

# Meeting Summarization: A Survey of the State of the Art

**Lakshmi Prasanna Kumar**

IMRSV Data Labs., Ottawa, Canada  
lakshmi@imrsv.ai

**Arman Kabiri**

IMRSV Data Labs., Ottawa, Canada  
arman@imrsv.ai

## Abstract

**Information** overloading requires the need for summarizers to extract salient information from the text. Currently, there is an overload of dialogue data due to the rise of virtual communication platforms. The rise of Covid-19 has led people to rely on online communication platforms like Zoom, Slack, Microsoft Teams, Discord, etc. to conduct their company meetings. Instead of going through the entire meeting transcripts, people can use meeting summarizers to select useful data. Nevertheless, there is a lack of comprehensive surveys in the field of meeting summarizers. In this survey, we aim to cover recent meeting summarization techniques. Our survey offers a general overview of text summarization along with datasets and evaluation metrics for meeting summarization. We also provide the performance of each summarizer on a leaderboard. We conclude our survey with different challenges in this domain and potential research opportunities for future researchers.

## 1 Introduction

In the current pandemic era, numerous meetings are happening virtually thereby producing substantial amount of meeting data on a daily basis. Meeting participants need to do the laborious task of going through the meeting transcripts to extract the salient information which in turn has attracted researchers attention to the need of effective meeting summarizers. According to [Feng et al.](#), meeting summarization, the task of summarizing online multi-participant meetings, is a difficult task to perform due to its multi-speaker structure and dialogue style which lead to low information density. The fact that multiple participants in a meeting have different language styles and roles result in meeting heterogeneity. There are other challenges like lack of annotated datasets, length of the meeting transcripts, etc. This paper performs a thorough study of over forty papers covering different sum-

marization techniques for meetings and discuss the challenges in this field.

The paper starts by giving a background on the standard text summarization techniques followed by dialogue summarization and the metrics used for evaluation in section 2. In section 3, we discuss the meeting summarization techniques followed by the datasets commonly used for training and evaluation of the meeting summarization models in section 4. The discussions and challenges faced in meeting summarization is explained in sections 5 and 6, respectively. We conclude the survey in section 7 by giving future research directions for the potential researchers who are interested to work in this domain.

## 2 Background

In this section we briefly describe the automatic text summarization methods followed by an overview of dialogue summarization tasks. A brief discussion of evaluation metrics is also discussed here.

### 2.1 Automatic Text Summarization

Automatic text summarization is an inevitable task to process and understand vast amount of textual data which is available in the form of web contents, scientific papers, legal documents, news articles, medical documents, etc. According to ([Hovy and Marcu, 2005](#)), a summary is defined as a text that is produced out of one or more texts that contain a significant portion of the information of the original text(s), and that is no longer than half of the original text(s). It is a difficult task for computers to understand the entire context of a document and find the significant data in it as compared to humans ([Al-lahyari et al., 2017](#)). Scientists have started the research in text summarization as early as 1950, and they are still seeking new techniques to achieve a summary as close to human summary ([Gambhir and Gupta, 2017](#)).

Automatic text summarization(ATS) can be classified in different ways (El-Kassas et al., 2021). Some of the important categories are discussed below. ATS can be single document summarization or multi-document summarization depending on the type of input source. A multi-document summarization system extracts salient information from multiple documents according to some topic and presents them in a coherent concise manner (Lin and Hovy, 2002). Besides, based on the summarization approach, ATS can be divided into extractive and abstractive summarization. Extractive approach extracts the most important sentences from the input text and generates the summary by concatenating them, whereas abstractive approach generates a summary by paraphrasing the contents of the input text (Nallapati et al., 2017). Most of the time, extractive summaries are simpler and more accurate than the latter. Maximal Marginal Relevance (MMR) algorithm suggested by Carbonell and Goldstein was one of the initial methods for sentence selection in extractive text summarization followed by integer linear programming (ILP) (McDonald, 2007), submodular-based approaches (Lin and Bilmes, 2012), etc. For sentence scoring, there are graph-based approaches like LexRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004). Some neural network-based approaches are also proposed for sentence scoring (Cao et al., 2015; Ren et al., 2017). The following years saw a rise in popularity of using neural network-based models for both extractive and abstractive paradigms. Nallapati et al. and Nallapati et al. proposed abstractive and extractive techniques using recurrent neural networks(RNN) for generating multi-sentence summaries. In order to overcome the repetition of words and to increase the accuracy, See et al. suggested a hybrid pointer-generator network with coverage mechanism. With the introduction of transformer-based sequence-to-sequence models (Vaswani et al., 2017), there is a boom in the use of pretrained language models for text summarization tasks. The transformer architecture enables transfer learning for NLP by training language models on an unlabeled raw corpus and then fine-tuning for specific downstream tasks. Pre-trained language models like Bert (Liu and Lapata, 2019), GPT (Radford et al., 2019), Bart (Lewis et al., 2019), and Pegasus (Zhang et al., 2020) have been very successful in this domain.

## 2.2 Dialogue Summarization

Most of the published work on text summarization has targeted single-speaker formal documents like news, scientific articles, etc (Cachola et al., 2020; Yasunaga et al., 2019; Narayan et al., 2018; Zhang et al., 2020). Dialogue summarization is a domain which is gaining popularity due to the availability of high volume conversational data obtained when people use digital platforms and smartphones to exchange information. Dialogue domain can contain emails, meetings, online chats, customer service interactions, medical conversations between doctors and patients, podcasts, etc. (Feng et al., 2021). The sub-domain under dialogue summarization domain, Meeting-based summarization, will be explained further in section 3.

**Chat Summarization:** There is a growing interest in chat summarization because of the extensive usage of chat applications like Slack<sup>1</sup>, Discord<sup>2</sup>, etc. One of the earliest work is a personalized summarization based on user profiles from group chats, like Slack channels by a chat assistant service, Collabot (Tepper et al., 2018). Another early work, SAMSum (Gliwa et al., 2019), is a benchmark dataset for abstractive dialogue summarization with 16k dialogues which has facilitated much research in this domain. Chen and Yang extracted different views from the conversations using SAMSum and generated multi-view dialogue summary utilizing a conversation encoder and a multi-view decoder. The authors proposed a multi-view decoder that is a transformer-based decoder to integrate encoded representations from different views and further generate summaries. Most of the recent research focuses on improving the quality of abstractive summaries. Lee et al. proposes self-supervised strategies that focus on correcting the speaker names by a post correction model thereby generating a revised summary. To reduce hallucinations and to improve factual consistency Liu and Chen; Narayan et al.; Wu et al. proposed conditioning the summary with personal named entities, entity chain, and summary sketch, respectively. Graph based dialogue interaction models are also studied in some works. Summarization based on topic words using a guided graph model is one such work (Zhao et al., 2020). Another work under this category is by Feng et al. who integrated common sense knowledge in a heterogeneous graph along

<sup>1</sup><https://slack.com/>

<sup>2</sup><https://discord.com/>

with utterance and speaker nodes. To overcome the absence of sufficient data, [Gunasekara et al.](#), using pretrained language models, worked on generating conversations for a given summary by three approaches — supervised, reinforced, and turn-by-turn utterance generation in a controlled manner. This augmented conversation dataset can be used to improve the performance of summarization models. MediaSUM ([Zhu et al., 2021](#)) is another dataset for abstractive dialogue summarization which is worth mentioning. This dataset has more than 450K interview transcripts procured from national Public radio (NPR) and CNN annotated with description of topics and interview overviews as summaries.

**Email Summarization:** Initial works on email thread summarization focused on multiple approaches: extracting topic phrases ([Muresan et al., 2001](#)), forming message clusters based on groups of topics ([Newman and Blitzer, 2003](#)), extracting important sentences based on specific email-related features ([Rambow et al., 2004](#)), etc. Recently, some task-specific approaches have been proposed, such as automatic email subject-line generation ([Zhang and Tetreault, 2019](#)) and email action item extraction/To-Do item generation ([Mukherjee et al., 2020](#)).

**Medical conversation Summarization:** Medical conversation summarization is an important application of dialogue summarization. According to [Feng et al.](#), summarization methods in medical domain prefer to use a combination of extractive and abstractive methods to generate an accurate summary.

**Customer conversation Summarization:** In addition to these domains, dialogue summarization is popular in summarizing customer interactions. Some of the customer service summarization models use topic modeling to generate accurate summaries ([Liu et al., 2019](#); [Zou et al., 2020](#)). One of the datasets under this category is TODSum ([Zhao et al., 2021](#)) which has a state knowledge of dialogues like user intents, preferences, etc. along with the dialogues.

## 2.3 Metrics for summary evaluation

The most common standard evaluation metric used for text summarization is Rouge (Recall-Oriented Understudy for Gisting Evaluation) which is an automatic approach for measuring the quality of generated summaries. This metric compares a generated summary with a set of human-generated

summaries by counting the number of overlapping ngrams between the generated summary and reference summaries ([Lin, 2004](#)). F1 scores for R1(Rouge-1, unigram overlap), R2(Rouge-2, bigram overlap), and Rouge-L (longest common sequence) are some of the common reported metrics in most papers. BERTScore ([Zhang et al., 2019](#)) is another automatic metric which calculates semantic similarity with the help of contextual embeddings from pretrained BERT.

## 3 Meeting Summarizers

In this section we describe various summarization approaches for generating a concise summary from the meetings transcripts. This section is divided into extractive and abstractive based summarization approaches.

### 3.1 Extractive Meeting Summarization

One of the earliest work in extractive meeting summarization is by [Murray et al.](#) that focused on a study of text summarization techniques like MMR and Latent Semantic Analysis (LSA) on ICSI dataset ([Janin et al., 2003](#)). They have compared these approaches with feature-based summarization approaches that include prosodic and lexical features. The results showed that MMR and LSA outperformed the feature-based approaches. [Riedhammer et al.](#) suggested using key phrases as queries to MMR can significantly improve the model performance. In addition to the features like lexical, structural, discourse, and prosodic features, [Xie et al.](#) found topic-related features as useful. In ([Xie and Liu, 2008](#)), the authors proposed different ways to calculate similarity measures like centroid score and corpus-based similarity metrics ([Mihalcea et al., 2006](#)). Numerous works were proposed afterwards that performed better than the greedy search algorithm, MMR. We can divide extractive summarizers into MMR-based, graph-based, optimization-based, and supervised categories ([Bokaei et al., 2016](#)) as shown in Figure 1.

The second category is graph-based algorithms where nodes represent utterances, and the edges represent the similarity between the utterances. A random walk on the graph is then used for ranking the utterances to be included in the generated summary. An unsupervised graph-based approach, ClusterRank ([Garg et al., 2009](#)) extended TextRank algorithm ([Mihalcea and Tarau, 2004](#)) and per-

formed better than MMR, taking care of redundant and off-topic data. In this approach, nodes are clusters of adjacent utterances grouped based on similarity, and the scoring of each utterance is based on the cluster it is associated with. The research work (Bokaei et al., 2016), similar to ClusterRank, segments the meeting transcript using an unsupervised genetic algorithm and scores the utterances using a weighted scheme based on their association with each segment. The categorization of the segments is based on the participation of speakers which can be a monologue or discussion. A two layer graph consisting of utterance layer and speaker layer suggested by Chen and Metze uses three relations to score nodes namely, utterance-utterance, utterance-speaker, and speaker-speaker thereby giving importance to both speakers and utterances.

To overcome the non-optimality issue of MMR, optimization based approaches were introduced. Optimum utterances were selected using ILP algorithm (Lin et al., 2009; Gillick et al., 2009; Riedhammer et al., 2010) under the constraint of summary length. The fourth category is based on supervised algorithms which performs a binary classification to decide whether to include an utterance in the summary or not. Xie and Liu and Liu and Liu gave suggestions for improving supervised learning by incorporating different sampling techniques and speaker-related features, respectively. Prediction of summary worthy extracted dialogue acts (EDA) from the meeting corpora is another area which is explored by Lai et al.. They used a regression model using *turn taking* features like *participation equality*, *turn taking freedom*, *barge-in rate*, etc. for capturing various aspects of participation.

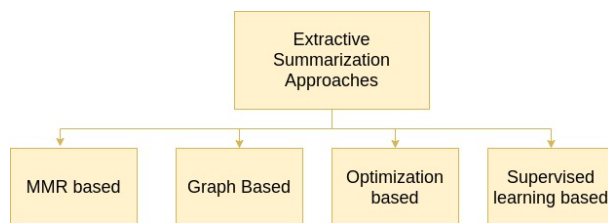


Figure 1: Extractive Summarization Approaches

For predicting summary-worthy extracts, multimodal features like gestures, facial expressions, language, prosody, etc. are also important. Nihei et al. discusses about fusing non-linguistic with linguistic features to perform a multimodal extractive summarization. Murray et al. suggested that

users prefer abstractive summaries over extractive summaries. They point that extractive summary, being a copy of the input text, copies the noise or grammatical mistakes present in the input document. Another weakness of extractive summaries is that they may lose their coherence. Thus, the user may need to review the original document to understand the context. Abstractive summaries are more preferable and are closer to human-generated summaries.

## 3.2 Abstractive Meeting Summarization

We can categorize abstractive summarization approaches as shown in Fig. 2.

### 3.2.1 Graph Based Summarization

Mehdad et al. proposed detecting utterance's communities and applying entailment graph to select significant utterances. The final abstractive summary is obtained by applying word graph fusion algorithm based on ranking on these salient sentences. Banerjee et al. used ILP based fusion approach to aggregate all the utterances that are important in each topic segment and generate a sentence summary for each topic segment. Utterances were represented as a dependency graph, and to increase the chance of merging, the author used anaphora resolution<sup>3</sup>. Anaphora resolution is used to replace pronouns with the original nouns from the previous utterance they referred to. This work was later extended in (Banerjee et al., 2015b). The authors created an end-to-end framework starting from topic segmentation, identifying important utterances in each topic segment and then applying utterance fusion to obtain the summary. For topic segmentation, LCSeg (Galley et al., 2003) and bayesian unsupervised topic segmentation (Eisenstein and Barzilay, 2008) methods were used, followed by an extractive summarization method to select important utterances. Contrary to their previous work, linguistic constraints were also introduced to the ILP problem to generate a readable summary. An unsupervised method using graph-based approach for abstractive summarization was proposed by Shang et al.. In this approach, after performing the pre-processing step, the utterances are grouped together to form communities. These communities are based on a topic or a subtopic discussed in the meeting. Next, from each community, a single abstractive sentence is generated

<sup>3</sup>Anaphora resolution is used to identify what a pronoun or noun phrase is referring to



using an extension of Multi-sentence compression Graph (Filippova, 2010). The final step, is to select most important sentences from this category using budgeted submodular maximization.

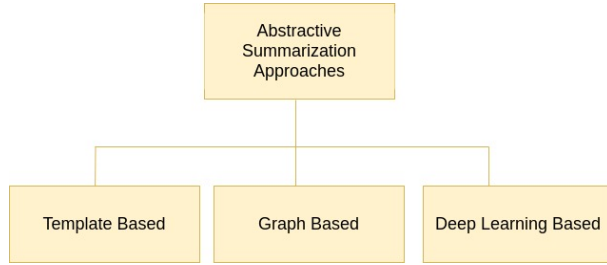


Figure 2: Abstractive Summarization Approaches

### 3.2.2 Template Based Summarization

The overall workflow of a template based summarization approach is shown in Figure 3. The basic idea is to generate a template from the manual summary and fill these templates with significant information extracted from the input text. From these filled templates, sentences are selected to generate the final summary based on a ranking scheme. Wang and Cardie proposed focused abstractive meeting summary generation that summarizes a particular aspect of the meeting using templates. The first step on their framework was to identify dialogue act clusters. Salient information was selected from this cluster in the form of relation instances using a discriminative classifier based on support vector machines (SVM). The templates were learnt from manual summary using an algorithm based on Multiple Sequence Alignments (MSA) (Durbin et al., 1998). *Overgenerate-and-Rank* strategy (Walker et al., 2001; Heilman and Smith, 2010) was used for template filling followed by statistical ranking of a candidate abstract. The final summary was generated after removing redundancies. The paper (Oya et al., 2014) proposed template generation based on hypernym labeling, clustering, and applying a word graph-based fusion algorithm. Instead of a classifier, they used topic segmentation followed by extraction of important phrases and speakers. This work also used a novel template selection algorithm that identifies the templates associated with communities similar to the topic segments. Similarly, Singla et al. also used template based abstractive summarization and explored the creation of communities using different heuristics.

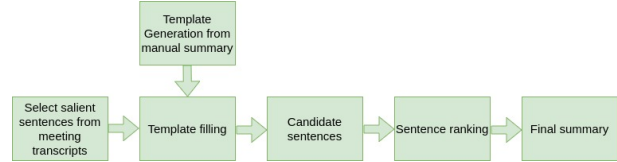


Figure 3: Workflow of template based summarization

### 3.2.3 Deep Learning Based Summarization

We can see a paradigm shift in the abstractive summarization approaches after the evolution of neural networks. A categorization of deep learning-based summarization approaches is shown in Fig. 4.

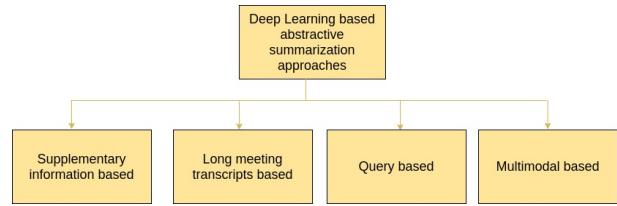


Figure 4: Deep learning based abstractive summarization approaches

**Query Based:** According to (Zhong et al., 2021b), meeting participants would like to get the summary of the meetings according to their area of interest. To serve this purpose the authors have proposed a summarization task based on a query. For this task they have also created a multi-domain summarization dataset with query-summary pairs. The summarization was done in a pipeline of two steps, locating the relevant span of meeting transcripts according to the query, and summarizing it with the help of state-of-art abstractive models like Pointer-Generator Network (See et al., 2017), BART (Lewis et al., 2020), HMNet (Zhu et al., 2020). The authors used two methods for locator, a Pointer Network (Vinyals et al., 2015) and a hierarchical ranking based model. Vig et al. conducted an exploratory study of various approaches on Query Focused Summarization. The authors also proposed two end-to-end neural network-based approaches which are composed of extractor models followed by abstractor models.

**Multi-modal based:** The lack of structure alongside grammatical mistakes of meeting data makes the natural language models to lose focus in summarizing. Li et al. proposed using multi-modal data in addition to the meeting transcripts for the task of summarization. They proposed to use Visual Focus Of Attention (VFOA) as a feature to select salient

data. The work also proposes a multi-modal hierarchical attention which focuses on topic segments, and then to the utterances and words in it. VFOA is obtained by observing the participants' video recordings for each frame and estimate the eye gaze direction vector and head pose angle. If a speaker receives more attention, we can consider their utterances as a summary-worthy utterance which should be paid more attention when generating a summary. Topic segmentation and summarization model are trained jointly in this work.

### **Summarization Models for Long meetings:**

There are a set of works which focus on summarization of long meeting transcripts. One of the methods enabling sequence-to-sequence models to summarize long transcripts is to use hierarchical attention. Zhao et al. recommended a hierarchical encoder-decoder network with adaptive segmentation of conversations for the summarization task. The semantic representation of meeting transcripts is learned by the encoder which is a bi-directional LSTM along with the conversation segmentation. A decoder trained with reinforcement learning framework is used to generate the abstractive summary. In (Zhu et al., 2020), the authors have discussed about using a hierarchical structure for handling long meeting transcripts and proposing a Hierarchical Meeting summarization Network (HMNet) which is based on transformer architecture (Vaswani et al., 2017). The attention is given in the token level and turn level of the meeting dialogues, and the role vector of speakers are added to the turn level encoding. To overcome the challenge of inadequate meeting summary data, this model is pretrained on various news summarization datasets. Here, the data from news domain is converted into the meeting format where different news articles mimic a multi-person meeting. Each sentence forms a turn and they are reshuffled to resemble a mixed order of speakers. Rohde et al. designed a Hierarchical Attention transformer (HAT) architecture by modifying the standard transformer architecture (Vaswani et al., 2017) and adding hierarchical attention. They added BOS tokens at the starting of each sentence and added a layer in the encoder block which attends to these BOS embeddings. This layer generates contextual representations for the BOS tokens which can be considered as sentence level representations. An additional attention module is added to the decoder to attend these BOS token embeddings. Zhang et al. per-

formed a comparative study of three different methods for long dialogue summarization. The three models are (1) Longformer — the extended transformer model (Beltagy et al., 2020) —, (2) *retrieve-then-summarize* pipeline models, and (3) the hierarchical HMNet model. Their study established that *retrieve-then-summarize* pipeline performs well compared to other models. Two main stages of the pipeline are retrieving relevant utterances and summarization. For retrieval, they experimented with TF-IDF (Jones, 1972), BM25 (Robertson and Zaragoza, 2009), and Locator (Zhong et al., 2021b), and for summarization, BART-large model (Lewis et al., 2020) was used. Some of the works focused on sliding window approach in which the long transcripts are broken down into small windows and a summary is generated from each window. Koay et al. proposed a sliding window approach along with a Bart based abstractive summarizer to obtain the partial summaries of each window. Using these local summaries, salient utterances were identified based on Rouge score. These selected utterances were then included to the main meeting summary. The results were promising when a sliding window approach was used. Liu and Chen also proposed a dynamic sliding window approach where the window stride size were predicted by their proposed model called Retrospective. The decoder predicts the context boundary along with the summary sentence. The final summary is obtained by concatenating the summary of each window segment. Zhang et al. suggested a new framework for handling long input data called Summ<sup>N</sup> which had a coarse-grained and fine-grained stage. The coarse-grained stage performed the compression of the input text by doing a segmentation followed by a coarse summary generation. Multiple coarse stages were repeated with different model parameters. Finally, the coarse summary is passed to the fine-grained stage to get the final summary. The underlying summarizer model used is BART (Lewis et al., 2020). In another study, a pre-trained model called DIALOGLM was suggested in (Zhong et al., 2021a) for understanding and summarizing long dialogues. The proposed model is a neural auto-encoder, and a window-based denoising method is used for pretraining the model. For this purpose, five types of noise are introduced based on the dialogue characteristics. The model is trained to restore the original content out of the given noisy content and the remaining dialogue. After pretrain-

ing, the model is applied on three downstream tasks including long dialogue summarization.

**Supplementary Information based:** Some of the works have used supplementary information as a guidance for creating summarizer models. RetrievalSum (An et al., 2021) retrieves high quality semantically-similar exemplars from a knowledge base which acts as a guidance of writing format and also provides background information. Given the query document, the retrieval module retrieves exemplars by a semantic matching process. Ganesh and Dingliwal and Feng et al. incorporated discourse relations into their summarizer pipeline. Ganesh and Dingliwal proposed a zero-shot learning based abstractive summarization method using discourse relations. These discourse relations are used to restructure the conversations into a document. This is followed by a document summarization process using two models: pointer-generator networks (See et al., 2017) and transformer-based model BART (Lewis et al., 2020). Discourse labels generated using Conditional Random fields (Okazaki, 2007) is used for dialogue restructuring into a document. Feng et al. presented a Dialogue Discourse-Aware Meeting Summarizer (DDAMS) that models the dialogue discourse relations among utterances. The meeting utterances with discourse relations are converted into a meeting graph which is then modeled using a graph encoder. Koay et al. annotated the dataset with domain terminologies (jargon terms) and studied their impact on summarizers performance. Goo and Chen leveraged dialogue acts for dialogue summarization using sentence-gated modeling. Dialogue act is based on the interactive patterns between speakers which can be very informative for summarization. In a different study, Beckage et al. explored incremental temporal summarization with a focus on previous summaries generated by humans. They also suggest semantic role labelling to extract information from past summaries.

## 4 Datasets

AMI (Carletta et al., 2005) and ICSI (Janin et al., 2003) are the two main corpora used for meeting summarization task. The meeting transcripts are obtained using Automatic Speech Recognition (ASR). Both of the datasets contain human annotated extractive and abstractive summaries. The AMI dataset meetings are about the product de-

sign of a new remote controller. The ICSI dataset however contains academic discussions between research students. AMI and ICSI datasets have 289 and 464 turns with 4757 and 10,189 words, respectively (Zhu et al., 2020). Each summary in AMI and ICSI datasets contains 322 and 534 words on average, respectively. QMSum (Zhong et al., 2021b) is a multi-domain dataset which contains data from three domains such as academic meetings, product meetings, and committee meetings. It has 1,808 query-summary pairs over 232 meetings — 137 AMI meetings, 59 ICSI meetings, and 36 Parliament committee meetings. ConvoSumm (Fabbri et al., 2021) is a suite of four datasets of online conversations based on news comments, discussion forums, community question answering forums, and email threads. Using these datasets, the authors also have benchmarked state-of-the-art summarization models. In this survey, we try to cover all publicly reported results for any of these datasets.

## 5 Discussions

The leaderboard of the meeting summarizers is shown in the table 1. The results are reported from published literature reviews including (Feng et al., 2021), (Fabbri et al., 2021), and (Zhong et al., 2021a), and the original papers. Missing values in the leaderboard indicate that the results are not available for the corresponding dataset. Besides, in all of these cases, the code is not publicly made available to reproduce the experiments on the other datasets for which results are not reported. For dataset AMI, the Rouge-1 score of RetrievalSum is the highest(56.26) when compared to other deep learning based abstractive methods. Models such as DIALOGLM and Longformer-BART-arg follows RetrievalSum in the Rouge-1 score with the values 54.49 and 54.47. The highest Rouge-1 score for RetrievalSum can be attributed to the two mechanisms — Group Alignment and Rouge Credit which helps to learn the structure and style of summary from exemplars in addition to the background knowledge. DIALOGLM cannot be compared with RetrievalSum as the former is a pre-trained model specifically designed for long dialogue understanding and summarization. For ICSI dataset, DIALOGLM outperforms all other models with a ROUGE-1 score of 49.56. For example, this model exceeds ROUGE-1 score of HM-Net by 2.97 points. Longformer-BART-arg is a longformer model initialised with BART param-

Model	AMI			ICSI		
	R-1	R-2	R-L	R-1	R-2	R-L
<b>Extractive Methods</b>						
TextRank (Mihalcea and Tarau)	35.19	6.13	15.7	30.72	4.69	12.97
SummaRunner (Nallapati et al.)	30.98	5.54	13.91	27.6	3.7	12.52
<b>Abstractive Methods (DL-based)</b>						
Topicseg (Li et al.)	51.53	12.23	25.47	-	-	-
Topicseg+vfoa (Li et al.)	53.29	13.51	26.9	-	-	-
HMNet(Zhu et al.)	53.02	18.57	-	46.28	10.6	-
HAT-CNNNDM(Rohde et al.)	52.27	20.15	50.57	43.98	10.83	41.36
Sliding Window(Koay et al.)	51.2	27.6	-	-	-	-
BART-SW-Dynamic(Liu and Chen)	52.83	21.77	26.01	-	-	-
SUMM^N(Zhang et al.)	53.44	20.3	51.39	48.87	12.17	46.38
DIALOGLM (l = 5, 120)(Zhong et al.)	54.49	20.03	51.92	49.25	12.31	46.8
DIALOGLM-sparse (l = 8192)(Zhong et al.)	53.72	19.61	51.83	49.56	12.53	47.08
Sentence-Gated (Goo and Chen)	49.29	19.31	24.82	39.37	9.57	17.17
Retrievalsum (An et al.)	56.26	34.9	52.51	-	-	-
DDAMS(Feng et al.)	53.15	22.32	25.67*	40.41	11.02	19.18
Domain Terminology(Koay et al.)	-	-	-	60.7	37.1	-
Longformer-BART (Fabbri et al.)	54.2	20.72	51.36	43.03	12.14	40.26
Longformer-BART-arg (Fabbri et al.)	54.47	20.83	51.74	44.17	11.69	41.33

Table 1: Leaderboard. We adopt reported results from literature. Results with \* indicate that ROUGE-L is calculated without sentence splitting.

ters and trained with argument-mining based input. This model allows up to 16K tokens in the text input. The results of this model are comparable to those of DIALOGLM for AMI but the latter performs better for ICSI dataset. For ICSI dataset, (Koay et al., 2020b) outperforms every other model with ROUGE-1 value of 60.7, despite its simplicity. The results show the impact of domain terminology on meeting summarization.

## 6 Challenges

An abstractive text summarizer often suffers from a factual inconsistency problem. These problems are also called as hallucinations. Meeting summarizers also can suffer from this issue. Besides, lengthy meeting transcripts often pose a challenge to the meeting summarizers. The number of tokens in a meeting transcript is typically more than what can be handled by a transformer architecture. Many recent works have focused on this issue trying to effectively summarize lengthy inputs. However, this still is a challenge which needs more attention. In addition to this, there is not enough annotated datasets for meetings as most of the meetings performed in industry are proprietary in nature. Lack

of structure is a problem faced in meeting summarization task. There are multiple participants in a meeting and the information is distributed across dialogues between the speakers. The dialogues are more verbose in nature and also suffer from lack of coherency.

## 7 Conclusion

The pandemic era resulted in a rise of remote working and the usage of video conferencing tools. These virtual meetings could be transcribed easily with the help of ASR systems, thereby increasing the demand for automatic summarization systems for meetings. This paper thoroughly surveys meeting summarization work that have happened till date. This literature covers the state-of-the-art extractive summarization and abstractive summarization models for meetings. We also discuss the challenges faced by researchers and shed light into future research work. Some of the future works can focus on the factual inconsistency problem and the summarization of lengthy meeting transcripts. Exploring the the impact of speakers’ roles in the generated summary could also be a potential future direction for researchers.



## References

- Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.
- Chenxin An, Ming Zhong, Zhichao Geng, Jianqiang Yang, and Xipeng Qiu. 2021. Retrievalsum: A retrieval enhanced framework for abstractive summarization. *arXiv preprint arXiv:2109.07943*.
- Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015a. Abstractive meeting summarization using dependency graph fusion. In *Proceedings of the 24th International Conference on World Wide Web*, pages 5–6.
- Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015b. Generating abstractive summaries from meeting transcripts. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, pages 51–60.
- Nicole Beckage, Shachi H Kumar, Saurav Sahay, and Ramesh Manuvinakurike. 2021. Context or no context? a preliminary exploration of human-in-the-loop approach for incremental temporal summarization in meetings. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 96–106.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Mohammad Hadi Bokaei, Hossein Sameti, and Yang Liu. 2016. Extractive summarization of multi-party meetings through discourse segmentation. *Natural Language Engineering*, 22(1):41–72.
- Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. *arXiv preprint arXiv:2004.15011*.
- Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.
- Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.
- Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.
- Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. *arXiv preprint arXiv:2010.01672*.
- Yun-Nung Chen and Florian Metze. 2012. Two-layer mutually reinforced random walk for improved multi-party meeting summarization. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 461–466. IEEE.
- Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.
- Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343.
- Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.
- Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.
- Alexander R Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. *arXiv preprint arXiv:2106.00829*.
- Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.
- Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2020a. Dialogue discourse-aware graph model and data augmentation for meeting summarization. *arXiv preprint arXiv:2012.03502*.
- Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2020b. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. *arXiv preprint arXiv:2010.10044*.
- Katja Filippova. 2010. Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*, pages 322–330.
- Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569.

- Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66.
- Prakhar Ganesh and Saket Dingliwal. 2019. Restructuring conversations using discourse relations for zero-shot abstractive dialogue summarization. *arXiv preprint arXiv:1902.01615*.
- Nikhil Garg, Benoit Favre, Korbinian Reidhammer, and Dilek Hakkani Tür. 2009. Clusterrank: a graph based method for meeting summarization. Technical report, Idiap.
- Dan Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. 2009. A global optimization framework for meeting summarization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4769–4772. IEEE.
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsun corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.
- Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742. IEEE.
- Chulaka Gunasekara, Guy Feigenblat, Benjamin Szneider, Sachindra Joshi, and David Konopnicki. 2021. Summary grounded conversation generation. *arXiv preprint arXiv:2106.03337*.
- Michael Heilman and Noah A Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617.
- Eduard Hovy and Daniel Marcu. 2005. Automated text summarization. *The Oxford Handbook of computational linguistics*, 583598.
- Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.
- Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.
- Jia Jin Koay, Alexander Roustai, Xiaojin Dai, Dillon Burns, Alec Kerrigan, and Fei Liu. 2020a. How domain terminology affects meeting summarization performance. *arXiv preprint arXiv:2011.00692*.
- Jia Jin Koay, Alexander Roustai, Xiaojin Dai, Dillon Burns, Alec Kerrigan, and Fei Liu. 2020b. [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. *arXiv preprint arXiv:2104.12324*.
- Catherine Lai, Jean Carletta, Steve Renals, K Evanini, and K Zechner. 2013. Detecting summarization hot spots in meetings using group level involvement and turn-taking features. In *INTERSPEECH*, pages 2723–2727.
- Dongyub Lee, Jungwoo Lim, Taesun Whang, Chanhee Lee, Seungwoo Cho, Mingun Park, and Heui-Seok Lim. 2021. Capturing speaker incorrectness: Speaker-focused post-correction for abstractive dialogue summarization. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 65–73.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- 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. 2019a. 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.
- Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019b. [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.
- Chin-Yew Lin and Eduard Hovy. 2002. From single to multi-document summarization. In *Proceedings of*

- the 40th annual meeting of the association for computational linguistics, pages 457–464.
- Hui Lin, Jeff Bilmes, and Shasha Xie. 2009. Graph-based submodular selection for extractive summarization. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 381–386. IEEE.
- Hui Lin and Jeff A Bilmes. 2012. Learning mixtures of submodular shells with application to document summarization. *arXiv preprint arXiv:1210.4871*.
- Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1957–1965.
- Fei Liu and Yang Liu. 2010. Exploring speaker characteristics for meeting summarization. In *Eleventh Annual Conference of the International Speech Communication Association*.
- Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.
- Zhengyuan Liu and Nancy F Chen. 2021a. Controllable neural dialogue summarization with personal named entity planning. *arXiv preprint arXiv:2109.13070*.
- Zhengyuan Liu and Nancy F Chen. 2021b. Dynamic sliding window for meeting summarization. *arXiv preprint arXiv:2108.13629*.
- Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *European Conference on Information Retrieval*, pages 557–564. Springer.
- Yashar Mehdad, Giuseppe Carenini, Frank Tompa, and Raymond Ng. 2013. Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 136–146.
- Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Aaai*, volume 6, pages 775–780.
- Rada Mihalcea and Paul Tarau. 2004. Texttrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.
- Sudipto Mukherjee, Subhabrata Mukherjee, Marcello Hasegawa, Ahmed Hassan Awadallah, and Ryen White. 2020. Smart to-do: Automatic generation of to-do items from emails. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8680–8689, Online. Association for Computational Linguistics.
- Smaranda Muresan, Evelyne Tzoukermann, and Judith L Klavans. 2001. Combining linguistic and machine learning techniques for email summarization. In *Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning (ConLL)*.
- 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*.
- Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings.
- Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.
- 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. *arXiv preprint arXiv:1808.08745*.
- Shashi Narayan, Yao Zhao, Joshua Maynez, Gonalo Simoes, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *arXiv preprint arXiv:2104.07606*.
- Paula S Newman and John C Blitzer. 2003. Summarizing archived discussions: a beginning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 273–276.
- Fumio Nihei, Yukiko I Nakano, and Yutaka Takase. 2018. Fusing verbal and nonverbal information for extractive meeting summarization. In *Proceedings of the Group Interaction Frontiers in Technology*, pages 1–9.
- Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).
- 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.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Owen C Rambow, Lokesh Shrestha, John Chen, and Chirsty Lauridsen. 2004. Summarizing email threads.

- Pengjie Ren, Zhumin Chen, Zhaochun Ren, Furu Wei, Jun Ma, and Maarten de Rijke. 2017. Leveraging contextual sentence relations for extractive summarization using a neural attention model. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 95–104.
- Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. 2008. A keyphrase based approach to interactive meeting summarization. In *2008 IEEE Spoken Language Technology Workshop*, pages 153–156. IEEE.
- Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. 2010. Long story short—global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 52(10):801–815.
- Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.
- Tobias Rohde, Xiaoxia Wu, and Yinhan Liu. 2021. Hierarchical learning for generation with long source sequences. *arXiv preprint arXiv:2104.07545*.
- Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.
- Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Jean-Pierre Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. *arXiv preprint arXiv:1805.05271*.
- Karan Singla, Evgeny Stepanov, Ali Orkan Bayer, Giuseppe Carenini, and Giuseppe Riccardi. 2017. Automatic community creation for abstractive spoken conversations summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 43–47.
- Naama Tepper, Anat Hashavit, Maya Barnea, Inbal Ronen, and Lior Leiba. 2018. Collabot: Personalized group chat summarization. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 771–774.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Jesse Vig, Alexander R Fabbri, and Wojciech Kryściński. 2021. Exploring neural models for query-focused summarization. *arXiv preprint arXiv:2112.07637*.
- Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *arXiv preprint arXiv:1506.03134*.
- Marilyn Walker, Owen Rambow, and Monica Rogati. 2001. Spot: A trainable sentence planner. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
- 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.
- Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable abstractive dialogue summarization with sketch supervision. *arXiv preprint arXiv:2105.14064*.
- Shasha Xie and Yang Liu. 2008. Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4985–4988. IEEE.
- Shasha Xie and Yang Liu. 2010. Improving supervised learning for meeting summarization using sampling and regression. *Computer Speech & Language*, 24(3):495–514.
- Shasha Xie, Yang Liu, and Hui Lin. 2008. Evaluating the effectiveness of features and sampling in extractive meeting summarization. In *2008 IEEE Spoken Language Technology Workshop*, pages 157–160. IEEE.
- Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7386–7393.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.
- Rui Zhang and Joel Tetreault. 2019. This email could save your life: Introducing the task of email subject line generation. *arXiv preprint arXiv:1906.03497*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed H Awadallah, Dragomir Radev, and Rui Zhang. 2021a. Summ<sup>n</sup>: A multi-stage summarization framework for long input dialogues and documents. *arXiv preprint arXiv:2110.10150*.



- Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir Radev. 2021b. An exploratory study on long dialogue summarization: What works and what’s next. *arXiv preprint arXiv:2109.04609*.
- Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449.
- Lulu Zhao, Fujia Zheng, Keqing He, Weihao Zeng, Yuejie Lei, Huixing Jiang, Wei Wu, Weiran Xu, Jun Guo, and Fanyu Meng. 2021. Todsum: Task-oriented dialogue summarization with state tracking. *arXiv preprint arXiv:2110.12680*.
- Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference*, pages 3455–3461.
- Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021a. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *arXiv preprint arXiv:2109.02492*.
- Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021b. 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. *arXiv preprint arXiv:2103.06410*.
- Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. *arXiv preprint arXiv:2004.02016*.
- Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2020. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. *arXiv preprint arXiv:2012.07311*.