MeetingBank: A Benchmark Dataset for Meeting Summarization

Yebowen Hu,[†] Tim Ganter,[‡] Hanieh Deilamsalehy,[‡] Franck Dernoncourt,[‡]
Hassan Foroosh,[†] Fei Liu[§]

[†]University of Central Florida [‡]Adobe Research [§]Emory University
`huye@knights.ucf.edu` `hassan.foroosh@ucf.edu`
`{ganter,deilamsa,franck.dernoncourt}@adobe.com` `fei.liu@emory.edu`

Abstract

As the number of recorded meetings increases, it becomes increasingly important to utilize summarization technology to create useful summaries of these recordings. However, there is a crucial lack of annotated meeting corpora for developing this technology, as it can be hard to collect meetings, especially when the topics discussed are confidential. Furthermore, meeting summaries written by experienced writers are scarce, making it hard for abstractive summarizers to produce sensible output without a reliable reference. This lack of annotated corpora has hindered the development of meeting summarization technology. In this paper, we present *MeetingBank*, a new benchmark dataset of city council meetings over the past decade. *MeetingBank* is unique among other meeting corpora due to its divide-and-conquer approach, which involves dividing professionally written meeting minutes into shorter passages and aligning them with specific segments of the meeting. This breaks down the process of summarizing a lengthy meeting into smaller, more manageable tasks. The dataset provides a new testbed of various meeting summarization systems and also allows the public to gain insight into how council decisions are made. We make the collection, including meeting video links, transcripts, reference summaries, agenda, and other metadata, publicly available to facilitate the development of better meeting summarization techniques.[1]

1 Introduction

An astonishing 55 million meetings happen in the U.S. each week (Flynn, 2022). With the extensive use of video conferencing software, e.g., Microsoft Teams, Google Meet and Zoom, it has become easier than ever before to record meetings. While these recordings provide a wealth of human intelligence and actionable knowledge, the temporal nature of sound makes it difficult for users to navigate and search for specific content (Bengio and Bourlard, 2004). A summarization system that produces text summaries from transcripts can help, by providing users with great flexibility in navigating recordings, including but not limited to: meetings, interviews, podcasts, lectures, movies and TV series (Papalampidi et al., 2020; Zhu et al., 2021; Song et al., 2022; Chen et al., 2022; Cho et al., 2022).

Effective meeting summarization requires annotated datasets. Most summarizers, including few-shot and prompt-based (Goyal et al., 2022), will benefit directly from benchmark datasets containing hundreds of thousands of document-summary pairs such as XSum (Narayan et al., 2018), Multi-News (Fabbri et al., 2019), GovReport (Huang et al., 2021), PLoS (Goldsack et al., 2022). However, datasets for meeting summarization are relatively scarce, small, or unrepresentative. ICSI and AMI are two benchmark datasets (Janin et al., 2003; Carletta et al., 2006) that consist of only 75 and 140 meetings, respectively. Other existing datasets for meetings are developed for speech recognition, or are in languages other than English and do not have reference summaries (Tardy et al., 2020; Kratochvil et al., 2020).

Creating an annotated meeting dataset poses several challenges. First, meetings often contain confidential or proprietary information, making it difficult to share them publicly. Moreover, accurately annotating meeting summaries is a labor-intensive process, even for experienced writers familiar with the meeting topics (Renals et al., 2007). Effective meeting summaries should capture key issues discussed, decisions reached, and actions to be taken, while excluding irrelevant discussions (Zechner, 2002; Murray et al., 2010). There thus is a growing need for innovative approaches to construct a meeting dataset with minimal human effort to support advanced meeting solutions.

An increasing number of *city governments* are releasing their meetings publicly to encourage trans-

[1]Our dataset can be accessed at: `meetingbank.github.io`

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 16409–16423

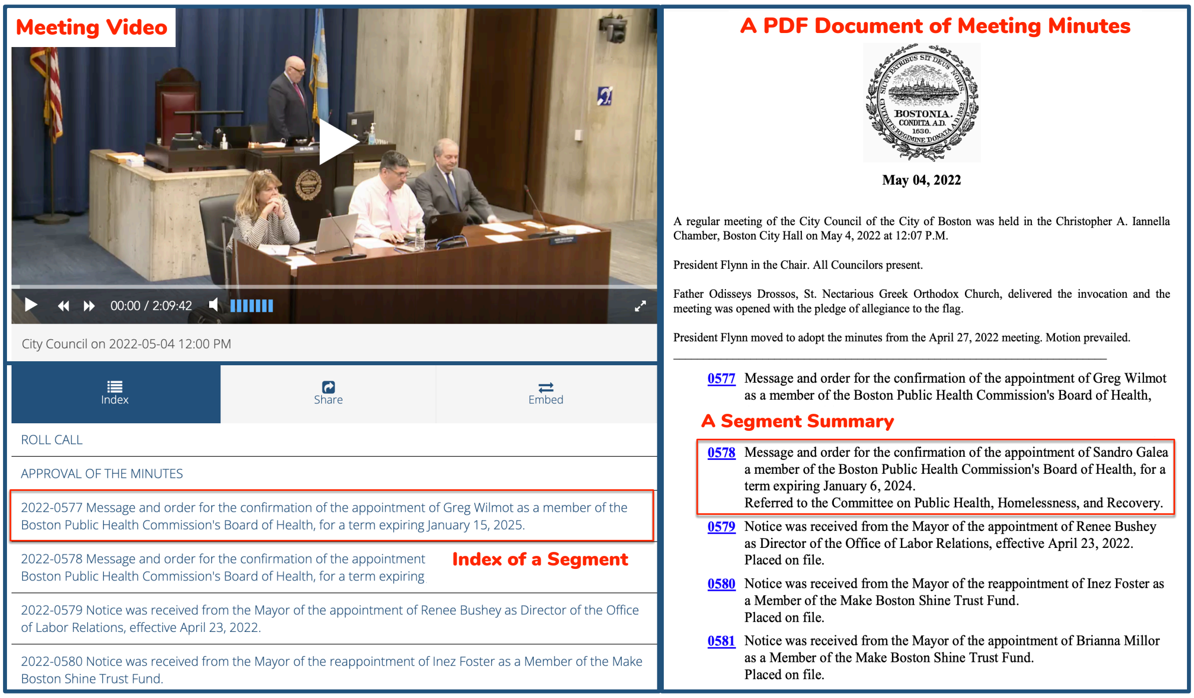

Figure 1: A screenshot of a city council meeting of the City of Boston held on May 4, 2022. The meeting video is shown on the left, its corresponding minutes document on the right. The meeting includes discussions of multiple ordinances and resolutions. A summary of the discussion on item 2022-0578 is highlighted in red.

parency and engage residents in their decision making process. In this paper, we present a systematic approach to develop ***a city council meeting dataset***. A city council is the legislative branch of local government. The council members are responsible for making decisions on a range of issues that affect the city and its citizens. These decisions may include creating and approving annual budgets, setting tax rates, confirming appointments of officers, enacting and enforcing ordinances, and setting policies on issues such as land use, public safety, and community development. Figure 1 provides an example of a regular meeting of the City Council of Boston held on May 4, 2022.

We present ***MeetingBank***, a benchmark dataset created from the city councils of 6 major U.S. cities to supplement existing datasets. It contains 1,366 meetings with over 3,579 hours of video, as well as transcripts, PDF documents of meeting minutes, agenda, and other metadata. It is an order of magnitude larger than existing meeting datasets (Carletta et al., 2006). On average, a council meeting is 2.6 hours long and its transcript contains over 28k tokens, making it a valuable testbed for meeting summarizers and for extracting structure from meeting videos. To handle the max sequence length constraint imposed by abstractive summarizers, we introduce a *divide-and-conquer strategy* to divide lengthy meetings into segments, align these segments with their respective summaries from minutes documents, and keep the segments simple for easy assembly of a meeting summary. This yields 6,892 segment-level summarization instances for training and evaluating of performance. Our repository can be further enhanced through community efforts to add annotations such as keyphrases and queries (Zhong et al., 2021). To summarize, this paper presents the following contributions:

- We have curated a repository of city council meetings, *MeetingBank*, to advance summarization in an understudied domain. We detail our process of examining 50 major U.S. cities, accessing their city councils' websites for meeting videos and minutes, and obtaining permission to use their data for research purposes. As more cities participate in open government initiatives and release their council meetings, MeetingBank has the potential to continue growing.
- We test various summarizers including extractive, abstractive with fine-tuning, and GPT-3 with prompting on this task. They are provided with the transcript of a meeting segment and is tasked with generating a concise summary. Experiments with automatic metrics and expert annotators sug-

gest that meeting summarizers should prioritize capturing the main points of meeting discussions and maintaining accuracy to the original.

2 Existing Datasets

In this section, we review existing meeting datasets, discuss the techniques used to create reference summaries for them and identify research challenges that require attention in this area.

ICSI and **AMI** are widely used datasets for meetings. ICSI (Janin et al., 2003) is a benchmark of 75 meetings that occurred naturally among speech researchers in a group seminar setting. Each meeting lasts approximately an hour. AMI (Carletta et al., 2006) contains 100 hours of recorded meetings, including 140 scenario-based meetings where groups of four participants assume roles within a fictitious company to design a product. Meetings typically last around 30-40 minutes, with roles including a project manager, user interface designer, production designer, and marketing expert. A wide range of annotations are performed by annotators, including speech transcriptions, dialogue acts, topics, keywords, extractive and abstractive summaries. Although small in size, these datasets offer a valuable testbed for evaluating meeting summarization systems (Wang and Cardie, 2013; Oya et al., 2014; Shang et al., 2018; Li et al., 2019; Koay et al., 2020, 2021; Zhang et al., 2022).

Our study complements recent datasets for meetings such as **ELITR** and **QMSum**. Nedoluzhko et al. (2022) developed ELITR, a dataset of 120 English and 59 Czech technical project meetings spanning 180 hours of content. Their minutes are created by participants and specially-trained annotators. Zhong et al. (2021) developed the QMSum system which extracts relevant utterances from transcripts and then uses the utterances as input for an abstractor to generate query-focused summaries. Human annotators are recruited to collect queries and compose summaries. They annotate a limited number of 25 committee meetings from the Welsh Parliament, 11 from the Parliament of Canada, as well as AMI and ICSI meetings for query-focused summarization.

Summarization datasets have been developed for genres similar to meetings, such as podcasts, interviews, livestreams and TV series. **Spotify** (Clifton et al., 2020) released a dataset of 100,000 podcasts to support podcast search and summarization. The dataset includes automatic transcripts with word-level time alignments and creator-provided podcast descriptions are used as reference summaries. **MediaSum** (Zhu et al., 2021) is a dataset of media interviews containing 463.6k transcripts from NPR and CNN, with overview and topic descriptions used as reference summaries. **StreamHover** (Cho et al., 2021) used crowd workers to annotate 370 livestream videos for both extractive and abstractive summaries. **SummScreen** (Chen et al., 2022) consists of TV series transcripts and human-written recaps extracted from community websites. Unlike meetings, these datasets have a smaller number of participants, usually one or two, and do not involve decision making.

Meetings can also occur through online chats, as opposed to face-to-face. **SAMSum** (Gliwa et al., 2019) is a dataset of 16k chat dialogs with manually annotated abstractive summaries. The conversations are short, ranging from 3 to 30 utterances. **ForumSum** (Khalman et al., 2021) is another dataset that includes online posts collected from inter-

Speaker 4:	Thank you. Next item, please.
Speaker 0:	Item 18 Report from Human Resources. Recommendation to purchase access workers compensation insurance for a total premium amount not to exceed 505,134 citywide.
Speaker 4:	Any public comment?
Speaker 0:	No public comment on this item.
Speaker 4:	OC motion, but counter appearances are a second. Taken my customers and they asked Customer Pearce, do you want to? He said, You want to. Short staff update.
Speaker 1:	Yes. I'm not sure which which item it is if it's 18, 19 or 20, but I just. Was hoping to get a brief staff report on on the insurance.
Speaker 4:	On which item?
Speaker 1:	We'll do it on 18 is fine.
Speaker 4:	Okay
Speaker 7:	Alex Vasquez will get the step forward.
Speaker 0:	Good evening, Mayor and city council. I'm going to turn it over to Jolene Richardson.
Speaker 1:	She's our risk manager and she'll give a brief overview of this particular report. Even the mayor and council. This is for the city's annual renewal, for the excess workers compensation insurance, which is important for us to continue to provide coverage for our employees. It also helps us to reduce our negative financial consequences for our high exposures or losses that may result from injuries or deaths due to accidents, fire or terrorist attacks and earthquakes during work hours. This coverage will be obtained through the city's casualty.
Speaker 0:	Broker for a record.
Speaker 1:	Alliant Insurance Services. This year's policy for excess workers compensation will continue to provide 150 million and coverage access of 5 million self-insured retention at a premium of $505,134, which represents an increase of approximately 6.6% from the expiring policy due to increase in city's payroll. I think if there's any questions, we'd be happy to answer ...

Reference Summary: Recommendation to authorize City Manager, or designee, to purchase, through Alliant Insurance Services, excess workers' compensation insurance with Safety National Casualty Corporation, for a total premium amount not to exceed $505,134, for the period of July 1, 2020 through July 1, 2021.

Figure 2: An example of a transcript snippet for a meeting segment, which serves as the source text for our summarizer. Similar to BillSum (Kornilova and Eidelman, 2019), a short description of the discussed bill serves as the segment-level reference summary. Source: Long Beach, 6/23/2022.

CITY	MEETING-LEVEL #Mtgs	#Hrs	#Tks	#Speakers	Period	INST. #Segs	SOURCE #Snts	#Tks	SUMMARY #Snts	#Tks
Denver	401	979	25,460	[3,20]	2014–22	1,506	204	5,100	3.32	111
Seattle	327	446	15,045	[3,14]	2015–22	1,497	54	1,499	1.06	78
Long Beach	310	1103	39,618	[4,19]	2014–22	2,695	146	3,826	1.90	86
Alameda	164	730	47,981	[2,15]	2015–22	672	251	6,452	2.04	67
King County	132	247	20,552	[2,10]	2016–22	223	196	5,358	1.00	78
Boston	32	72	23,291	[4,11]	2021–22	299	63	1,422	1.98	77
TOTAL COUNT: **1,366** meetings, **3,579** hours transcribed, **6,892** summarization instances collected										

Table 1: Dataset statistics. Our dataset includes a total of 1,366 city council meetings. We present the number of meetings (`#Mtgs`), their cumulative duration in hours (`#Hrs`), the average number of tokens per meeting (`#Tks`), and the number of speakers per meeting (`#Speakers`) for each city. We also provide the number of summarization instances gathered (`#Segs`) for each city, as well as the average number of sentences (`#Snts`) and tokens (`#Tks`) in both source and summary texts. On average, across all cities, a meeting has an average duration of 2.6 hours and 28k tokens. A meeting segment has 2,892 tokens in the source transcript and 87 tokens in the summary.

net forums, with associated human-written summaries. **DialogSum** (Chen et al., 2021) contains 13k dialogs gathered from multiple websites. Medical consultations conducted through online chats have also been used to create consultation summaries (Zeng et al., 2020; Laleye et al., 2020; Moramarco et al., 2021; Gao and Wan, 2022). Our paper focuses on developing a dataset of naturally-occurring meeting conversations to aid in the development of summarization systems from transcripts.

3 Creation of *MeetingBank*

There is a growing need to make public meetings more accessible and inclusive for citizens to engage with their local officials and have their voices heard. A 2020 report from the American Academy of Arts and Sciences[2] reveals that only 11% of Americans attend public meetings to discuss local issues. Organizations such as the CouncilDataProject.org and CodeforAmerica.org are working to improve the search infrastructure of public meetings. Creation of a city council meeting dataset could provide a valuable testbed for meeting summarizers, and an effective summarizer could make public meetings more accessible by allowing citizens to navigate meetings more efficiently, thus promoting citizen engagement.

We begin by compiling a list of the top 50 cities in the U.S. by population[3] and narrow it down to include only cities that regularly release meetings with accompanying minutes documents, and have downloadable videos on their city council websites. An example of a public meeting and its minutes can be seen in Figure 1. We consult with our legal team and reach out to city councils when necessary to ensure compliance with licensing and data policies.[4] Our dataset for this release includes **1,366 meetings** from six cities or municipalities spanning over a decade, including *Seattle, Washington; King County, Washington; Denver, Colorado; Boston, Massachusetts; Alameda, California; and Long Beach, California*.

Using minutes documents as-is for the development of summarization systems can be challenging. This is because minutes are often provided in PDF format and do not always align with the flow of meeting discussions. For instance, minutes may include a section on the Mayor's update that provides detailed information on appointments of officers, but this is only briefly mentioned in the meeting. In general, *minutes* are more formal and comprehensive records of meetings, including information such as the date, location, attendees, summary of main points discussed, decisions made, and action items assigned. Minutes are distributed to the stakeholders after the meeting. In contrast, *meeting summaries* tend to be shorter and less formal, focusing on the key points discussed in a meeting.

We propose a divide-and-conquer strategy for

[2] `amacad.org/ourcommonpurpose/recommendations`

[3] `www.infoplease.com/us/cities/top-50-cities-us-population-and-rank`

[4] For example, we have excluded the City of San Francisco from our dataset as the city has advised us that meeting videos may be reposted or edited with attribution, but minutes and agenda are official public documents that are not permitted to be reposted or edited.

creating reference meeting summaries. It involves dividing lengthy meetings into segments, aligning them with their corresponding summaries from minutes documents, and keeping the segments simple for easy assembly of a meeting summary. To start, we extract a list of Council Bill (CB) numbers discussed at the meeting by parsing the minutes and city council websites.[5] For each bill, we then identify a short description that summarizes its content, which serves as the reference summary. Next, we use the bill number to obtain the corresponding meeting segment, including its start and end time, by referencing the index of the meeting on the city council website (see Figure 1). The transcript of that segment serves as the source text for the summarizer. After filtering out noisy and too short segments,[6] we have a total of **6,892 segment-level instances** in our dataset (Table 1).

We use Speechmatics.com's speech-to-text API to automatically transcribe 3,579 hours of meetings, an order of magnitude larger than existing datasets. Our transcripts include word-level time alignment, casing, punctuation, and speaker diarization. City council meetings range from 2 to 19 speakers, with an average duration of 2.6 hours and 28,358 tokens per transcript. On average, across all cities, a meeting segment has 2,892 tokens in the transcript and 87 tokens in the summary. The resulting compression rate is 97%. For every council meeting, we collect the following information, represented using their attribute name and sample value:[7]

1. Title of the meeting ("*Full Council 12/14/15*")
2. Meeting ID ("*SeattleCityCouncil_12142015*")
3. Link to the specific meeting (`https://www.seattlechannel.org/FullCouncil?videoid=x60447&Mode2=Video`)
4. Link to the meeting video (`https://video.seattle.gov/media/council/full_121415V.mp4`)
5. Link to the meeting minutes (`https://seattle.legistar.com/View.ashx?M=M&ID=449835&GUID=712D0B7C-A536-498E-8C99-D3037AE814D9`)
6. ID of a specific topic discussed ("*CB 118549*")
7. Type of the ID number ("*Ordinance*")
8. Start and end times in the video where the topic was discussed ("*00:06:24*" to "*00:18:19*")
9. Full transcript of the video, along with start and end points of each segment of the meeting (Figure 2)
10. Reference summary for each meeting segment

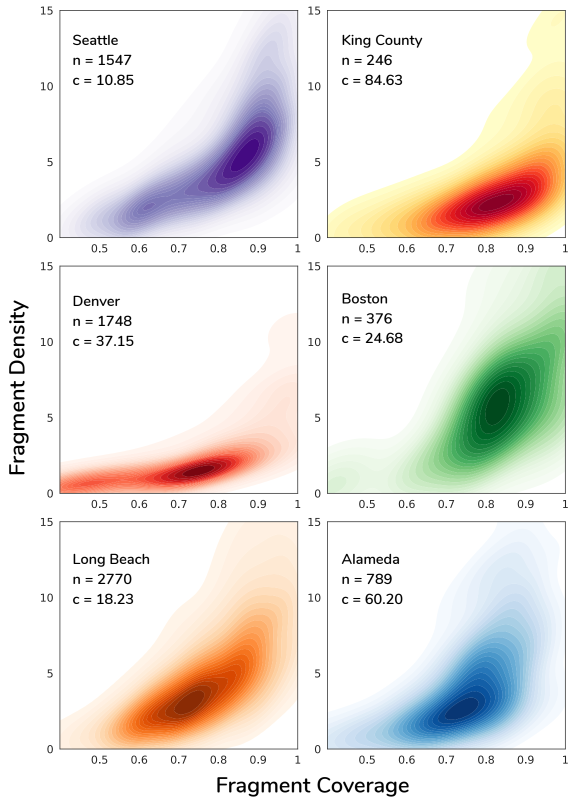

Figure 3: *Coverage* and *Density* scores for segment-level summarization instances, plotted for individual cities. Seattle and Boston have the highest density scores among the cities studied, while Denver has the lowest, indicating that the minutes for this city have undergone a high degree of editing.

4 Data Analysis

We measure the level of abstraction in meeting summaries by quantifying the amount of reused text. Higher abstraction poses more challenges to the meeting summarization systems. We employ two common measures, *Coverage* and *Density* (Grusky et al., 2018), to evaluate segment-level reference summaries. Results are illustrated in Figure 3, with coverage on x-axis and density on y-axis.

The *Coverage* score measures the percentage of summary words that appear in the source transcript. E.g., a summary of 10 words that includes 8 words from the source transcript and 2 new words has a coverage score of 0.8 (= 8/10). As shown in the figure, the *Coverage* score for city council meeting summaries is in the range of 0.7-0.9 for most cities.

[5] E.g., `boston.legistar.com/MeetingDetail.aspx?ID=958849&GUID=CAD14B15-407D-4552-AF01-4BD64314AD2D`

[6] We require a minimum length of 60 seconds for a segment to be included in our dataset, as segments shorter than this are too brief to be summarized. The reference summary for a segment should contain at least 10 words.

[7] We gather the meeting agenda and other supporting documents when available. Our focus is on summarizing transcripts of spontaneous speech where natural language is the primary means of information conveyance. We do not attempt to obtain non-verbal cues such as eye gazes, facial expressions, laughter, or understand persuasive argumentation.

	ROUGE				BLEU + MET.		EMBEDDINGS		QA	SUMM
MODEL	R-1	R-2	R-L	R-We	BLEU	METEOR	BERTS.	MoverS.	QAEval	LEN.
Extr Oracle	61.82	46.60	52.61	55.60	22.99	52.35	69.54	63.15	21.69	64.89
Lead-3	28.15	19.53	25.75	23.77	7.90	23.53	50.20	54.56	9.62	40.79
LexRank	24.61	10.68	19.06	15.98	5.86	17.70	48.55	53.23	6.53	53.70
TexRank	30.25	15.97	24.37	21.91	9.16	22.10	52.32	54.65	8.33	61.81
BART w/o FT	31.02	16.76	23.93	23.11	8.07	16.63	53.04	53.91	13.63	140.65
HMNet	50.55	34.22	45.07	45.05	13.93	46.80	66.38	60.34	12.96	50.44
Longformer	59.89	48.23	55.66	56.15	40.04	50.92	75.31	65.27	23.54	82.86
BART	62.81	51.66	58.84	59.32	41.46	53.24	77.17	66.74	26.87	89.46
Pegasus	68.54	59.28	66.09	65.75	33.29	**70.24**	80.70	70.44	**27.13**	49.90
DialogLM	**70.30**	**60.12**	**67.54**	**67.55**	**45.42**	66.44	**81.61**	**71.56**	25.75	66.36
GPT3-D3	36.37	16.95	26.82	26.14	8.80	25.41	56.53	55.61	10.88	60.41

Table 2: Evaluation of state-of-the-art summarization systems on the test split of our city council dataset. The final column shows the average length of system summaries, which are generated by each individual summarizer using their default settings.

This suggests that, unlike news, these meeting summaries tend to include discussion points verbatim rather than performing abstraction. Given a compression rate of over 90%, an effective summarizer should focus on accurately identifying content to be included in the summary.

A $\mathcal{Density}$ score evaluates how much a summary can be characterized as a set of extractive fragments. For example, a summary of 10 words made up of two extractive fragments of length 1 and 6 and three new words would have a density score of 3.7 = $(1^2 + 6^2)/10$. A summary with long consecutive text fragments taken from the source transcript would yield a high density score. We observe that Seattle and Boston have the highest density scores among all cities studied, while Denver has the lowest score, indicating a high degree of editing is performed to produce the minutes for Denver. We note that certain resolutions and ordinances are read out plainly at the council meetings and included in the minutes, making the summaries often have higher density scores than those of news documents.

The $\mathcal{Coverage}$ and $\mathcal{Density}$ measures can be influenced by a range of factors such as the length and complexity of meetings and the preferences of the minute-takers. The diversity of meeting summaries highlights the complexity of this task.

5 Performance of Existing Systems

We evaluate state-of-the-art summarization systems on city council meetings, focusing on segments of the meetings rather than entire transcripts due to the length constraint imposed by abstractive summarizers. We split our dataset into train, validation and test sets, containing 5169, 861, 862 instances respectively. Each summarizer is given the transcript of a meeting segment and tasked with generating a concise summary. The results are reported for the test set of our meeting dataset.

Extractive. Our extractive methods include the `Oracle`, `LEAD`, `LexRank` and `TextRank` (Erkan and Radev, 2004; Mihalcea and Tarau, 2004). The `Extractive Oracle`[8] selects the highest-scoring sentences from the input transcript, adding one sentence at a time until the combined R1 and R2 score can no longer be improved. The `LEAD-N` baseline selects the first N sentences of the input. `LexRank` and `TextRank`, both graph-based methods, determine the importance of sentences by analyzing their centrality in the graph structure. Both methods are set to extract two sentences from a transcript segment, which is the average number of sentences in the reference summaries.

Abstractive with fine-tuning. We investigate five best-performing neural abstractive summarizers. These include `BART`-large (Lewis et al., 2020), a denoising autoencoder that is trained to reconstruct original text from corrupted input, `Pegasus` (Zhang et al., 2020a), a model that is trained to regenerate missing key sentences, `Longformer` (Beltagy et al., 2020), a model designed to handle long sequences

[8] `github.com/pltrdy/extoracle_summarization`

TRAIN	#Inst.	TEST SET (ALL) R-1	R-2	R-L	TEST SET (BY CITY) L.B.	Denver	Seattle	Alameda	Boston	K.C.
w/o L.B.	3,157	60.36	48.57	56.48	**21.30**↓	1.95↑	4.39↑	1.72↓	3.92↓	3.71↓
w/o Denver	3,951	58.74	46.90	54.31	2.84↓	**32.43**↓	1.96↓	0.22↓	0.22↓	1.64↑
w/o Seattle	4,115	**53.42**	**40.05**	**48.28**	3.62↓	3.34↑	**31.12**↓	2.19↓	1.06↓	2.79↓
w/o Alameda	4,698	61.54	50.31	57.48	2.41↓	4.07↑	3.00↑	**19.32**↓	2.54↓	1.68↓
w/o Boston	4,936	62.70	51.62	58.82	1.21↑	0.75↑	7.52↓	1.02↑	**42.82**↓	4.96↑
w/o K.C.	4,988	63.60	52.71	59.87	3.60↓	3.32↑	4.37↑	6.14↓	3.76↓	**11.6**↓

Table 3: Evaluation of the BART summarizer using a series of ablations. LEFT: we remove all the training instances from a single city and fine-tune the model with the remaining instances, denoted by `#Inst`. We find that although the City of Seattle only contributes a moderate number of training instances, removing them has led to a substantial decrease in summarization performance. RIGHT: we evaluate the performance of the BART summarizer on a city-by-city basis. We show the variance in R-2 F-scores for each test city when training instances from the same city are included vs. when they are excluded. ↓ indicates a performance drop and ↑ a performance gain.

through windowed attention, `DialogLM` (Zhong et al., 2022), a summarizer developed for summarizing long dialogues and pretrained using window-based denoising and `HMNet` (Zhu et al., 2020), a hierarchical model that uses role vectors to distinguish between speakers. We evaluate `BART`-large with and without fine-tuning on our dataset, and compare the results to other models.

GPT-3 with prompting. Large language models like GPT-3, T5 and PaLM (Brown et al., 2020; Raffel et al., 2020; Chowdhery et al., 2022) have developed advanced capabilities due to increased computation, parameters, training data size (Wei et al., 2022). When prompted, GPT-3 can generate a summary of a source text by identifying the most important information. The `text-davinci-003` version of GPT-3 is used in this study, with a prompt asking the model to summarize the text in two sentences (Goyal et al., 2022).

Evaluation Metrics. We use a variety of automatic evaluation metrics to assess the quality of transcript summaries. These metrics are broadly grouped into three categories: (a) traditional metrics comparing system and reference summaries based on lexical overlap, including ROUGE (Lin, 2004), ROUGE-we (w/ embeddings), BLEU (Post, 2018) and METEOR (Banerjee and Lavie, 2005); (b) new metrics making use of contextualized embeddings to measure semantic similarity, e.g., BertScore (Zhang et al., 2020b) and MoverScore (Zhao et al., 2019); (c) question answering-based metrics, where the hypothesis is that high-quality summaries should contain informative content and act as a surrogate for the original document in satisfying users' information needs. We leverage summarization evaluation toolkits provided by Fabbri et al.(2021), SacreBLEU (Post, 2018), QAEval (Deutsch et al., 2021) and SummerTime (Ni et al., 2021) to report results using these metrics.

Table 2 shows our experimental results. We observe that the Extractive Oracle yields a high R-2 F-score of 46.60%, indicating that the content of reference summaries mostly comes from the source transcripts, and extractive summarization methods could be promising. However, it would be desirable to develop more sophisticated methods than LexRank and TextRank, despite their outstanding performance on news articles, they do not perform well on this task. We find that DialogLM performs the best among abstractive summarizers. This is not surprising as it is designed for summarizing long dialogues. Pegasus also demonstrates strong performance, its results are on par with those of DialogLM. Fine-tuning BART on in-domain data yields substantial improvement on its performance. Finally, GPT-3 with prompting does not perform well according to automatic metrics, but we have interesting findings during human assessment (§7).[9]

6 City-by-City Analysis

We investigate the characteristics that make effective training instances for meeting summarization by conducting a series of ablations. We begin by

[9] We find that some automatic metrics are affected by the extractiveness of summaries, such as MoverScore and ROUGE, while others, such as BERTScore and QAEval, are less sensitive. Metrics that are sensitive to extractiveness give varying scores across different datasets, and those that are insensitive tend to produce scores in similar ranges.

Informativeness:	How well does the summary capture the main points of the meeting segment? A good summary should contain all and only the important information of the source.
Factuality:	Are the facts provided by the summary consistent with facts in the meeting segment? A good summary should reproduce all facts accurately and not make up untrue information.
Fluency:	Consider the individual sentences of the summary, are they well-written and grammaticall?
Coherence:	Consider the summary as a whole, does the content fit together and sound natural? A good summary should not just be a collection of related information, but should build from sentence to sentence to a coherent body of information about a topic.
Redundancy:	Does the summary contain redundant content? A good summary should not have unnecessary word or phrase repetitions in a sentence or semantically similar sentences.

Table 4: Human evaluation criteria, adapted from Fabbri et al. (2021).

removing all training instances from a single city and using the remaining instances to fine-tune the BART summarizer. The results are shown in Table 3 (left panel), where we present the R-1, R-2, and R-L F-scores. We find that although the City of Seattle only contributes a moderate number of training instances, removing them has led to a substantial decrease in summarization performance, resulting in an R-2 F-score of 40.05%. It suggests that these training instances are quite effective and the City Council of Seattle might have implemented a better practice of producing high-quality meeting minutes compared to other cities.

We evaluate the performance of the BART summarizer on a city-by-city basis. We show the variance in R-2 F-scores for each test city when training instances from the same city are included vesus when they are excluded, as seen in the right panel of Table 3. For instance, we observe a performance drop (↓) of 32.43% for the City of Denver when all training instances from the same city are removed from fine-tuning.[10] We observe that Seattle, Boston, and Denver benefit more from fine-turning using same-city training data. Particularly, Seattle and Boston have shorter source transcripts and their reference summaries tend to reuse texts from the source. It suggests that different cities may have varying levels of discussions in council meetings and different styles of meeting minutes, and that training instances from the same city are crucial for achieving the best performance.

7 Human Evaluation

We evaluate the performance of seven state-of-the-art summarization systems, including fine-tuned abstractive models `HMNet`, `BART`, `Pegasus`, `DialogLM`, `GPT-3` with prompting, and traditional extractive models `LexRank` and `LEAD` to best assess the effectiveness of system-generated meeting summaries. All abstractive models have been fine-tuned on the train split of our city councils dataset to achieve the best possible results.

To ensure high quality in the assessment of summaries, we have worked with `iMerit.net`, a labor sourcing company, to recruit experienced evaluators from the U.S. and India to perform annotations. The workers are registered on `Appen.com`, a crowdsourcing platform, to complete the tasks and deliver results. A total of three workers from the United States and six workers from India participate in our evaluations, including pilot annotations.[11]

The workers are asked to watch a video segment, typically 30 minutes or less, read the transcript, and then evaluate the quality of each system summary based on five criteria: *informativeness*, *factuality*, *fluency*, *coherence*, and *redundancy*. These criteria are outlined in Table 4. Importantly, summaries are presented in a random order to prevent workers from making assumptions about quality based on the order they are presented.

In Table 5, we present the performance of summarization systems on 200 randomly selected instances. A 5-point Likert scale is used to evaluate each criterion. The scores are then averaged, and standard deviation is also reported. We find that among the five criteria, redundancy is the least of concern. Furthermore, we observe that abstractive systems perform stronger than extractive systems. The best-performing abstractive system is `Pegasus`. We believe its effectiveness is attributed to the pre-training method of masking key sentences within

[10] We fine-tune the BART summarizers for the same number of steps in both cases to mitigate the impact of varying number of training instances.

[11] All workers have excellent English proficiency, with U.S. workers being native speakers. After a pilot annotation, we decide to work with only U.S. workers due to their high quality of work. They are compensated at $27.50/hr.

Criterion	Extractive		Abstractive w/ Finetuning				Prompting
	LEAD	**LexRank**	**HMNet**	**BART**	**DialogLM**	**Pegasus**	**GPT-3**
Informativeness	1.72±1.22	1.90±1.15	2.44±1.15	3.43±1.13	3.50±1.15	3.65±1.10	**3.74**±1.17
Factuality	1.92±1.35	2.42±1.31	2.65±1.18	3.45±1.19	3.58±1.09	3.79±1.13	**3.82**±1.09
Fluency	2.89±1.50	3.22±1.32	2.78±1.32	3.58±1.13	3.47±1.14	3.71±1.23	**4.52**±0.83
Coherence	2.14±1.34	2.70±1.39	2.74±1.39	3.64±1.15	3.72±1.13	3.88±1.13	**4.41**±1.00
Redundancy	3.79±1.38	3.53±1.40	3.42±1.40	3.54±1.41	3.96±1.35	4.15±1.25	**4.57**±0.75
Average Score	2.49	2.75	2.81	3.53	3.65	3.84	**4.21**

Table 5: Human evaluation results. We observe that abstractive systems perform stronger than extractive systems. GPT-3 is well received in human assessments, but still falls short in terms of informativeness and factuality.

a document and using the remaining sentences to regenerate them, making it particularly well-suited for this task and effective at identifying important content from the transcripts.

We find that `GPT-3` achieves the highest overall score of 4.21 according to human evaluations across all criteria. This aligns with recent studies that demonstrate GPT-3's near-human performance in news summarization (Goyal et al., 2022; Zhang et al., 2023). On our meeting dataset, `GPT-3` shows exceptional performance in terms of fluency and coherence, receiving scores of 4.52 and 4.41 respectively. However, its results are less impressive in terms of informativeness and factuality with scores of 3.74 and 3.82, but still on par with the best abstractive model, `Pegasus`. Our findings suggest that meeting summarization solutions should continue to focus on capturing the main discussion points and staying true to the original content.

8 Conclusion

We created a benchmark dataset from city council meetings and tested various summarization systems including extractive, abstractive with fine-tuning, and GPT-3 with prompting on this task. Our findings indicate that GPT-3 is well received in human assessments, but it falls short in terms of informativeness and factual consistency. Our MeetingBank dataset could be a valuable testbed for researchers designing advanced meeting summarizers and for extracting structure from meeting videos.

9 Limitations

We present a new dataset for meeting summarization that has the potential to improve the efficiency and effectiveness of meetings. However, we note that the dataset is limited to city council meetings from U.S. cities over the past decade and licensing issues have restricted our ability to include certain city council meetings in the dataset. For example, we contacted the City Council of San Francisco and were informed that they do not allow the redistribution of meeting minutes. Moreover, our dataset does not include non-verbal cues such as eye gazes, gestures and facial expressions, which may make it less suitable for developing summarization systems that rely on these cues. Despite these limitations, we believe that the dataset is of high quality and will be a valuable resource for the development of meeting summarization systems.

10 Ethical Considerations

The city council meetings included in this dataset are publicly accessible. We obtain meeting videos, minutes documents, and other metadata from publicly available sources. We consult with our legal team and reach out to city councils as necessary to ensure compliance with licensing and data policies. We release this dataset to facilitate the development of meeting summarization systems and have made efforts to ensure that the dataset does not include confidential information. Our dataset is intended for research purposes only.

Acknowledgements

We are grateful to the reviewers for their insightful feedback, which has helped enhance the quality of our paper. This research has been partially supported by the NSF CAREER award, #2303655.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Pro-*

ceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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

Samy Bengio and Hervé Bourlard. 2004. *Machine Learning for Multimodal Interaction: First International Workshop*. Springer Berlin, Heidelberg.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, and Vasilis Karaiskos et al. 2006. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39. Springer Berlin Heidelberg.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Sangwoo Cho, Franck Dernoncourt, Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, and Fei Liu. 2021. StreamHover: Livestream transcript summarization and annotation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6457–6474, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sangwoo Cho, Kaiqiang Song, Xiaoyang Wang, Fei Liu, and Dong Yu. 2022. Toward unifying text segmentation and long document summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 106–118, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, and Sebastian Gehrmann et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jack Flynn. 2022. 28 incredible meeting statistics: Virtual, zoom, in-person meetings and productivity. *https://www.zippia.com/advice/meeting-statistics/*.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. *arXiv preprint arxiv:2210.09932*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI meeting corpus. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Misha Khalman, Yao Zhao, and Mohammad Saleh. 2021. ForumSum: A multi-speaker conversation summarization dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4592–4599, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jia Jin Koay, Alexander Roustai, Xiaojin Dai, Dillon Burns, Alec Kerrigan, and Fei Liu. 2020. How domain terminology affects meeting summarization performance. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5689–5695, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jia Jin Koay, Alexander Roustai, Xiaojin Dai, and Fei Liu. 2021. A sliding-window approach to automatic creation of meeting minutes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 68–75, Online. Association for Computational Linguistics.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Jonas Kratochvil, Peter Polak, and Ondrej Bojar. 2020. Large corpus of Czech parliament plenary hearings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6363–6367, Marseille, France. European Language Resources Association.

Fréjus A. A. Laleye, Gaël de Chalendar, Antonia Blanié, Antoine Brouquet, and Dan Behnamou. 2020. A French medical conversations corpus annotated for a virtual patient dialogue system. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 574–580, Marseille, France. European Language Resources Association.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Aleksandar Savkov, and Ehud Reiter. 2021. A preliminary study on evaluating consultation notes with post-editing. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 62–68, Online. Association for Computational Linguistics.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010. Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR minuting corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3174–3182, Marseille, France. European Language Resources Association.

Ansong Ni, Zhangir Azerbayev, Mutethia Mutuma, Troy Feng, Yusen Zhang, Tao Yu, Ahmed Hassan Awadallah, and Dragomir Radev. 2021. SummerTime: Text summarization toolkit for non-experts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 329–338, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. 2014. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 45–53, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. 2020. Screenplay summarization using latent narrative structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1920–1933, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Steve Renals, Thomas Hain, and Herve Bourlard. 2007. Recognition and understanding of meetings the ami and amida projects. In *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, pages 238–247.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia. Association for Computational Linguistics.

Kaiqiang Song, Chen Li, Xiaoyang Wang, Dong Yu, and Fei Liu. 2022. Towards abstractive grounded summarization of podcast transcripts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4407–4418, Dublin, Ireland. Association for Computational Linguistics.

Paul Tardy, David Janiszek, Yannick Estève, and Vincent Nguyen. 2020. Align then summarize: Automatic alignment methods for summarization corpus creation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6718–6724, Marseille, France. European Language Resources Association.

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405, Sofia, Bulgaria. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arxiv:2301.13848*.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arxiv:2104.05938*.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.

A Experimental Settings

Our implementation details and hyperparameter settings for both extractive systems and abstractive systems with fine-tuning are shown in Table 6. We use `text-davinci-003` version of GPT-3 in our experiments. We follow the convention of Goyal et al. (2022) and use the following prompt asking the model to summarize a transcript in two sentences: *Article:{{article}}*

Summarize the above article in N sentences.

B Comparison to ICSI/AMI

By introducing a corpus, we aim to spur research and development in the area of meeting summarization. However, meetings often pertain to specialized domains and exhibit unique structures. Our preliminary experiments suggest that a BART summarizer fine-tuned for our dataset does not perform optimally on the ICSI/AMI datasets. In particular, the ICSI meetings pose a challenge as they are research seminars conducted by a group of speech researchers, whereas our dataset is collected from city councils in the U.S.

`Extractive Oracle`
We use the implementation provided by Paul Tardy: `github.com/pltrdy/extoracle_summarization` (a) "-length_oracle" sets the output to have the same number of sentences as the reference summary. (b) "-method greedy -length 999" allows the greedy algorithm to select an optimal number of sentences that yield the highest (R1+R2) scores. In this paper, we report results using option (b).
`LexRank` and **`TextRank`**
We use SummerTime's implementation of LexRank and TextRank with default parameters. `https://github.com/Yale-LILY/SummerTime` For each meeting segment, 2 sentences are extracted.
`BART`
The BART model is initialized using bart-large-cnn: The default model parameters are used, with some of the important ones listed below. - max input sequence length: 1,024 tokens - min output length: 56 tokens - max output length: 142 tokens - beam width: 4 - length penalty: 2.0 - initial learning rate: 2.5e-6
`Pegasus`
Pegasus is initialized using google/pegasus-xsum - max sequence length: 512 - max output length: 64 - beam width: 8 - length penalty: 0.6 - initial learning rate: 2.5e-6
`Longformer`
Longformer is initialized using patrickvonplaten/ longformer2roberta-cnn_dailymail-fp16 - max sequence length: 4,098 - min output length: 56 - max output length: 142 - beam width: 1 - length penalty: 1.0 - Initial Learning Rate: 2.5e-6
`HMNet`
We use the implementation of Zhu et al. (2020). HMNet is initialized using HMNet-pretrained - max sequence length: 8300 - min output length: 10 - max output length: 300 - beam width: 6 - initial learning rate: 1e-4
`DialogLM`
We use the original DialogLM source implementation. - max sequence length: 5,632 - min output length: 10 - max output length: 300 - beam width: 6 - Initial Learning Rate: 7e-5

Table 6: Implementation details and hyperparameter settings for extractive systems and abstractive systems.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 9

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 10

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 10

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 10

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 3 and 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Sections 3 and 4

C ☑ Did you run computational experiments?

Section 5

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Supplementary Materials

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 7

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Supplementary Materials

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 7

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 7 and Supplementary Materials

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 7

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.