[229] Georgios Petasis and Vangelis Karkaletsis. Identifying argument components through textrank. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*, 2016. URL https://www.aclweb.org/anthology/W16-2811/.

[230] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1202. URL https://doi.org/10.18653/v1/n18-1202.

[231] Maxime Peyrard. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Associ-*

ation for Computational Linguistics, pages 1059–1073, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1101.

[232] Joseph J. Pollock and Antonio Zamora. Automatic abstracting research at chemical abstracts service. *J. Chem. Inf. Comput. Sci.*, 15(4): 226–232, 1975. doi: 10.1021/ci60004a008. URL https://doi.org/10.1021/ci60004a008.

[233] Hanieh Poostchi and Massimo Piccardi. Cluster labeling by word embeddings and wordnet's hypernymy. In Sunghwan Mac Kim and Xiuzhen Jenny Zhang, editors, *Proceedings of the Australasian Language Technology Association Workshop 2018, Dunedin, New Zealand, ALTA 2018, December 10-12, 2018*, pages 66–70, 2018. URL https://aclanthology.org/U18-1008/.

[234] Cristian Popa and Traian Rebedea. BART-TL: weakly-supervised topic label generation. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1418–1425. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.121. URL https://doi.org/10.18653/v1/2021.eacl-main.121.

[235] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

[236] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. Clickbait Detection. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 2016)*, volume 9626 of *Lecture Notes in Computer Science*, pages 810–817, Berlin Heidelberg New York, March 2016. Springer. doi: 10.1007/978-3-319-30671-1_72.

[237] PurdueOWL. Journalism and journalistic writing: The inverted pyramid structure, 2019. URL http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm.

[238] Zheng Lin Qingyi Si. Alpaca-cot: An instruction fine-tuning platform with instruction data collection and unified large language models interface. https://github.com/PhoebusSi/alpaca-CoT, 2023.

[239] Minghui Qiu and Jing Jiang. A latent variable model for viewpoint discovery from threaded forum posts. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1031–1040, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1123.

[240] Dragomir R. Radev and Kathleen R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500, 1998. URL https://aclanthology.org/J98-3005.

[241] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*, 2000. URL https://aclanthology.org/W00-0403.

[242] Dragomir R. Radev, Eduard H. Hovy, and Kathleen R. McKeown. Introduction to the special issue on summarization. *Comput. Linguistics*, 28(4):399–408, 2002. doi: 10.1162/089120102762671927. URL https://doi.org/10.1162/089120102762671927.

[243] Dragomir R. Radev, Hongyan Jing, Magorzata Sty, and Daniel Tam. Centroid-based summarization of multiple documents. *Inf. Process. Manag.*, 40(6):919–938, 2004. doi: 10.1016/j.ipm.2003.10.006. URL https://doi.org/10.1016/j.ipm.2003.10.006.

[244] 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.

[245] 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. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[246] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In

Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/d16-1264. URL https://doi.org/10.18653/v1/d16-1264.

[247] Sarvesh Ranade, Jayant Gupta, Vasudeva Varma, and Radhika Mamidi. Online debate summarization using topic directed sentiment analysis. In Erik Cambria, Bing Liu, Yongzheng Zhang, and Yunqing Xia, editors, *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM 2013, Chicago, IL, USA, August 11, 2013*, pages 7:1–7:6. ACM, 2013. doi: 10.1145/2502069.2502076. URL https://doi.org/10.1145/2502069.2502076.

[248] Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), pages 131–136, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-2024.

[249] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

[250] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1054. URL https://www.aclweb.org/anthology/P19-1054.

[251] Zhaochun Ren, Jun Ma, Shuaiqiang Wang, and Yang Liu. Summarizing web forum threads based on a latent topic propagation process. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 879–884, New York, NY, USA, 2011. Association for Computing Machinery. ISBN

9781450307178. doi: 10.1145/2063576.2063703. URL https://doi.org/10.1145/2063576.2063703.

[252] Carole Rich. *Writing and reporting news: A coaching method*. Cengage Learning, 2015.

[253] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In Donna K. Harman, editor, *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST), 1994. URL http://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[254] François Role and Mohamed Nadif. Beyond cluster labeling: Semantic interpretation of clusters' contents using a graph representation. *Knowl. Based Syst.*, 56:141–155, 2014. doi: 10.1016/j.knosys.2013.11.005. URL https://doi.org/10.1016/j.knosys.2013.11.005.

[255] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *Trans. Assoc. Comput. Linguistics*, 8:264–280, 2020. URL https://transacl.org/ojs/index.php/tacl/article/view/1849.

[256] Vasile Rus and Mihai C. Lintean. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In Joel R. Tetreault, Jill Burstein, and Claudia Leacock, editors, *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, BEA@NAACL-HLT 2012, June 7, 2012, Montréal, Canada*, pages 157–162. The Association for Computer Linguistics, 2012. URL https://aclanthology.org/W12-2018/.

[257] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 379–389. The Association for Computational Linguistics, 2015. doi: 10.18653/v1/d15-1044. URL https://doi.org/10.18653/v1/d15-1044.

[258] Horacio Saggion and Guy Lapalme. Generating indicative-informative summaries with SumUM. *Computational Linguistics*, 28

(4):497–526, 2002. doi: 10.1162/089120102762671963. URL https://aclanthology.org/J02-4005.

[259] Evan Sandhaus. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752, 2008.

[260] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

[261] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

[262] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3715–3734. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.272. URL https://doi.org/10.18653/v1/2022.naacl-main.272.

[263] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova

del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022. doi: 10.48550/arXiv.2211.05100. URL https://doi.org/10.48550/arXiv.2211.05100.

[264] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.34. URL https://aclanthology.org/2021.naacl-main.34.

[265] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. On the effect of sample and topic sizes for argument mining datasets, 2022. URL https://arxiv.org/abs/2205.11472.

[266] Natalie Schluter. The limits of automatic summarisation according to ROUGE. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 41–45. Association for Computational Linguistics, 2017. doi: 10.18653/v1/e17-2007. URL https://doi.org/10.18653/v1/e17-2007.

[267] Falk Scholer, Andrew Turpin, and Mark Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft, editors, *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 1063–1072. ACM, 2011. doi: 10.1145/2009916.2010057. URL https://doi.org/10.1145/2009916.2010057.

[268] Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. Answers unite! unsupervised metrics for reinforced summarization models. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3244–3254. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1320. URL https://doi.org/10.18653/v1/D19-1320.

[269] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. Questeval: Summarization asks for fact-based evaluation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6594–6604. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.529. URL https://doi.org/10.18653/v1/2021.emnlp-main.529.

[270] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1099. URL https://doi.org/10.18653/v1/P17-1099.

[271] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.704. URL https://doi.org/10.18653/v1/2020.acl-main.704.

[272] Holli A. Semetko and Patti M. Valkenburg Valkenburg. Framing European politics: A Content Analysis of Press and Television News. *Journal of Communication*, 50(2):93–109, 01 2006. ISSN 0021-9916. doi: 10.1111/j.1460-2466.2000.tb02843.x. URL https://doi.org/10.1111/j.1460-2466.2000.tb02843.x.

[273] Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond K. Wong, and Fang Chen. Summarization evaluation in the absence of human model summaries using the compositionality of word embeddings. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 905–914. Association for Computational Linguistics, 2018. URL https://www.aclweb.org/anthology/C18-1077/.

[274] Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Ido Dagan, and Yael Amsterdamer. Interactive query-assisted summarization via deep reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2551–2568, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.184. URL https://aclanthology.org/2022.naacl-main.184.

[275] 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.

[276] Sam Shleifer and Alexander M Rush. Pre-trained summarization distillation. *arXiv preprint arXiv:2010.13002*, 2020.

[277] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, 1996. doi: 10.1109/VL.1996.545307.

[278] Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar, October 2014. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D14-1006.

[279] Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In Jan Hajic and Junichi Tsujii, editors, *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1501–1510. ACL, 2014. URL https://www.aclweb.org/anthology/C14-1142/.

[280] Julius Steen and Katja Markert. How to evaluate a summarizer: Study design and statistical analysis for manual linguistic quality evaluation. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1861–1875. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.160. URL https://doi.org/10.18653/v1/2021.eacl-main.160.

[281] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. *CoRR*, abs/2009.01325, 2020. URL https://arxiv.org/abs/2009.01325.

[282] Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. How to compare summarizers without target length? pitfalls, solutions and reexamination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2303. URL https://aclanthology.org/W19-2303.

[283] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

[284] Shahbaz Syed, Michael Völske, Nedim Lipka, Benno Stein, Hinrich Schütze, and Martin Potthast. Towards summarization for social media - results of the tl;dr challenge. In Kees van Deemter, Chenghua Lin, and Hiroya Takamura, editors, *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 523–528. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-8666. URL https://aclanthology.org/W19-8666/.

[285] Shahbaz Syed, Roxanne El Baff, Johannes Kiesel, Khalid Al Khatib, Benno Stein, and Martin Potthast. News editorials: Towards sum-

marizing long argumentative texts. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5384–5396. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.470. URL https://doi.org/10.18653/v1/2020.coling-main.470.

[286] Shahbaz Syed, Khalid Al Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. Generating informative conclusions for argumentative texts. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3482–3493. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.306. URL https://doi.org/10.18653/v1/2021.findings-acl.306.

[287] Shahbaz Syed, Tariq Yousef, Khalid Al Khatib, Stefan Jänicke, and Martin Potthast. Summary explorer: Visualizing the state of the art in text summarization. In Heike Adel and Shuming Shi, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 185–194. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-demo.22. URL https://doi.org/10.18653/v1/2021.emnlp-demo.22.

[288] Shahbaz Syed, Dominik Schwabe, and Martin Potthast. SUMMARY WORKBENCH: unifying application and evaluation of text summarization models. In Wanxiang Che and Ekaterina Shutova, editors, *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 - System Demonstrations, Abu Dhabi, UAE, December 7-11, 2022*, pages 232–241. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-demos.23.

[289] Shahbaz Syed, Dominik Schwabe, Khalid Al Khatib, and Martin Potthast. Indicative summarization of long discussions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2752–2788.

Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.166. URL https://doi.org/10.18653/v1/2023.emnlp-main.166.

[290] Shahbaz Syed, Timon Ziegenbein, Philipp Heinisch, Henning Wachsmuth, and Martin Potthast. Frame-oriented Summarization of Argumentative Discussions. In *24th Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics, September 2023.

[291] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao, editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 613–624. ACM, 2016. doi: 10.1145/2872427.2883081. URL https://doi.org/10.1145/2872427.2883081.

[292] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[293] Sansiri Tarnpradab, Fei Liu, and Kien A Hua. Toward extractive summarization of online forum discussions via hierarchical attention networks. In *The Thirtieth International Flairs Conference*, 2017.

[294] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 107–118. Association for Computational Linguistics, 2020. URL https://doi.org/10.18653/v1/2020.emnlp-demos.15.

[295] Simone Teufel, Jean Carletta, and Marc Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the ninth conference on European chapter of the Association for*

Computational Linguistics, pages 110–117. Association for Computational Linguistics, 1999.

[296] Paul Thomas, Gabriella Kazai, Ryen White, and Nick Craswell. The crowd is made of people: Observations from large-scale crowd labelling. In David Elsweiler, editor, *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, March 14 - 18, 2022*, pages 25–35. ACM, 2022. doi: 10.1145/3498366.3505815. URL https://doi.org/10.1145/3498366.3505815.

[297] Almer S. Tigelaar, Rieks op den Akker, and Djoerd Hiemstra. Automatic summarisation of discussion fora. *Nat. Lang. Eng.*, 16(2):161–192, 2010. doi: 10.1017/S135132491000001X. URL https://doi.org/10.1017/S135132491000001X.

[298] Kunsman Todd. 36 user-generated content statistics that you can't ignore. https://everyonesocial.com/blog/user-generated-content-statistics/. Accessed: 11.08.2023.

[299] Anastasios Tombros and Mark Sanderson. Advantages of Query Biased Summaries in Information Retrieval. In *Proceedings of SIGIR 1998*, pages 2–10, 1998.

[300] Stephen E Toulmin. *The uses of argument*. Cambridge university press, 2003.

[301] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/arXiv.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

[302] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. In Phil Blunsom, Antoine Bordes, Kyunghyun Cho, Shay B. Cohen, Chris Dyer, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, and Scott Yih, editors, *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 191–200. Association for Computational Linguistics, 2017. doi:

10.18653/v1/w17-2623. URL https://doi.org/10.18653/v1/w17-2623.

[303] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. doi: 10.18653/v1/p16-1008. URL https://doi.org/10.18653/v1/p16-1008.

[304] Teun A Van Dijk. Racism and argumentation: Race riot rhetoric in tabloid editorials. *Argumentation illuminated*, pages 242–259, 1992.

[305] Teun A Van Dijk. Opinions and ideologies in editorials. In *4th International Symposium of Critical Discourse Analysis, Language, Social Life and Critical Thought, Athens*, pages 14–16, 1995.

[306] Teun A Van Dijk et al. Recalling and summarizing complex discourse. *Text processing*, pages 49–93, 1979.

[307] Oleg V. Vasilyev and John Bohannon. Is human scoring the best criteria for summary evaluation? In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2184–2191. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.192. URL https://doi.org/10.18653/v1/2021.findings-acl.192.

[308] Oleg V. Vasilyev, Vedant Dharnidharka, and John Bohannon. Fill in the BLANC: human-free quality estimation of document summaries. *CoRR*, abs/2002.09836, 2020. URL https://arxiv.org/abs/2002.09836.

[309] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.

[310] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087.

[311] Jesse Vig, Wojciech Kryscinski, Karan Goel, and Nazneen Fatema Rajani. Summvis: Interactive visual analysis of models, data, and evaluation for text summarization. *CoRR*, abs/2104.07605, 2021. URL https://arxiv.org/abs/2104.07605.

[312] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2692–2700, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/29921001f2f04bd3baee84a12e98098f-Abstract.html.

[313] Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. Tl;dr: Mining reddit to learn automatic summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu, editors, *Proceedings of the Workshop on New Frontiers in Summarization, NFiS@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pages 59–63. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-4508. URL https://doi.org/10.18653/v1/w17-4508.

[314] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 315–323. ACM, 1998. doi: 10.1145/290941.291017. URL https://doi.org/10.1145/290941.291017.

[315] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational Argumentation Quality Assessment in Natural Language. In *15th Conference of the European Chapter of the*

Association for Computational Linguistics (*EACL 2017*), pages 176–187, April 2017. URL http://aclweb.org/anthology/E17-1017.

[316] Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5106. URL https://www.aclweb.org/anthology/W17-5106.

[317] Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an Argument Search Engine for the Web. In Kevin Ashley, Claire Cardie, Nancy Green, Iryna Gurevych, Ivan Habernal, Diane Litman, Georgios Petasis, Chris Reed, Noam Slonim, and Vern Walker, editors, *4th Workshop on Argument Mining* (*ArgMining 2017*) *at EMNLP*, pages 49–59. Association for Computational Linguistics, September 2017. URL https://www.aclweb.org/anthology/W17-5106.

[318] Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. Argumentation Synthesis following Rhetorical Strategies. In *The 27th International Conference on Computational Linguistics* (*COLING 2018*). Association for Computational Linguistics, August 2018.

[319] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the Best Counterargument without Prior Topic Knowledge. In Iryna Gurevych and Yusuke Miyao, editors, *56th Annual Meeting of the Association for Computational Linguistics* (*ACL 2018*), pages 241–251. Association for Computational Linguistics, July 2018. URL http://aclweb.org/anthology/P18-1023.

[320] Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 812–817. European Language Re-

sources Association (ELRA), 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/summaries/1078.html.

[321] Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

[322] Xiaojun Wan. An exploration of document impact on graph-based multi-document summarization. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 755–762. ACL, 2008. URL https://www.aclweb.org/anthology/D08-1079/.

[323] Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5008–5020. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.450. URL https://doi.org/10.18653/v1/2020.acl-main.450.

[324] Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. Vizseq: a visual analysis toolkit for text generation tasks. In Sebastian Padó and Ruihong Huang, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019 - System Demonstrations*, pages 253–258. Association for Computational Linguistics, 2019. URL https://doi.org/10.18653/v1/D19-3043.

[325] Lu Wang and Wang Ling. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1007. URL https://www.aclweb.org/anthology/N16-1007.

[326] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *CoRR*,

abs/2212.10560, 2022. doi: 10.48550/arXiv.2212.10560. URL https://doi.org/10.48550/arXiv.2212.10560.

[327] Mark Wasson. Using leading text for news summaries: Evaluation results and implications for commercial summarization applications. In Christian Boitet and Pete Whitelock, editors, *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, August 10-14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference*, pages 1364–1368. Morgan Kaufmann Publishers / ACL, 1998. doi: 10.3115/980691.980791. URL https://aclanthology.org/P98-2222/.

[328] Wikipedia. Abstract (summary) — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Abstract%20(summary)&oldid=1152807295, 2023. [Online; accessed 09-May-2023].

[329] Peter N Winograd. Strategic difficulties in summarizing texts. *Reading Research Quarterly*, pages 404–425, 1984.

[330] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[331] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*, 2023.

[332] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Reproducible ranking baselines using lucene. *ACM J. Data Inf. Qual.*, 10(4):16:1–16:20, 2018. doi: 10.1145/3239571. URL https://doi.org/10.1145/3239571.

[333] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5754–5764, 2019. URL http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.

[334] Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. Recent advances in document summarization. *Knowledge and Information Systems*, 53(2):297–336, 2017. URL https://doi.org/10.1007/s10115-017-1042-4.

[335] Tariq Yousef and Stefan Jänicke. A survey of text alignment visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27 (2):1149–1159, 2021. doi: 10.1109/TVCG.2020.3028975.

[336] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html.

[337] Amy X. Zhang, Lea Verou, and David R. Karger. Wikum: Bridging discussion forums and wikis using recursive summarization. In Charlotte P. Lee, Steven E. Poltrock, Louise Barkhuus, Marcos Borges, and Wendy A. Kellogg, editors, *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017*, pages 2082–2096. ACM, 2017. doi: 10.1145/2998181.2998235. URL https://doi.org/10.1145/2998181.2998235.

[338] Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. Pretraining-based natural language generation for text summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 789–797, Hong Kong, China, November 2019. As-

sociation for Computational Linguistics. doi: 10.18653/v1/K19-1074. URL https://www.aclweb.org/anthology/K19-1074.

[339] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR, 2020. URL http://proceedings.mlr.press/v119/zhang20ae.html.

[340] Shiyue Zhang, David Wan, and Mohit Bansal. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. *CoRR*, abs/2209.03549, 2022. doi: 10.48550/arXiv.2209.03549. URL https://doi.org/10.48550/arXiv.2209.03549.

[341] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

[342] Yang Zhang, Yunqing Xia, Yi Liu, and Wenmin Wang. Clustering sentences with density peaks for multi-document summarization. In Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar, editors, *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1262–1267. The Association for Computational Linguistics, 2015. doi: 10.3115/v1/n15-1136. URL https://doi.org/10.3115/v1/n15-1136.

[343] Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. 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)*, pages 1298–1308, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1131. URL https://www.aclweb.org/anthology/N19-1131.

[344] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In Kentaro Inui,

Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 563–578. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1053. URL https://doi.org/10.18653/v1/D19-1053.

[345] Zheng Zhao, Shay B. Cohen, and Bonnie Webber. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.203.

[346] Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. Modeling Appropriate Language in Argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4344–4363. Association for Computational Linguistics, July 2023. doi: 10.18653/v1/2023.acl-long.238. URL https://aclanthology.org/2023.acl-long.238.pdf.

[347] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL http://arxiv.org/abs/1909.08593.

[348] Markus Zopf. Estimating summary quality with pairwise preferences. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1687–1696. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1152. URL https://doi.org/10.18653/v1/n18-1152.