SUMREN: Summarizing Reported Speech about Events in News

Revanth Gangi Reddy[1*], **Heba Elfardy**[2], **Hou Pong Chan**[3], **Kevin Small**[2], **Heng Ji**[2]

[1]University of Illinois Urbana-Champaign
[2]Amazon Alexa
[3]University of Macau
revanth3@illinois.edu, {helfardy,smakevin,jihj}@amazon.com, hpchan@um.edu.mo

Abstract

A primary objective of news articles is to establish the factual record for an event, frequently achieved by conveying both the details of the specified event (i.e., the 5 Ws; Who, What, Where, When and Why regarding the event) and how people reacted to it (i.e., reported statements). However, existing work on news summarization almost exclusively focuses on the event details. In this work, we propose the novel task of summarizing the reactions of different speakers, as expressed by their reported statements, to a given event. To this end, we create a new multi-document summarization benchmark, SUMREN, comprising 745 summaries of reported statements from various public figures obtained from 633 news articles discussing 132 events. We propose an automatic silver-training data generation approach for our task, which helps smaller models like BART achieve GPT-3 level performance on this task. Finally, we introduce a pipeline-based framework for summarizing reported speech, which we empirically show to generate summaries that are more abstractive and factual than baseline query-focused summarization approaches.

1 Introduction

In news, attribution occurs when the journalist reports the statements of a third party either by directly quoting them (i.e., direct quotation) or paraphrasing what they said (i.e. indirect quotation). Reported speech serves as a central resource for tracking public figures' stance, opinions, and worldviews, making it of general interest to news readers. For example, readers are likely to be interested in knowing President Biden's view on the 2022 Ukraine crisis or the latest guidance from the Center for Disease Control and Prevention regarding a new COVID-19 variant. In addition, reported statements cover a significant portion of the information presented in news articles – as part of our annotation exercise (described later in section 3.1), we found that 45% of the overall article content corresponds to reported statements. However, current news summarization datasets such as CNN-DM (Hermann et al. 2015), Multi-News (Fabbri et al. 2019), and Timeline$_{100}$ (Li et al. 2021) largely disregard summarizing these reported statements.

To bridge this gap, we introduce the new task of **Sum**marizing **R**eported speech about **E**vents in **N**ews and create a new benchmark, **SUMREN**, for this task. Formally, given a set of news articles related to a specified event, the task is to summarize the statements made by a given speaker about this event (e.g., "What did Chuck Schumer say about passing the Inflation Reduction Act of 2022?"). The aim of the task is to provide news readers with the reactions of various public figures towards different events. Table 1 shows an example from SUMREN, along with the reported statements and the corresponding reference summary.

Event: *Power Outage in Texas*	**Speaker:** *Nateghi*
Reported Statements	
An issue facing all power grid operators, Nateghi of Purdue said, is adequately preparing for changes in climate.	
They're also not taking into account inter-dependencies in the system: You need water to generate electricity, and you need electricity to transport water, and so on, Nateghi said.	
And when the system is really stressed from an extreme event like it is in Texas, then we're seeing natural gas shortages which exacerbate the whole impact, she said.	
Nateghi, who researches sustainability and resilience of infrastructure, said other solutions such as upgraded equipment and infrastructure may not be as cost-effective, but are still crucial.	
"If we continue down the paradigm of what we've done before we will see more extremes," Nateghi said. "These stories are going to just keep playing, and perhaps more frequently."	
Summary: *Nateghi said that interdependencies in the system are not being considered, and the problem of gas shortages could be seen in the power outage in Texas. Solutions such as upgraded equipment and infrastructure maybe less cost-effective but crucial. She also said that power grid operators needed to make changes before extreme situations became more frequent.*	

Table 1: An example from SUMREN showing reported statements from the speaker "Nateghi" about the "Power outage in Texas" along with the corresponding summary.

Summarizing reported speech in news brings a set of unique challenges. As opposed to traditional news summarization datasets where the most salient information about the event is normally discussed in the first few sentences of a given article, generally referred to as "*lead bias*" (Jung

*Work primarily done during an internship at Amazon Alexa.

et al. 2019; Zhu et al. 2021), reported speech from the same speaker can be scattered across the entire article. Statements can be split across multiple sentences (i.e., "running quotations") and speakers are often referred to by their nominal and pronominal mentions, requiring modelling of long-term dependencies and reliable co-reference resolution. Additionally, generating concise summaries from a set of reported statements requires a higher level of abstraction. This is also verified empirically, as we find that summaries in SUMREN are considerably more abstractive compared to existing news summarization datasets, as shown in Table 2. Finally, factual consistency is paramount in reported speech summarization, as misquoting or misrepresenting statements from public figures can be particularly harmful.

To address the above challenges, we propose a pipeline-based approach for the task of summarizing reported speech in news articles. The pipeline involves first identifying individual statements and corresponding local speaker mentions, then resolving speaker mentions globally using co-reference resolution to group statements from the same speaker together, and finally summarizing the extracted reported statements by the given speaker. We hypothesize that, in a pipeline-based framework, having an explicit extractive component that can identify relevant context helps the summarization model better attend to the key information from the given articles.

In addition, we introduce a cost-effective approach to generating training data for the reported speech summarization task. Specifically, we leverage large-scale pre-trained language models, such as GPT-3 (Brown et al. 2020), to generate silver-standard summaries for statements obtained from automatic reported speech extraction systems. This follows recent work that uses large language models to create training data (Schick and Schütze 2021), although previously explored for discriminative tasks such as Natural Language Inference. We show that training with such silver-standard data can help smaller language models, such as BART (Lewis et al. 2020) achieve GPT-3-level performance on this task.

To summarize, the contributions of this work include:

- introducing a new challenging task of summarizing reported speech about events in news and releasing the first multi-document summarization benchmark, SUMREN[1], for the task. SUMREN contains 745 instances annotated over 633 news articles discussing 132 events,
- empirically demonstrating that large-scale language models can be leveraged to create cost-efficient silver-standard training data for the reported speech summarization task,
- proposing a pipeline-based reported speech summarization framework and showing that it is capable of generating summaries that are considerably more abstractive than query-focused approaches, while also improving the factual consistency of the generated summaries with the source documents.

[1]Code and data at: https://github.com/amazon-science/sumren

2 Related Work

Our work draws from multiple related research veins as itemized in this section.

News Summarization: Summarizing news articles has been extensively studied in existing literature with multiple existing datasets. Single document summarization datasets include CNN/Daily Mail (Hermann et al. 2015), News Room corpus (Grusky, Naaman, and Artzi 2018) and the XSum dataset (Narayan, Cohen, and Lapata 2018). Fabbri et al. (2019) introduce a large-scale dataset, Multi-News, to extend news summarization to a multi-document setting. Timeline summarization (Steen and Markert 2019; Li et al. 2021) adds a temporal aspect to news summarization by generating a sequence of major news events with their key dates. Another line of work lies around news headline generation (Banko, Mittal, and Witbrock 2000), which involves generating representative headlines for a given news story, explored in both single- (Hayashi and Yanagimoto 2018) and multi-document settings (Gu et al. 2020). However, these datasets all largely focus on summarizing the event details and neglect the reported speech related to these events.

Query-Focused Summarization: Query-focused summarization (QFS) aims to produce a summary that answers a specific query about the source document(s). Conceptually, reported speech summarization corresponds to the query, *"What did X say about Y?"*. Prior work builds large-scale QFS datasets by obtaining reference summaries by scraping them from the web or using pseudo-heuristics. For example, WikiSum (Liu et al. 2018) and AQuaMuSe (Kulkarni et al. 2020) directly extract paragraphs from Wikipedia articles as reference summaries. On the other hand, manually annotated QFS datasets are small – DUC 2006 and 2007 (Dang 2005) contain up to only 50 examples. QMSum (Zhong et al. 2021b) focuses on summarizing meeting dialogue transcripts and is most similar to our work. However, QMSum transcripts contain a considerable amount of informal conversations and do not contain focused informative content like the reported statements in SUMREN.

Since QFS datasets usually come with only source-summary pairs, most prior work either use end-to-end approaches (Vig et al. 2022; Xu and Lapata 2022) or follow a two-step extract-then-abstract framework (Xu and Lapata 2021; Vig et al. 2022), with the extractor trained to identify text spans that are similar to the reference summary in terms of ROUGE scores. Conversely, SUMREN additionally provides the corresponding relevant content, reported statements in this case, that was used to annotate the summaries. Thereby, our proposed pipeline-based approach can leverage this to build and evaluate an extractive component that is independent of the reference summary, while still ensuring the generated summary has high input fidelity in terms of factual consistency.

Attribution in News: Attribution has been well-studied with multiple available datasets. Elson and McKeown (2010); Zhang and Liu (2021) study attribution of direct quotations along with their speakers. Pareti (2012); Pareti et al. (2013) extend this notion by including indirect quota-

Step 1: Identifying salient spans in statements
While Republicans look inward at the aftermath of the Capitol Hill riots after President Trump's address Wednesday, Democrats are adding to the division, Fox News contributor Charles Hurt told "Fox & Friends."
"I get this rush to want to blame everything on President Trump. Everything that is going on right now has been in the making for years and decades, of which politicians on Capitol Hill have been a part," Hurt, Washington Times opinion editor, told co-host Brian Kilmeade.
He added: "The last thing they want to do is take stock of themselves and try to figure out, 'OK, what have I done to make this worse or to create this situation?"'
Within seconds of reconvening Wednesday night, Democrats on Capitol Hill started "accusing Republicans of treason and sedition," Hurt said.
"They get caught up in their own mob mentality, they're all trying to outdo one another on Twitter to see who can make the most outrageous charge or make the most outrageous demand of the other side," Hurt said.
While this might be a good time for soul-searching for both parties, Hurt concluded, "There is no indication from Democrats on Capitol Hill that any one of them has any intention of doing that and certainly not from Joe Biden or Kamala Harris."

Step 2: Grouping salient spans into sentences.
"blame everything on President Trump" + "accusing the Republicans of treason and sedition" → *"blame everything on President Trump and accuse the Republicans."*
"Democrats are adding to the division" + "they get caught up in their own mob mentality, they're all trying to outdo one another" → *"Democrats get caught up in trying to outdo one another and are adding to the division."*
"last thing they want to do is take stock of themselves" + "this might be a good time for soul-searching" + "no indication from Democrats that any one of them has any intention of doing that" → *"They don't seem to have any intention of doing any soul-searching"*

Step 3: Combining sentences into a summary
Charles Hurt suggested that Democrats are rushing to blame everything on President Trump and accuse the Republicans. He said that Democrats get caught up in trying to outdo one another and are adding to the division. Finally, they don't seem to have any intention of doing any soul-searching.

Figure 1: Walk-through example showing the process of annotating summaries given a set of reported statements. Salient spans within the statements are shown in red and sentences copied over from step 2 into the summary in step 3 are shown in blue.

tions and create the PARC3 corpus. More recently, PolNeAR (Newell, Margolin, and Ruths 2018) was created to improve upon PARC3 by doubling the recall and improving inter-annotator agreement. However, all of these lines of work solely deal with identifying attribution and do not aggregate extracted statements from specific speakers to help with a downstream task. More direct uses of quotations in news include opinion mining (Balahur et al. 2009) and sentiment analysis (Balahur et al. 2013). In contrast, our proposed task involves attribution to identify reported statements in news articles, which are then aggregated and summarized to convey the reactions to events in news.

3 SumREN Benchmark

The SumREN benchmark aims to assist in the development and evaluation of models for the reported speech summarization task. In this section, we describe the task of reported speech summarization, the benchmark construction process, as well as present statistics of the constructed dataset.

Given a set of news articles about a specific event and the speaker name, the goal is to generate a succinct summary for the statements made by the speaker in the source content.

3.1 Benchmark Construction

The first step in our benchmark construction process involves collecting a news corpus discussing a large set of events. We split the news articles according to the discussed event and from each cluster of news articles, we then extract all reported statements along with the speakers of each of these statements. Finally, a summary is written for each group of statements by the same speaker.

News Corpus Acquisition: We first identified a list of 132 major news events between 2013-2021 that were mentioned in Wikipedia and other sources. We then collected a list of news articles discussing these events and retained articles that are present in Common Crawl (CC) News.[2] We ended up with a total of 633 news articles corresponding to 132 major events.

Reported Statements Annotation: To annotate the reported statements and the speakers, we used Amazon Mechanical Turk and collected three annotations per HIT. The annotation tasks were restricted to annotators in English-speaking countries and who passed the custom qualification test for the corresponding task – reported statement span selection or speaker identification.[3] Overall, 12% of the annotators that took the test were qualified. In addition, we blocked spammers that spent less than a specified number of seconds per task or that consistently provided low-quality annotations. For the reported statement span selection task, annotators were provided with a snippet from the news article and were asked to highlight the spans containing reported statements. Contiguous sentences with statements from the same speaker were considered to be parts of the same reported statement. After collecting the annotations, we grouped reported statements (and associated articles) by a specific speaker about each event.

Summary Generation: For summary generation, we relied on expert annotators since it is a more challenging task and hence less suitable for MTurk. Two reference summaries produced by two different annotators were created for each cluster of reported statements. An abridged version of the annotation guidelines is presented below and a walk-through example of the annotation process is shown in Figure 1.

- **Step 1:** Identify salient spans in the given statements.
- **Step 2:** Group similar salient spans – that discuss related aspects of the event – together and combine these into a

[2]For articles between 2013 and 2016, we relied on WayBack Machine since CC News is not available for these years.

[3]Please refer to appendix of the Arxiv version for detailed annotation guidelines.

single sentence; *using paraphrasing if needed.*

- **Step 3:** Combine these sentences into a summary.

3.2 Statistics

Our benchmark has 745 examples in total, with a train/dev/test split of 235/104/406 respectively. On average, the summaries have a length of 57 words and each summary comes from 5.3 reported statements. 57% of the examples have a single source news article, with 26% having 2 source articles and remaining 17% having 3-5 source articles. The average combined source length is 2,065 words. Overall, the news corpus contains 633 articles with a total of 10,762 reported statements from 3,725 unique speakers. Further, we observe that the summaries in our benchmark are relatively more abstractive compared to existing summarization datasets. Table 2 shows the percentage of novel n-grams, with SUMREN containing considerably higher novel trigrams and 4-grams. To account for this relatively higher abstractiveness and also variance in generation, each example in our benchmark has two reference summaries.

Datasets	**unigram**	**bigram**	**trigram**	**4-gram**
CNN-DM (S)	17.0	53.9	72.0	80.3
NY Times (S)	22.6	55.6	71.9	80.2
MultiNews (M)	17.8	57.1	75.7	82.3
WikiSum (M)	18.2	51.9	69.8	78.2
SumREN (M)	16.8	63.1	86.4	93.4

Table 2: Percentage of novel n-grams in the reference summaries of different summarization datasets. (S) and (M) denote single and multi-document summarization respectively. Numbers for SumREN are computed by averaging over the two reference summaries.

3.3 Silver Training Data Generation

Given the cost associated with annotating statements and writing summaries, we automatically generate large-scale silver-standard training data for our task. Specifically, we leverage GPT-3 (Brown et al. 2020) to automatically generate abstractive silver-standard summaries of the reported statements. This can be achieved by prompting (Liu et al. 2021a), which involves decomposing the task into an instruction (or a 'prompt') that is then provided to the model along with the input as the context. In our scenario, the input would be the reported statements and a speaker– automatically identified through the reported speech system that we build and describe in Section 4.2 and the prompt would be *"Summarize what <speaker> said:"*. Similar to the gold-standard dataset, statements corresponding to the same speaker are grouped together before prompting GPT-3 to generate the summary. Overall, we generate 10,457 examples for our silver training set.

4 Models

Here, we describe our proposed pipeline-based approach along with several strong baselines that we experiment with.

4.1 Query-Focused Summarization Baselines

Our proposed task requires generating a summary of the reported statements, given a set of news articles and the name of the speaker as input. To leverage existing models, our reported-speech summarization task can be approached as query-focused summarization – by generating a summary of the given text conditioned upon a query. Specifically, given the name of the speaker, the corresponding query can be formulated as: *"Summarize what <speaker> said."*. Following this, we explore multiple query-focused summarization approaches, which we describe below.

- **GR-SUM** (Wan 2008) uses an unsupervised graph-based extractive method where each source sentence is treated as a node.[4] It uses a random-walk algorithm to rank the input sentences based on the adjacency weights and the topic relevance vectors for each node.
- **RelReg** (Vig et al. 2022) uses a two-step process. First a regression model is used to extract a contiguous span within the input that is relevant to the input query. The extracted context is then passed along with the query to a BART model to generate a summary. Both the regression and BART models are trained on QMSum (Zhong et al. 2021a), a query-focused meeting summarization dataset.
- **SegEnc** (Vig et al. 2022) is an end-to-end generative model that first splits the source documents into overlapping text segments. Each of these segments is then concatenated with the input query and independently encoded by a Transformer encoder. The encoded segments are then concatenated into a sequence of vectors and fed into a Transformer decoder to generate the summary. The model is pre-trained on WikiSum dataset (Liu et al. 2018) and finetuned on QMSum dataset (Zhong et al. 2021b).
- **GPT-3**: In addition to these baselines, we also explore directly providing the source news articles as input to GPT-3 and using the query as the prompt.

4.2 Pipeline-Based Summarization Framework

We utilize a pipeline-based approach for summarizing reported speech. The proposed pipeline involves three main steps; (1) extracting reported statements and their speakers from the given set of news articles, (2) grouping statements together that come from the same speaker, and (3) generating a summary for each group of reported statements.

Reported Speech Extraction: Given a collection of news articles and a speaker, we aim to identify all reported statements along with the corresponding speakers. To this end, we build a span-tagging system that leverages a Transformer-based encoder to identify the spans of statements and the corresponding speaker. The model is trained using the PolNeAR corpus (Newell, Margolin, and Ruths 2018) which provides annotated triples of *source* (i.e. speaker), *cue* (i.e. words that indicate the presence of attribution), and *content* (i.e. the statements made by the speaker) for statements made in the news.

[4] We use the source code from Chan, Wang, and King (2021).

Setting	Model	Approach	Dev				Test			
			R-1	R-2	R-L	BertScore	R-1	R-2	R-L	BertScore
Baselines (Zero-shot)	GR-SUM	QFS	38.73	15.32	24.70	16.45	35.99	12.05	22.18	14.96
	RelReg		35.40	11.88	22.97	21.64	31.49	8.38	20.02	17.24
	SegEnc		38.53	14.99	24.98	26.26	36.62	11.77	22.99	23.26
	GPT-3		42.34	16.71	29.12	34.08	39.45	13.78	26.72	31.16
Zero-shot	BART	Pipeline	40.85	16.99	27.63	30.38	37.28	13.16	24.45	29.36
	GPT-3	Pipeline	44.49	18.51	31.21	**40.12**	42.29	16.02	29.33	**37.68**
	GPT-3	Pipeline (Oracle)	47.27	20.74	33.98	42.65	45.45	17.89	31.27	40.29
+ Silver Training	SegEnc	QFS	47.09	20.05	31.99	38.64	44.35	17.47	**29.69**	36.26
	BART	Pipeline	46.14	18.92	31.37	34.17	43.00	15.95	28.66	34.55
+ Gold Finetuning	SegEnc	QFS	**48.30**	**22.45**	**32.98**	39.95	**45.06**	**18.45**	29.43	36.71
	BART	Pipeline	46.59	20.38	32.31	37.78	44.38	17.53	29.62	35.72
	BART	Pipeline (Oracle)	51.11	24.23	35.92	42.28	47.82	20.23	32.20	39.61

Table 3: ROUGE and BertScore performance of various models on the SumREN benchmark. We explore both query-focused (QFS) and pipeline-based approaches under zero-shot, silver-training and gold-fine-tuning settings. *Pipeline (Oracle)* corresponds to using the gold reported statements as input to the summarization model and is reported for the best setup for each of the zero-shot and fine-tuned models.

Given an input paragraph of length T, we use a BERT encoder to learn the representation $H \in R^{TXD}$ – of hidden dimension D – for the input sequence. We then add a binary classification head to identify whether or not the input paragraph contains a reported statement and a BIO sequence labeling head to identify the spans of the statement and the speaker. The binary classification y^{cls} and the token label $Y_i^{span} \in R^K$ probabilities are calculated as follows:

$$y^{cls} = \sigma(w^{cls} \cdot H_{CLS} + b^{cls}) \quad (1)$$

$$Y_i^{span} = \text{softmax}(W^{sp} H_i + b^{sp}) \quad (2)$$

where $w^{cls} \in R^D$ and $W^{sp} \in R^{K \times D}$ are the weights, b^{cls} and b^{sp} are the bias terms, K is the total number of BIO tags, H_{CLS} and H_i denote the representation of the CLS token and the i-th token respectively.

Finally, the model is trained with a multi-task learning objective by using a joint loss that performs a weighted sum of the classification – binary cross entropy (BCE) – and the sequence labeling head – Cross Entropy (CE) – losses.

$$L = \alpha \cdot BCE(y^{cls}, \hat{y}^{cls}) + \beta \cdot CE(Y^{sp}, \hat{Y}^{sp}) \quad (3)$$

where y^{cls} and $\hat{y}^{cls}$ correspond to the predicted and ground-truth classification label respectively, Y^{sp} and $\hat{Y}^{sp}$ denote the predcited and ground-truth token labels respectively, α and β are tunable hyper-parameters.[5]

Speaker Co-reference Resolution: In order to group the statements by the speaker, we need to perform co-reference resolution since speakers can be referred to by different nominal (e.g., Biden, Joe Biden, Joe R. Biden) and pronominal (e.g., He) mentions. To achieve this, we utilize an existing information extraction system (Li et al. 2020), and updated it with a co-reference resolution from Lai, Bui, and Kim (2022). As we show later, using co-reference resolution considerably increases the coverage of reported statements by a given speaker.

Summary Generation: Given a set of reported statements for a speaker, we aim to generate a concise summary of the statements. The summary generation process for the extracted reported statements of a given speaker is akin to single-document summarization. The reported statements are concatenated before getting passed as input to a BART (Lewis et al. 2020) model. The summarization model, trained on CNN-DailyMail (Hermann et al. 2015), is first used in a zero-shot setting. This model then undergoes silver-training and gold-finetuning, the details of which are provided in Section 5.1.

5 Experiments

5.1 Training Setup and Metrics

We explore two methods for fine-tuning our base summary generation models: Silver Training and Gold Fine-tuning. During silver training, the models are fine-tuned on the silver-standard training data. For gold fine-tuning, we add a second fine-tuning step using the gold data.

For evaluation, we use ROUGE (Lin 2004) and choose the best models based on ROUGE-L performance on the development set.[6] We also report BertScore (Zhang* et al. 2020) which leverages pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity. As opposed to ROUGE which measures the lexical similarity between the source and generated summaries, BertScore is capable of capturing the semantic similarity.

5.2 Results

Table 3 compares the performance of our proposed pipeline-based approach against the QFS baselines with and without fine-tuning using our silver and gold data. For the baselines, GPT-3 performs best, justifying the choice of using it

[5]In our experiments, we set α to 1 and β to 0.4.

[6]We use the SCORE_MULTI function from the ROUGE_SCORE python package: https://pypi.org/project/rouge-score/

for generating silver-standard training data. We find that using silver training data for fine-tuning improves the performance of both the query-focused SegEnc and pipeline-based BART considerably, even outperforming GPT-3 in terms of ROUGE. Finally, we see that the models further benefit by fine-tuning on the gold human-annotated training data.

We also find that using the pipeline approach, where we first extract the reported statements before passing them to GPT-3, achieves considerably better scores than passing the raw articles to GPT-3. However, GPT-3 has relatively lower ROUGE scores than smaller models (SegEnc and BART) that have been fine-tuned using gold data. We hypothesize that this could be attributed to the fact that GPT-3 generates more abstractive summaries (as will be shown in Table 8) thereby leading to higher scores for metrics that are capable of capturing semantic similarity.

In zero-shot settings, the pipeline-based model considerably outperforms query-focused baselines, showing the benefit from explicitly extracting reported statements. However, in both silver training and gold fine-tuning settings, the SegEnc model consistently outperforms the pipeline-based models, suggesting that it may be possible to implicitly identify reported statements within an end-to-end approach. Nevertheless, when using the oracle reported statements, BART surpasses SegEnc – implying that employing better reported speech and co-reference resolution systems will considerably improve the pipeline-based approach.

Reported Speech Extraction Performance
Next, we analyze the performance of the proposed reported speech extraction component to identify areas of improvement. We compare our span tagging approach with a Semantic Role Labeling (SRL) baseline to identify reported statements and the corresponding speakers and evaluate using character-level offset F1-score of the extracted span. SRL outputs the verb predicate-argument structure of a sentence such as who did what to whom. Given a paragraph as an input, we filter out verb predicates matching a pre-defined set of cues that signal attribution (e.g., *say*, *believe*, *deny*) and identify these sentences as containing reported statements.[7] The sentences encompassing ARG-1 of the predicate are considered as the reported statement and the span corresponding to ARG-0 (agent) is used as the speaker.

Model	Dev			Test		
	P	**R**	**F1**	**P**	**R**	**F1**
SRL	84.3	42.7	56.7	83.3	40.8	54.8
+ co-reference	82.2	68.1	74.5	83.1	68.3	75.0
Span Tagging	80.3	48.6	60.5	80.2	45.0	57.6
+ co-reference	78.7	69.9	74.1	78.2	73.0	75.5

Table 4: Performance (in %) of different approaches for identifying reported statements corresponding to a given speaker for the summaries in SumREN. "+ *co-reference*" corresponds to adding co-reference resolution for the speaker mention extracted by the system.

[7]Full list of used cues is in the appendix of the Arxiv version.

Model	Dev		Test	
	Exact-Match	**F1**	**Exact-Match**	**F1**
SRL	20.8	48.1	16.1	44.4
+ co-reference	62.9	73.7	69.1	77.1
Span Tagging	22.3	51.2	18.8	49.3
+ co-reference	63.3	74.8	69.8	78.4

Table 5: Performance (in %) of the proposed span tagging component – against the baseline – on identifying the speakers corresponding to the given reported statements with and without co-reference resolution.

As shown in Table 4, our proposed span tagging model outperforms SRL, especially in terms of recall which ensures better coverage of information for the summarization step. We also find that incorporating co-reference resolution for speaker identification considerably improves recall with almost the same or slightly lower precision. Table 5 measures the performance of the proposed span-tagging approach for speaker extraction against the SRL baseline. We report both string exact-match and F1-score, both of which are commonly used in extractive question answering (Rajpurkar et al. 2016). We find that the performance of different approaches for identifying the speaker of a given reported statement improves significantly when using co-reference resolution. This is crucial for correctly grouping statements from the same speaker together.

	Dev		Test	
	R-1/2/L	**BertS**	**R-1/2/L**	**BertS**
Full FT	51.1/24.2/35.9	42.3	47.8/20.2/32.3	39.6
Gold FT	50.0/23.9/35.1	40.9	47.2/19.5/31.4	38.4
PE FT	50.7/24.6/36.1	42.0	47.8/20.2/31.8	39.4

Table 6: Comparison of performance of parameter-efficient fine-tuning for BART when used for summarization with oracle reported statements. *Full FT* corresponds to silver training + gold FT.

Parameter-Efficient versus Direct Fine-tuning
In addition to full fine-tuning methods, we also explore leveraging parameter-efficient fine-tuning approaches to directly fine-tune on the small-scale gold training data. We use LORA (Hu et al. 2021), an efficient fine-tuning technique that injects trainable low-rank decomposition matrices into the layers of a pre-trained model. Table 6 compares the performance of three different fine-tuning strategies, namely *Full FT* (silver training + gold fine-tuning), *Gold FT* (direct gold fine-tuning) and PE FT (parameter-efficient gold fine-tuning). We find that the benefit of incorporating the silver-standard training data can be seen from the fact that Full FT considerably outperforms Gold FT. We also observe that *PE FT* with LORA, which fine-tunes only 0.3% of model parameters, can achieve a comparable performance to Full FT while also consistently outperforming *Gold FT*. This shows that parameter-efficient fine-tuning is

Event: 2017 Solar Eclipse Speaker: Jamie Yuccas News Article: https://tinyurl.com/ycxbpbsu
Reported Statements
"They're expecting about a million people to enter the state, a million out-of-towners are supposed to come to the state of Oregon," said CBS News correspondent Jamie Yuccas. "Where we're located in Madras, they're expecting between 100,000 and 200,000 people."
She said the local residents have been "really, really nice and accommodating."
"What the mayor said to me was kind of funny," Yuccas said. "He said 'you know, I think it's going to be one of those situations that you might not get your newspaper, you might not have your daily Starbucks and if that happens, I guess it's a first-world problem, and you're going to have to figure out your own survival skills."'
She laughed, saying "there are going to be some minor inconveniences, but I actually think they had a pretty good plan together."
SegEnc (QFS): Jamie Yuccas has said that the local residents of Madras, Oregon have been very accommodating towards people who are coming to see the solar eclipse. She laughed when she heard that the mayor of the town thought it would be a first-world problem if the eclipse did not occur. She said that they are expecting about a million people to enter the state, and that a million out-of-towners are supposed to come to the state.
GPT-3 (QFS): Jamie Yuccas is a CBS News correspondent who is reporting from Madras, on the upcoming solar eclipse. She says that the city is expecting between 100,000 and 200,000 visitors for the event, and that the locals have been very accommodating. Yuccas also says that the eclipse is expected to be the most observed eclipse in history.
GPT-3 (Pipeline): Jamie Yuccas said that the city of Portland is expecting about a million out-of-towners to come to Oregon for the eclipse, and that the locals have been very accommodating. She said that the mayor told her that there may be some minor inconveniences, but that they have a good plan in place.
BART (Pipeline): According to the correspondent, the state of Oregon is expecting a million people from out-of-towns to come to the state, and the local residents have been very nice and accommodating. However, there will be some minor inconveniences, but the state had a good plan in place.
Gold: Jamie Yuccas said that a million tourists are supposed to come to the state of Oregon and between 100,000 and 200,000 people are expected in Madras, where she is located. She also mentioned minor inconveniences could occur derived from the event, but overall, they had a good plan set in place.

Table 7: Model outputs for an example in SumREN, along with the gold reported statements. Summaries from the QFS approaches contain factually inconsistent fragments, while those from pipeline-based approaches better match the gold summary.

effective for our pipeline-based reported speech summarization framework, with future work potentially benefiting from better PE approaches (Liu et al. 2021b).

Abstractiveness and Factuality of Generated Summaries We investigate the effect of using silver and gold data for fine-tuning, on both the abstractiveness and factuality of generated summaries. There is generally a trade-off between abstractiveness and factual consistency of the summary against the source input (Dreyer et al. 2021). Hence, the goal of any abstractive summarization system is to generate more abstractive summaries while maintaining a high level of factual consistency with the source.

For abstractiveness, we measure it through the percentage of novel n-grams (uni, bi and tri-grams), as well as MINT (*M*etric for lexical *IN*dependence of generated *T*ext) (Dreyer et al. 2021) which is computed based on the n-gram precision and longest common sub-sequence length of the generated summary. As shown in Table 8, we find that models in zero-shot settings are considerably more extractive, and that abstractiveness of generated summary significantly increases from both silver training and gold fine-tuning. Further, we notice that our pipeline-based approach is considerably more abstractive than the QFS approach, demonstrating that incorporating an explicit statement extraction component helps the summarization model focus on paraphrasing and synthesizing the selected statements into the summary.

For factuality, we use FactCC (Kryściński et al. 2020),

	Model	Setting	Uni	Bi	Tri	MINT
QFS	SegEnc	Zero-Shot	1.0	6.6	13.1	11.1
		+ Silver Train	2.8	22.8	39.3	31.2
		+ Gold FT	3.6	26.6	46.6	38.4
	GPT-3	Zero-Shot	3.8	26.2	44.2	38.9
Pipeline	BART	Zero-Shot	1.9	11.5	20.5	15.3
		+ Silver Train	3.3	24.8	41.6	32.9
		+ Gold FT	4.7	30.6	52.1	43.5
	GPT-3	Zero-Shot	5.7	35.2	56.6	49.6

Table 8: Abstractiveness and novelty scores – measured by % of novel ngrams – of the generated summaries using silver and gold data for fine-tuning the models. The novelty is computed with respect to the source news articles.

which Pagnoni, Balachandran, and Tsvetkov (2021) show to correlate most with human factuality labels. In addition, Entity Precision (Nan et al. 2021) is calculated based on the percentage of named entities in the generated summary that are present in the gold reported statements. In Table 9, we observe that while our proposed pipeline-based approach is considerably more abstractive than the QFS baselines, it still maintains high entity precision and a slightly higher FactCC score. As expected, we see that using gold (oracle) statements as input to the summarization step improves the factual consistency scores.

Approach	Model	FactCC	Entity P	MINT
QFS	GPT-3	45.4	61.7	38.9
	SegEnc	50.8	**75.4**	38.4
Pipeline	GPT-3	50.2	73.2	**49.6**
	BART	**52.1**	74.6	43.5
Pipeline (Oracle)	GPT-3	52.0	78.9	51.3
	BART	55.0	84.6	44.0

Table 9: Comparison of factuality (measured by FactCC and Entity Precision) of generated summaries relative to abstractiveness (measured by MINT). Models considered are after silver train + gold FT, except GPT-3 which is not fine-tuned.

Human Evaluation

We also performed a human study of the summaries generated using GPT-3 via both pipeline-based and QFS approaches. We chose GPT-3 summaries since they have consistently high scores across Rouge-L, BertScore and abstractiveness. Annotators were presented with the summaries along with the ground-truth reported statements, and were asked to evaluate on a scale of 1-3 for factual consistency, informativeness and coherence[8]. Evaluation for *factual consistency* involves looking for major or minor factual errors in the summary, *informativeness* is about how well the summary expresses the main point of the reported statements, and *coherence* is mainly checking whether the summary has a good flow and facts are presented in a logical order. The annotations were crowd-sourced via MTurk. Table 10 shows results from the human study. We find that summaries from the pipeline-based approach have considerably better factual consistency with the ground-truth reported statements, with slight improvements in informativeness. Concurring with recent observations (Goyal, Li, and Durrett 2022; Zhang et al. 2023) on the quality of summaries from large language models, we see that the summaries based on the two approaches, which both come from GPT-3, are very coherent.

Approach	Consistent	Informative	Coherent
GPT-3 (QFS)	2.76	2.92	2.99
GPT-3 (Pipeline)	2.92	2.95	2.99

Table 10: Results from human study on summaries from GPT-3 via pipeline-based and QFS approaches, when evaluated for factual consistency, informativeness and coherence.

5.3 Manual Error Analysis

Table 7 shows outputs from different models for an example in SUMREN. We see that summaries from the query-focused approaches contain factually inconsistent fragments: SegEnc output suggests that the mayor *"thought it would be a first-world problem if the eclipse did not occur"* whereas the mayor actually refers to *"people not getting their newspapers or their daily Starbucks"* as the first-world problems; GPT-3 (QFS) misattributes the statement *"the eclipse is expected to be the most observed eclipse in history"*. On the other hand, summaries from pipeline-based approaches match the gold summary better, with those from BART and GPT-3 (Pipeline) being fairly similar in quality.

We also analyzed some of the errors made by the reported speech extraction component of the proposed pipeline. As Table 3 shows, there is still a considerable room for improving our pipeline-based approach with better reported speech extraction systems. We found that the same entity can be referred to by different aliases that the co-reference system sometimes fails to capture (e.g., "Islamic State" and "ISIS" or "Anthony M. Kennedy" and "Justice Kennedy"). Utilizing entity-linking (Ayoola et al. 2022) will likely improve co-reference for entities with different aliases. In addition, we found that spelling variations; e.g., Nikos vs. Nicos, Muhammad vs. Mohammed or Sergey vs. Sergei, were also frequently missed by the system. We believe that incorporating character-level features into the co-reference resolution system will make it more robust to such variations.

Finally, to analyze the informativeness of the generated summaries, we calculated the percentage of input reported statements covered in the output summary. To obtain alignments between source-summary pairs, we leveraged SuperPAL (Ernst et al. 2021) which aligns OpenIE-extracted (Stanovsky et al. 2018) propositions in the generated summary with those in the source sentences. We found that human summaries cover considerably more percentage of the input reported statements (57.5%) compared to summaries from BART (51.2%) and GPT-3 (46.5%) in *Pipeline (Oracle)* settings. In order to explicitly improve coverage, future work can explore incorporating more control into the generation output by clustering the salient spans within the reported statements and separately generating summaries for each cluster, similar to Ernst et al. (2022).

6 Conclusion & Future Work

In this work, we introduce a new challenging task of summarizing reported speech in news and release SUMREN to promote more research in this direction. We propose a pipeline-based framework for summarizing reported statements and show that the proposed approach can generate summaries that are both more factual and abstractive than QFS. Future work involves improving reported speech extraction performance by leveraging entity-linking and by incorporating character-level features for speaker co-reference resolution. Another direction is to improve the coverage of salient spans in reported statements by adding more explicit control into the generation process.

Acknowledgments

We would like to thank members of the Alexa Web Information team, especially Markus Dreyer and Sandeep Atluri, for useful discussions and feedback. We would also like to thank the anonymous reviewers for their insightful comments.

References

Ayoola, T.; Tyagi, S.; Fisher, J.; Christodoulopoulos, C.; and Pierleoni, A. 2022. ReFinED: An Efficient Zero-shot-

[8]Detailed guidelines are in the appendix of the Arxiv version.

capable Approach to End-to-End Entity Linking. *arXiv preprint arXiv:2207.04108*.

Balahur, A.; Steinberger, R.; Kabadjov, M.; Zavarella, V.; Van Der Goot, E.; Halkia, M.; Pouliquen, B.; and Belyaeva, J. 2013. Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202*.

Balahur, A.; Steinberger, R.; Van Der Goot, E.; Pouliquen, B.; and Kabadjov, M. 2009. Opinion mining on newspaper quotations. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 3, 523–526. IEEE.

Banko, M.; Mittal, V. O.; and Witbrock, M. J. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 318–325.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33.

Chan, H. P.; Wang, L.; and King, I. 2021. Controllable Summarization with Constrained Markov Decision Process. *Trans. Assoc. Comput. Linguistics*, 9: 1213–1232.

Dang, H. T. 2005. Overview of DUC 2005. In *Proceedings of the document understanding conference*, volume 2005, 1–12.

Dreyer, M.; Liu, M.; Nan, F.; Atluri, S.; and Ravi, S. 2021. Analyzing the abstractiveness-factuality tradeoff with nonlinear abstractiveness constraints. *arXiv preprint arXiv:2108.02859*.

Elson, D. K.; and McKeown, K. R. 2010. Automatic attribution of quoted speech in literary narrative. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Ernst, O.; Caciularu, A.; Shapira, O.; Pasunuru, R.; Bansal, M.; Goldberger, J.; and Dagan, I. 2022. 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*, 1765–1779.

Ernst, O.; Shapira, O.; Pasunuru, R.; Lepioshkin, M.; Goldberger, J.; Bansal, M.; and Dagan, I. 2021. Summary-Source Proposition-level Alignment: Task, Datasets and Supervised Baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, 310–322.

Fabbri, A. R.; Li, I.; She, T.; Li, S.; and Radev, D. 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*, 1074–1084.

Goyal, T.; Li, J. J.; and Durrett, G. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Grusky, M.; Naaman, M.; and Artzi, Y. 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)*, 708–719.

Gu, X.; Mao, Y.; Han, J.; Liu, J.; Wu, Y.; Yu, C.; Finnie, D.; Yu, H.; Zhai, J.; and Zukoski, N. 2020. Generating representative headlines for news stories. In *Proceedings of The Web Conference 2020*, 1773–1784.

Hayashi, Y.; and Yanagimoto, H. 2018. Headline generation with recurrent neural network. In *New Trends in E-service and Smart Computing*, 81–96. Springer.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Jung, T.; Kang, D.; Mentch, L.; and Hovy, E. 2019. Earlier Isn't Always Better: Sub-aspect Analysis on Corpus and System Biases in Summarization. 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)*, 3324–3335.

Kryściński, W.; McCann, B.; Xiong, C.; and Socher, R. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9332–9346.

Kulkarni, S.; Chammas, S.; Zhu, W.; Sha, F.; and Ie, E. 2020. AQuaMuSe: Automatically Generating Datasets for Query-Based Multi-Document Summarization. *CoRR*, abs/2010.12694.

Lai, T. M.; Bui, T.; and Kim, D. S. 2022. End-to-end neural coreference resolution revisited: A simple yet effective baseline. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8147–8151. IEEE.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 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*, 7871–7880.

Li, M.; Ma, T.; Yu, M.; Wu, L.; Gao, T.; Ji, H.; and McKeown, K. 2021. Timeline Summarization based on Event Graph Compression via Time-Aware Optimal Transport. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6443–6456.

Li, M.; Zareian, A.; Lin, Y.; Pan, X.; Whitehead, S.; Chen, B.; Wu, B.; Ji, H.; Chang, S.-F.; Voss, C.; et al. 2020. Gaia: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Liu, P. J.; Saleh, M.; Pot, E.; Goodrich, B.; Sepassi, R.; Kaiser, L.; and Shazeer, N. 2018. Generating Wikipedia by Summarizing Long Sequences. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Liu, X.; Ji, K.; Fu, Y.; Du, Z.; Yang, Z.; and Tang, J. 2021b. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Nan, F.; Nallapati, R.; Wang, Z.; dos Santos, C. N.; Zhu, H.; Zhang, D.; McKeown, K.; and Xiang, B. 2021. Entity-level Factual Consistency of Abstractive Text Summarization. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, 2727–2733. Association for Computational Linguistics.

Narayan, S.; Cohen, S. B.; and Lapata, M. 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*, 1797–1807.

Newell, E.; Margolin, D.; and Ruths, D. 2018. An attribution relations corpus for political news. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Pagnoni, A.; Balachandran, V.; and Tsvetkov, Y. 2021. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In *NAACL-HLT*.

Pareti, S. 2012. A Database of Attribution Relations. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3213–3217.

Pareti, S.; O'keefe, T.; Konstas, I.; Curran, J. R.; and Koprinska, I. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 989–999.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Schick, T.; and Schütze, H. 2021. Generating Datasets with Pretrained Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6943–6951.

Stanovsky, G.; Michael, J.; Zettlemoyer, L.; and Dagan, I. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 885–895.

Steen, J.; and Markert, K. 2019. Abstractive timeline summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 21–31.

Vig, J.; Fabbri, A. R.; Kryscinski, W.; Wu, C.; and Liu, W. 2022. Exploring Neural Models for Query-Focused Summarization. In Carpuat, M.; de Marneffe, M.; and Ruíz, I. V. M., eds., *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*. Association for Computational Linguistics.

Wan, X. 2008. Using only cross-document relationships for both generic and topic-focused multi-document summarizations. *Information Retrieval*, 11(1): 25–49.

Xu, Y.; and Lapata, M. 2021. Generating Query Focused Summaries from Query-Free Resources. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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*, 6096–6109. Association for Computational Linguistics.

Xu, Y.; and Lapata, M. 2022. Document Summarization with Latent Queries. *Trans. Assoc. Comput. Linguistics*, 10: 623–638.

Zhang*, T.; Kishore*, V.; Wu*, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhang, T.; Ladhak, F.; Durmus, E.; Liang, P.; McKeown, K.; and Hashimoto, T. B. 2023. Benchmarking Large Language Models for News Summarization. *arXiv preprint arXiv:2301.13848*.

Zhang, Y.; and Liu, Y. 2021. DirectQuote: A Dataset for Direct Quotation Extraction and Attribution in News Articles. *arXiv preprint arXiv:2110.07827*.

Zhong, M.; Yin, D.; Yu, T.; Zaidi, A.; Mutuma, M.; Jha, R.; Awadallah, A. H.; Celikyilmaz, A.; Liu, Y.; Qiu, X.; and Radev, D. R. 2021a. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tür, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *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*, 5905–5921. Association for Computational Linguistics.

Zhong, M.; Yin, D.; Yu, T.; Zaidi, A.; Mutuma, M.; Jha, R.; Hassan, A.; Celikyilmaz, A.; Liu, Y.; Qiu, X.; et al. 2021b. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5905–5921.

Zhu, C.; Yang, Z.; Gmyr, R.; Zeng, M.; and Huang, X. 2021. Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1462–1471.