A Survey on Dialogue Summarization: Recent Advances and New Frontiers

Xiachong Feng, **Xiaocheng Feng**, **Bing Qin**

Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China

{xiachongfeng, xcfeng, bqin}@ir.hit.edu.cn

Abstract

With the development of dialogue systems and natural language generation techniques, the resurgence of dialogue summarization has attracted significant research attentions, which aims to condense the original dialogue into a shorter version covering salient information. However, there remains a lack of comprehensive survey for this task. To this end, we take the first step and present a thorough review of this research field. In detail, we provide an overview of publicly available research datasets, summarize existing works according to the domain of input dialogue as well as organize leaderboards under unified metrics. Furthermore, we discuss some future directions and give our thoughts. We hope that this first survey of dialogue summarization can provide the community with a quick access and a general picture to this task and motivate future researches.

1 Introduction

Dialogue summarization aims to distill the most important information from a dialogue into a shorter passage, which can help people quickly capture the highlights of a semi-structured and multi-participant dialogue without reviewing the complex dialogue context [Gurevych and Strube, 2004].

Conventional works mainly focus on single-participant document summarization, such as news and scientific papers [Nallapati *et al.*, 2016; Cohan *et al.*, 2018]. Thanks to the neural models, especially the sophisticated pre-trained language models, which have advanced these tasks significantly [Dong *et al.*, 2019]. Despite the success of single-participant document summarization, these methods can not be easily transferred to the multi-participant dialogue summarization. Firstly, the key information of one dialogue is often scattered and spanned multiple utterances and turns from different participants, which lead to low information density. Secondly, the dialogue contains multiple participants, inherently topic drifts, frequent coreferences, diverse interactive signals and domain terminologies. Both above characteristics make dialogue summarization a challenging task.

To this end, recent researchers put their efforts on solving this challenging task and more than 60 papers cover-
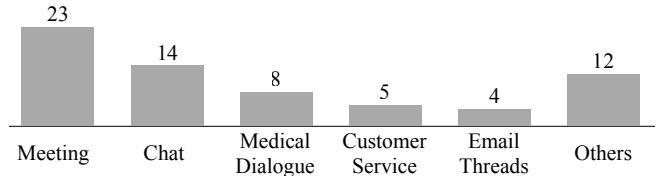

Figure 1: The number of dialogue summarization papers published over the past 5 years for each domain.

ing various dialogue domains have been published over the past 5 years, as shown in Figure 1. To review the current progress and help new researchers, we present this first survey for dialogue summarization. As the preliminary, we quickly overview the recent progress in general summarization and capture several key time points and key techniques, this serves as a strong background before we dive into the dialogue summarization. As the core content, we provide an overview of publicly available research datasets and summarize existing works according to the domain of dialogue, mainly covering meetings, chats, email threads, customer service and medical dialogues. For each type of dialogue, we thoroughly go through related research works, organize them according to their unique challenges and provide suggestions for future works. Besides, we also briefly describe other types of dialogues, such as podcasts and debates. Especially for chat and meeting summarization, we also carefully organize leaderboards under the unified evaluation metric. Based on analyses of existing works, we present several research directions, including faithfulness in dialogue summarization, multi-modal and multi-domain dialogue summarization.

To sum up, our contributions are as follows: (1) we are the first to present a comprehensive survey for dialogue summarization; (2) we thoroughly summarize existing works according to different types of dialogues and carefully organize leaderboards under the unified evaluation metric; (3) we suggest some new frontiers and highlight their challenges.

The rest of this survey is organized as follows. In section 2, we outline the background of dialogue summarization. In section 3, we summarize the existing works and organize leaderboards. In section 4, we discusses the new frontiers for this task. Finally, we give the conclusions in section 5.

2 Background

In this section, we give an overview on the summarization task, then describe the commonly used evaluation metrics.

2.1 Overview of Summarization

Automatic summarization is a fundamental task in natural language processing and has been continuously studied for decades [Paice, 1990; Kupiec *et al.*, 1999]. It aims to condense the original input into a shorter version covering salient information, which can help people quickly grasp the core content without diving into the details. It is mainly divided into two paradigms.: extractive and abstractive. Extractive methods select vital sentences as the summary, which is more accurate and faithful, while abstractive methods generate the summary using novel words, which improves the conciseness and fluency of the summary. Previous works adopt machine learning algorithms to perform extractive summarization [Carbonell and Goldstein-Stewart, 1998; Radev *et al.*, 2000; Erkan and Radev, 2004; Mihalcea and Tarau, 2004]. With sophisticated neural architectures, data-driven approaches have made much progress in both two paradigms [Nallapati *et al.*, 2016; Nallapati *et al.*, 2017]. Especially for abstractive methods, sequence-to-sequence learning combined with attention mechanism are adopted as the backbone architecture for solving this task [See *et al.*, 2017; Zhou *et al.*, 2018]. Recently, with the great success of pre-trained models in a wide range of natural language processing tasks, these models also become the *de facto* way for summary generation and have achieved many state-of-the-art results [Liu and Lapata, 2019; Lewis *et al.*, 2020; Zhang *et al.*, 2020a; Raffel *et al.*, 2020; Dong *et al.*, 2019; Bi *et al.*, 2020]. Moreover, new training methods [Liu and Liu, 2021] and task formulations [Liu *et al.*, 2021a] are proposed, which bring new insights to the summarization task.

2.2 Evaluation Metrics

ROUGE [Lin, 2004] is conventionally adopted as the standard metric for evaluating summarization tasks, which mainly involves the F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L that measure the word-overlap, bigram-overlap and longest common sequence between the ground truth and the generated summary respectively. Instead of exact surface form matching, recently, some works propose new metrics based on contextualized representations, such as BERTScore [Zhang *et al.*, 2020c] and MoverScore [Zhao *et al.*, 2019a], which show higher correlation with human judgements.

3 Taxonomy

In this section, we describe the taxonomy of dialogue summarization according to the domain of input dialogue, including meeting, chat, email thread, customer service, medical dialogue and other types of dialogue. Table 1 lists currently available datasets for dialogue summarization.

3.1 Meeting Summarization

Meeting plays an essential part in our daily life, which allows us to share information and collaborate with others. With the

Name	Domain	Language
ICSI [Janin *et al.*, 2003]	Meeting	English
AMI [Carletta *et al.*, 2005]	Meeting	English
QMSum [Zhong *et al.*, 2021]	Meeting	English
SUMMSCREEN [Chen *et al.*, 2021a]	TV Show	English
CRD3 [Rameshkumar and Bailey, 2020]	TV Show	English
SAMSum [Gliwa *et al.*, 2019]	Chat	English
GupShup [Mehnaz *et al.*, 2021]	Chat	Hindi-English
ADSC [Misra *et al.*, 2015]	Debate	English
[Song *et al.*, 2020]	Medical	Chinese
SumTitles [Malykh *et al.*, 2020]	Movie	English
LCSPIRT [Xi *et al.*, 2020]	Police	Chinese
MEDIASUM [Zhu *et al.*, 2021]	Interview	English
DIALOGSUM [Chen *et al.*, 2021b]	Spoken	English
EMAILSUM [Zhang *et al.*, 2020b]	Email	English
ConvoSumm [Fabbri *et al.*, 2021]	Mix	English

Table 1: Major datasets for dialogue summarization.

growing number of meetings, meeting summaries, aka meeting minutes could be of great value for both participants and non-participants to quickly grasp the main ideas. Thanks to the earlier publicly available datasets AMI [Carletta *et al.*, 2005] and ICSI [Janin *et al.*, 2003], meeting summarization has attracted extensive research attentions.

Precedent works focus on extractive meeting summarization. They adopt various features to detect importance utterances, such key phrases [Riedhammer *et al.*, 2008; Gillick *et al.*, 2009], topics [Xie *et al.*, 2008] and speaker characteristics [Liu and Liu, 2010]. However, due to the multi-participants nature, information is scattered and incoherent in the meeting, which makes the extractive methods unsuitable for meeting summarization [Murray *et al.*, 2010]. Therefore, Banerjee *et al.* [2015] turn towards abstractive methods and propose to fuse important contents based on the dependency graph. Singla *et al.* [2017] design automatic community creation method and use templates to get the final summary for each community. Shang *et al.* [2018] propose a graph-based method for summary generation in an unsupervised manner.

With the development of neural networks, many works have explored the application of deep learning in meeting summarization task and have achieve remarkable success [Jacquenet *et al.*, 2019; Doan *et al.*, 2020; Dammak and Benayed, 2021]. Although deep learning-based methods have strong modeling abilities, taking only literally information into consideration is not sufficient. This is because there are diverse interactive signals among meeting utterances and the long meeting transcripts further pose challenges to traditional sequence-to-sequence models. To this end, some works devote efforts to incorporate auxiliary information for better modeling meetings, such as dialogue discourse [Ganesh and Dingliwal, 2019; Feng *et al.*, 2020a], dialogue acts [Goo and Chen, 2018; Di *et al.*, 2020] and domain terminologies [Koay *et al.*, 2020]. To handle long meeting transcripts, Zhao *et al.* [2019b] propose a hierarchical adaptive segmental encoder-decoder network which automatically segments the meeting into topically consistent parts. Zhu *et al.* [2020] and Rohde *et al.* [2021] adopt hierarchical architectures, which model the meeting from token-level to turn-level. Koay *et al.* [2021] present a sliding-window approach to process the lengthy meeting progressively.

Model	AMI			ICSI		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Extractive Methods						
TextRank [Mihalcea and Tarau, 2004]	35.19	6.13	15.70	30.72	4.69	12.97
SummaRunner [Nallapati *et al.*, 2017]	30.98	5.54	13.91	27.60	3.70	12.52
Abstractive Methods						
UNS [Shang *et al.*, 2018]	37.86	7.84	13.72	31.73	5.14	14.50
PGN [See *et al.*, 2017]	42.60	14.01	22.62	35.89	6.92	15.67
Sentence-Gated [Goo and Chen, 2018]	49.29	19.31	24.82	39.37	9.57	17.17
TopicSeg [Li *et al.*, 2019a]	51.53	12.23	25.47	-	-	-
TopicSeg+VFOA [Li *et al.*, 2019a]	53.29	13.51	26.90	-	-	-
HMNet [Zhu *et al.*, 2020]	52.36	18.63	24.00	45.97	10.14	18.54
PGN($\mathcal{D}_{\text{ALL}}$) [Feng *et al.*, 2021]	50.91	17.75	24.59	-	-	-
DDAMS [Feng *et al.*, 2020a]	51.42	20.99	24.89	39.66	10.09	17.53
DDAMS+DDADA [Feng *et al.*, 2020a]	53.15	22.32	25.67	40.41	11.02	19.18
Pre-trained Language Model-based Methods						
Longformer-BART [Fabbri *et al.*, 2021]	54.81	20.83	25.98	43.40	12.19	19.29
Longformer-BART-arg [Fabbri *et al.*, 2021]	55.27	20.89	24.94	44.51	11.80	19.19

Table 2: Leaderboard of meeting summarization task on AMI [Carletta *et al.*, 2005] and ICSI [Janin *et al.*, 2003] datasets. We adopt reported results from published literatures [Feng *et al.*, 2020a] and corresponding publications. The results of Longformer [Fabbri *et al.*, 2021] are obtained by evaluating the output files provided by the author.

Instead of summarizing the whole meeting, generating summaries of a particular aspect of a meeting, such as decisions, actions, ideas and hypotheses, could also be concise and address specific needs. Previous works mainly generate decision summaries for the meeting [Bui *et al.*, 2009; Wang and Cardie, 2011]. Recently, Zhong *et al.* [2021] propose the query-based meeting summarization task, which aims to summarize the specific part of the meeting according the given general query and specific query.

In addition to multi-party and multi-stream characteristics, meeting summarization is also a multi-modal task [Renals, 2011]. Meetings can include various types of non-verbal information that is displayed by the participants, such as audio, visual and motion features. These features may be useful for detecting important utterances in a meeting. Therefore, a majority of works study the extractive multi-modal meeting summarization problem by fusing verbal and non-verbal features to enrich the representation of the utterance [Erol *et al.*, 2003; Murray *et al.*, 2005; Nihei *et al.*, 2018]. Recently, Li *et al.* [2019a] study the abstractive multi-modal meeting summarization problem. They propose to use the non-verbal feature, namely Visual Focus Of Attention (VFOA), to highlight the importance of one utterance.

Leaderboard: To unify this research, Fabbri *et al.* [2021] already attempt to create a benchmark. Based on their efforts, we present a more comprehensive benchmark for AMI and ICSI using *pyrouge* package, as shown in Table 2.

Highlight: Meetings always involve several participants with specific roles. Thus, it is necessary to model such distinctive role characteristics. Beside, the long transcripts also need the model to be capable of handling long sequences. Furthermore, the audio-visual recordings of meetings provide the opportunity for using multi-modal information. However, it is a double-edged sword. The error rate of automatic speech recognition systems and vision tools also pose challenges to the current models, which requires them to be more robust.

3.2 Chat Summarization

Online chat applications have become an indispensable way for people to communicate with each other. Especially due to the spread of COVID-19 worldwide, people are more dependent on online chatting and overwhelmed by massive amounts of chat information. The complex dialogue context poses a challenge to the new participant, since he may be unable to quickly review the main idea of the dialogue. Therefore, summarizing chats becomes a new trending direction.

Gliwa *et al.* [2019] introduce the first high-quality and manually annotated chat summarization corpus, namely, SAMSum, and conduct various baseline experiments, which rapidly sparks this research direction. Chen and Yang [2020] take the first step and propose a multi-view summarizer combing both topic segments and conversational stages. More importantly, they conduct a comprehensive study for challenges in this task, such as diverse interactions between multiple participants and the frequent occurrence of coreferences, which can advance this direction.

To model the interaction, some works adopt graph modeling strategies. Zhao *et al.* [2020] utilize fine-grained topic words as bridges between utterances to construct a topicword guided dialogue graph. Chen and Yang [2021] consider the inter-utterance dialogue discourse structure and intrautterance action triples to explicitly model the interaction. Feng *et al.* [2020b] view commonsense knowledge as cognitive interactive signals behind the dialogue and shows the effectiveness of integration of knowledge and heterogeneity modeling for different types of data.

Due to the multiple participants nature and frequent occurrence of coreferences, model-generated dialogue summaries always suffer from the factual inconsistent problem [Gabriel *et al.*, 2020]. To this end, Lei *et al.* [2021] put emphasis on modeling complex relationships among participants and their relative personal pronouns via speaker-aware self-attention mechanism. Liu *et al.* [2021b] explicitly incorporate coref-

erence information in dialogue summarization models. It is worth noting that they conduct data postprocessing to reduce incorrect coreference assignments caused by document coreference resolution model. From another perspective, Narayan *et al.* [2021] and Wu *et al.* [2021] both improve the factual consistency of dialogue summaries via coarse-to-fine generation. The final dialogue summary is controlled by a precedent, such as a sketch or an entity chain.

Since current dialogue summarization systems usually encode the text with additional information, Feng *et al.* [2021] propose an unsupervised DialoGPT annotator, which can perform three annotation tasks, including keywords extraction, redundancy detection and topic segmentation.

Even though SAMSum has become the benchmark for dialogue summarization, it can also be extended to others research directions. Gunasekara *et al.* [2021] explore the summary-to-dialogue generation problem and verify the augmented dialogue-summary pairs can do good to dialogue summarization. Mehnaz *et al.* [2021] transform the English dialogues in the SAMSum into Hindi-English dialogues and study chat summarization under the code-switch setting.

Leaderboard: Previous works have already achieved remarkable success on SAMSum dataset. However, due to the different versions of ROUGE evaluation package, there lacks benchmark results unifying all the scores. To this end, we present benchmark results using *py-rouge* package. The results are shown in Table 3.

Highlight: Thanks to the pre-trained language models, current methods are skilled at transforming the original chat into a simple summary realization. However, they still have difficulty to select the important parts and tend to generate hallucinations. In the future, powerful chat modeling strategies and reasoning abilities should be explored for this task.

3.3 Email Threads Summarization

Email thread is an asynchronous multi-party communication consist of coherent exchange of email messages among several participants, which is widely used in enterprise, academic and work settings. Compare with other types of dialogue, email has some unique characteristics. Firstly, it associates with the meta-data, including sender, receiver, main body and signature. Secondly, the email message always represents the intent of the sender, contains action items and may use quote to highlight the important part. Thirdly, unlike face-to-face spoken dialogue, replies in the email do not happen immediately. Such asynchronous nature may result in messages containing long sentences. To deal with email overload, email service providers seek for efficient summarization techniques to improve the user experience.

As a fundamental step, some works focus on single-email summarization. Muresan *et al.* [2001] use linguistic and content features to extract noun phrases as the email summary. Zhang and Tetreault [2019] study the task of email subject line generation given the email body.

Major efforts lie on email thread summarization. Pioneer works tackle this problem via extractive methods. Newman and Blitzer [2003] first group message sentences into topic clusters and then extract summaries from each cluster.

Model	R-1	R-2	R-L
Extractive Methods			
LONGEST-3	32.46	10.27	29.92
TextRank [Mihalcea and Tarau, 2004]	29.27	8.02	28.78
Abstractive Methods			
DynamicConv [Wu *et al.*, 2019]	33.69	10.88	30.93
Transformer [Vaswani *et al.*, 2017]	36.62	11.18	33.06
PGN [See *et al.*, 2017]	40.08	15.28	36.63
Fast Abs RL [Chen and Bansal, 2018]	41.95	18.06	39.23
D-HGN [Feng *et al.*, 2020b]	42.03	18.07	39.56
TGDGA [Zhao *et al.*, 2020]	43.11	19.15	40.49
Pre-trained Language Model-based Methods			
DialoGPT [Zhang *et al.*, 2020d]	39.77	16.58	38.42
UniLM [Dong *et al.*, 2019]	47.85	24.23	46.67
PEGASUS [Zhang *et al.*, 2020a]	50.50	27.23	49.32
BART [Lewis *et al.*, 2020]	52.98	27.67	49.06
S-BART [Chen and Yang, 2021]	50.70	25.50	48.08
FROST [Narayan *et al.*, 2021]	51.86	27.67	47.52
CODS [Wu *et al.*, 2021]	52.65	27.84	50.79
MV-BART [Chen and Yang, 2020]	53.42	27.98	49.97
BART($\mathcal{D}_{ALL}$) [Feng *et al.*, 2021]	53.70	28.79	50.81

Table 3: Leaderboard of chat summarization task on SAMSum [Gliwa *et al.*, 2019] dataset, where "R" is short for "ROUGE". We adopt reported results from published literatures [Gliwa *et al.*, 2019; Wu *et al.*, 2021] and corresponding publications. The results of MV-BART [Chen and Yang, 2020] are obtained via running the open-source code. The results of S-BART [Chen and Yang, 2021] are obtained by evaluating the output file provided by the author.

Nenkova and Bagga [2004] extract sentences from the root message and from each response to the root as the summary. They also reveal the importance of the subject of root email. Rambow *et al.* [2004] use content and structure features to select salient sentences as the summary.

On the basis of the study by the predecessors, some works present publicly available datasets to further facilitate this task. Carenini *et al.* [2007] collect 39 email threads from Enron email dataset [Klimt and Yang, 2004] and annotate them with extractive summaries. They propose an email fragment quotation graph based on the occurrence of clue words and conduct extractive summarization. Quotation plays an important role in the email that can directly highlight the salient part of the previous email. Kano *et al.* [2020] propose an unsupervised methods to predict implicit quotes as the summary. Ulrich *et al.* [2008] collect 40 email threads from W3C email dataset [Craswell *et al.*, 2005] and annotate them with abstractive and extractive summaries along with meta sentences, subjectivity and speech acts. Loza *et al.* [2014] collect 107 email threads from Enron email dataset [Klimt and Yang, 2004] and annotate them with extractive summaries, abstractive summaries and key phrases. Recently, Zhang *et al.* [2020b] present EMAILSUM, which contains 2549 email threads collected from Avocado Research Email Collection [Oard *et al.*,] associated with human-written short and long abstractive summaries. This dataset provides opportunities to data-hungry neural models.

Due to emails are always used for workflow organization and task tracking, some works explore action-focused email summarization, aka TO-DO item generation [Corston-Oliver

et al., 2004]. Mukherjee *et al.* [2020] propose Smart TO-DO system, which first detects commitment sentences and then generates to-do items using seq2seq models.

Highlight: Email is a specific genre of dialogue, which aims to organize the workflow. Therefore, an email frequently proposes requests, makes commitments and contains action items, which make the email intent understanding of vital importance. Future works should pay more attention to understand the fine-grained action items in the email and coarse-grained intent of the entire email.

3.4 Customer Service Summarization

Customer service is the direct one-on-one interaction between a customer and an agent, which frequently happens before and after the consumer behavior. Thus, it is important for growing business. To make the customer service more effective, automatic summarization is one way, which can provide the agent with quick solutions according to previous condensed summary. Therefore, customer service summarization gains a lot of research interest in recent years.

On the one hand, participants in customer service have strong intents and clear motivations to address issues, which makes the customer service inherently logical and surrounds specific topics. To this end, some works explore topic modeling for this task. Liu *et al.* [2019a] employ a coarse-to-fine generation framework, which first generates a sequence of key points (topics) to indicate the logic of the dialogue and then realize the detailed summary. For example, a key point sequence can be *question→solution→user approval→end*, which clearly shows the evolution of the dialogue. Instead of using explicitly pre-defined topics, Zou *et al.* [2021b] draw support from neural topic modeling and propose a multi-role topic modeling mechanism to explore implicitly topics. To alleviate data-insufficient problems, Zou *et al.* [2021a] propose an unsupervised framework called RankAE, in which topic utterances are first selected according to centrality and diversity simultaneously, and the denoising auto-encoder is further employed to produce final summaries.

On the other hand, customer service is a kind of task-oriented dialogues, which contains informative entities, covers various domains and involves two distinct types of participants. To integrate various information, Yuan and Yu [2019] propose Scaffold Pointer Network to utilize three information, including speaker role, semantic slot and dialogue domain. Previous works generate the summary in the third-person point of view. Since participants in customer service play distinct roles, Zhang *et al.* [2021] propose an unsupervised framework based on variational auto-encoder to generate summaries for the customer and the agent respectively.

Highlight: Customer service aims to address questions raised by agents. Therefore, it naturally has strong motivations, which makes the dialogue have a specific way of evolution following the interaction between two participants with distinctive characteristics: the customer and the agent. Thus, modeling participant roles, evolution chains and inherent topics are important for this task. Beside, some fine-grained information also should be taken into consideration, such as slots and intents [Qin *et al.*, 2021].

3.5 Medical Dialogue Summarization

Medical dialogues happen between patients and doctors. During this process, doctors are required to record digital version of a patient's health records, namely electronic health records (EHR), which leads to both patient dis-satisfaction and clinician burnout. To mitigate the above challenge, medical dialogue summarization is coming to the rescue.

From a coarse-grained perspective, a medical dialogue can be divided into several coherent segments according to different criteria. Liu *et al.* [2019b] specify the dialogue topics according to the symptoms, such as headache and cough, and design a topic-level attention mechanism to make the decoder focus on one symptom when generating one summary sentence. Kazi and Kahanda [2019] instead choose EHR categories to label each segment, such as family history and medical history. Specifically, Krishna *et al.* [2020] name the medical dialogue summary *SOAP note*, which stands for Subjective information reported by the patient; Objective observations; Assessments made by the doctor; and a Plan for future care, including diagnostic tests and treatments.

From a fine-grained perspective, several medical dialogue characteristics should be handled carefully. Firstly, question-answer pairs are the major discourse in medical dialogues and negations scatter in different utterances are notable parts. To this end, [Joshi *et al.*, 2020] encourage the model to focus on negation words via negation word attention and explicitly employ a gate mechanism to generate the [NO] word. Secondly, medical terminologies play an essential part of medical dialogues. Joshi *et al.* [2020] leverage compendium of medical concepts, known as unified medical language systems to identify the presence of terminologies and further use an indicator vector to influence the attention distribution. Thirdly, medical dialogue summary mainly describes core items and concepts in the dialogue, therefore, the summarization methods should bias towards extractive methods while keeping advantages of abstractive methods. Enarvi *et al.* [2020] and Joshi *et al.* [2020] both enhance the copy mechanism to facilitate copying from the input.

Highlight: Medical dialogue summarization mainly aims at helping doctors to quickly finish electronic health records and the medical dialogue summary should be more faithful rather than creative. Therefore, extractive methods combined with simple abstractive methods are preferred. The topic information can serve as a guideline for generating semi-structured summaries. Besides, terminologies and negations in the medical dialogue should be handled carefully.

3.6 Other Types of Dialogue Summarization

A dialogue is any discourse produced by more than one person [Ford, 1991]. In addition to the types of dialogues mentioned above, previous works also tackle the podcasts [Zheng *et al.*, 2020], online discussion forums [Tarnpradab *et al.*, 2021], legal debates [Duan *et al.*, 2019] and reader comment threads [Barker *et al.*, 2016] summarization tasks. Towards a more practical direction, Tepper *et al.* [2018] propose the personalized chat summarization task, which implicitly learns user interests via topic distribution and social graph connections, based on which, it provides a personalized summary.

4 New Frontiers

Section 3 mainly introduces single-domain and single-modal dialogue summarization tasks. In this section, we will discuss some new frontiers which meet actual application needs and fit in with real-world scenarios.

4.1 Faithfulness in Dialogue Summarization

Even though current state-of-the-art summarization systems have already made great progress, they suffer from the factual inconsistency problem, which distorts or fabricates the factual information in the article and is also known as hallucinations [Huang *et al.*, 2021]. Chen and Yang [2020] point out that the wrong reference is one of the main errors made by the dialogue summarization model, which means the generated summaries contain information that is not faithful to the original dialogue (e.g., associate one's actions or locations with a wrong speaker), as shown in Figure 2. This error largely hinders the application of dialogue summarization systems.

We argue that this problem is mainly caused by the multiple participants and diverse references in the dialogue. To remedy this issue, personal pronoun information [Lei *et al.*, 2021] and coreference information [Liu *et al.*, 2021b] are explicitly incorporated into the summarization model. Even so, the lack of high-quality coreference resolution model for dialogues results in low-quality resolution results, which further hinders the quality of summaries.

To mitigate the challenge, we can not only enhance the coreference resolution model with dialogue features, but also implicitly model the coreference by using contextual and discourse information.

4.2 Multi-modal Dialogue Summarization

Dialogues tend to occur in multi-modal situations, such as audio-visual recordings of meetings. Besides verbal information, non-verbal information can either supplement existing information or provide new information, which effectively enriches the representation of purely textual dialogues. According to whether different modalities can be aligned, the types of multi-modal information can be divided into two categories: synchronous and asynchronous.

Synchronous multi-modal dialogues mainly refer to meetings, which may contain textual transcripts, prosodic audios and visual videos. On the one hand, taking the aligned audio and video into consideration can enhance the representation of transcripts. On the other hand, both the audio and video can provide new insights, such as a person entering the room to join the meeting or an emotional discussion. However, facial features and voice print features have already become superior privacy for individuals, which makes them hard and sensitive to be acquired. Future works can consider multi-modal meeting summarization under the federal learning framework [Li *et al.*, 2019b].

Asynchronous multi-modal dialogues refer to different modalities happen at different times. With the development of communication technology, multi-modal messages, such as voice messages, pictures and emoji are frequently used in chat dialogues via applications like Messenger, WhatsApp and WeChat. These messages provide rich information, serving as one part of the dialogue flow. Future works should

Dialogue		Summary
Mary:	Are you going by car or train?	Tom rented a car. ✓
Tom:	I rented a car	
Mary:	This makes all of this much faster	**Mary** rented a car. ✗

Figure 2: Example dialogue with two summaries. The first summary is factual consistent. The second one is unfaithful to the dialogue, since it associates *rent car* with the wrong person *Mary*.

consider textual information of voice messages obtained via ASR systems, new entities provided by pictures and emotions associated with the emoji to produce meaningful summaries.

4.3 Multi-domain Dialogue Summarization

Multi-domain learning can mine shared information between different domains and further help the task of a specific domain, which is an effective learning method suitable for low-resource scenarios. Thanks to diverse summarization datasets, there are already some works explore the multi-domain learning or domain adaption for dialogue summarization [Sandu *et al.*, 2010; Zhu *et al.*, 2020; Yu *et al.*, 2021]. We divide this direction into two categories: macro multi-domain learning and micro multi-domain learning.

Macro multi-domain learning aims to use general domain summarization datasets, like news and scientific papers, to help the dialogue summarization task. The basis for this learning method to work is that no matter what data type they belong to, they aim to pick the core content of the original text. However, dialogues have some unique characteristics like more coreferences and participant-related features. Therefore, directly using these general datasets may reduce their effectiveness. Future works can first inject some dialogue specific features, like replacing names with personal pronouns, or transform the original general domain documents into turn-level documents at surface level.

Micro multi-domain learning aims to use dialogue summarization datasets to help one specific dialogue summarization task. For example, using meeting datasets to help email tasks. As shown in Table 1, diverse dialogue summarization datasets covering various domains have been proposed recent years. Future works can adopt meta-learning methods or rely on pre-trained language models to use different datasets.

5 Conclusion

This article presents the first comprehensive survey on the progress of dialogue summarization. We thoroughly summarize the existing works, highlight their challenges and provide leaderboards. Furthermore, we shed light on some new trends in this research field. We hope this survey can facilitate the research of the dialogue summarization.

References

[Banerjee *et al.*, 2015] Siddhartha Banerjee, P. Mitra, and Kazunari Sugiyama. Abstractive meeting summarization using dependency graph fusion. *Proceedings of the 24th International Conference on World Wide Web*, 2015.

[Barker *et al.*, 2016] Emma Barker, Monica Lestari Paramita, Ahmet Aker, Emina Kurtic, Mark Hepple, and Robert Gaizauskas. The SENSEI annotated corpus: Human summaries of reader comment conversations in on-line news. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 42–52, Los Angeles, 2016. Association for Computational Linguistics.

[Bi *et al.*, 2020] Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. PALM: Pretraining an autoencoding&autoregressive language model for context-conditioned generation. In *Proc. of EMNLP*, Online, 2020. Association for Computational Linguistics.

[Bui *et al.*, 2009] Trung Bui, Matthew Frampton, John Dowding, and Stanley Peters. Extracting decisions from multi-party dialogue using directed graphical models and semantic similarity. In *Proceedings of the SIGDIAL 2009 Conference*, pages 235–243, London, UK, 2009. Association for Computational Linguistics.

[Carbonell and Goldstein-Stewart, 1998] J. Carbonell and Jade Goldstein-Stewart. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98*, 1998.

[Carenini *et al.*, 2007] Giuseppe Carenini, Raymond T. Ng, and Xiaodong Zhou. Summarizing email conversations with clue words. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 91–100. ACM, 2007.

[Carletta *et al.*, 2005] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*. Springer, 2005.

[Chen and Bansal, 2018] Yen-Chun Chen and Mohit Bansal. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proc. of ACL*, pages 675–686, Melbourne, Australia, 2018. Association for Computational Linguistics.

[Chen and Yang, 2020] Jiaao Chen and Diyi Yang. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proc. of EMNLP*, Online, 2020. Association for Computational Linguistics.

[Chen and Yang, 2021] Jiaao Chen and Diyi Yang. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online, 2021. Association for Computational Linguistics.

[Chen *et al.*, 2021a] Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. Summscreen: A dataset for abstractive screenplay summarization. *arXiv preprint arXiv:2104.07091*, 2021.

[Chen *et al.*, 2021b] Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. Dialsumm: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*, 2021.

[Cohan *et al.*, 2018] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In *Proc. of NAACL-HLT*, pages 615–621, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[Corston-Oliver *et al.*, 2004] Simon Corston-Oliver, Eric Ringger, Michael Gamon, and Richard Campbell. Task-focused summarization of email. In *Text Summarization Branches Out*, pages 43–50, Barcelona, Spain, 2004. Association for Computational Linguistics.

[Craswell *et al.*, 2005] Nick Craswell, Arjen P De Vries, and Ian Soboroff. Overview of the trec 2005 enterprise track. In *Trec*, volume 5, pages 1–7, 2005.

[Dammak and Benayed, 2021] Nouha Dammak and Y. Benayed. Abstractive meeting summarization based on an attentional neural model. In *International Conference on Machine Vision*, 2021.

[Di *et al.*, 2020] Jiasheng Di, X. Wei, and Zhenyu Zhang. How to interact and change? abstractive dialogue summarization with dialogue act weight and topic change info. In *KSEM*, 2020.

[Doan *et al.*, 2020] Tung Doan, F. Jacquenet, Christine Largeron, and Marc Bernard. A study of text summarization techniques for generating meeting minutes. In *RCIS*, 2020.

[Dong *et al.*, 2019] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054, 2019.

[Duan *et al.*, 2019] Xinyu Duan, Yating Zhang, Lin Yuan, Xin Zhou, Xiaozhong Liu, Tianyi Wang, Ruocheng Wang, Qiong Zhang, Changlong Sun, and Fei Wu. Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning. In Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu, editors, *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*. ACM, 2019.

[Enarvi *et al.*, 2020] Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online, 2020. Association for Computational Linguistics.

[Erkan and Radev, 2004] Günes Erkan and Dragomir R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479, 2004.

[Erol *et al.*, 2003] B. Erol, Dar-Shyang Lee, and J. Hull. Multimodal summarization of meeting recordings. *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, 3:III–25, 2003.

[Fabbri *et al.*, 2021] Alexander R Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. Convosumm: Conversation summarization benchmark

and improved abstractive summarization with argument mining. *arXiv preprint arXiv:2106.00829*, 2021.

[Feng *et al.*, 2020a] Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. Dialogue discourse-aware graph model and data augmentation for meeting summarization. 2020.

[Feng *et al.*, 2020b] Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. *arXiv preprint arXiv:2010.10044*, 2020.

[Feng *et al.*, 2021] Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. Language model as an annotator: Exploring dialogpt for dialogue summarization. *arXiv preprint arXiv:2105.12544*, 2021.

[Ford, 1991] Cecilia E Ford. Linguistics: The cambridge survey: Volume 4. language: The socio-cultural context. frederick h. newmeyer (ed.). *Studies in Second Language Acquisition*, 13(3):412–413, 1991.

[Gabriel *et al.*, 2020] Saadia Gabriel, A. Çelikyilmaz, R. Jha, Yejin Choi, and Jianfeng Gao. Go figure! a meta evaluation of factuality in summarization. *ArXiv*, abs/2010.12834, 2020.

[Ganesh and Dingliwal, 2019] Prakhar Ganesh and Saket Dingliwal. Restructuring conversations using discourse relations for zero-shot abstractive dialogue summarization. *arXiv: Computation and Language*, 2019.

[Gillick *et al.*, 2009] D. Gillick, K. Riedhammer, B. Favre, and Dilek Z. Hakkani-Tür. A global optimization framework for meeting summarization. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

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

[Goo and Chen, 2018] Chih-Wen Goo and Yun-Nung Chen. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742, 2018.

[Gunasekara *et al.*, 2021] R. Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, S. Joshi, and D. Konopnicki. Summary grounded conversation generation. *ArXiv*, abs/2106.03337, 2021.

[Gurevych and Strube, 2004] Iryna Gurevych and Michael Strube. Semantic similarity applied to spoken dialogue summarization. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 764–770, Geneva, Switzerland, 2004. COLING.

[Huang *et al.*, 2021] Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*, 2021.

[Jacquenet *et al.*, 2019] F. Jacquenet, Marc Bernard, and Christine Largeron. Meeting summarization, a challenge for deep learning. In *IWANN*, 2019.

[Janin *et al.*, 2003] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. The icsi meeting corpus. In *ICASSP*. IEEE, 2003.

[Joshi *et al.*, 2020] Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online, 2020. Association for Computational Linguistics.

[Kano *et al.*, 2020] Ryuji Kano, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. Identifying implicit quotes for unsupervised extractive summarization of conversations. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 291–302, Suzhou, China, 2020. Association for Computational Linguistics.

[Kazi and Kahanda, 2019] Nazmul Kazi and Indika Kahanda. Automatically generating psychiatric case notes from digital transcripts of doctor-patient conversations. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 140–148, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics.

[Klimt and Yang, 2004] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer, 2004.

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

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

[Krishna *et al.*, 2020] Kundan Krishna, Sopan Khosla, Jeffrey P. Bigham, and Z. Lipton. Generating soap notes from doctor-patient conversations using modular summarization techniques. 2020.

[Kupiec *et al.*, 1999] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. *Advances in Automatic Summarization*, pages 55–60, 1999.

[Lei *et al.*, 2021] Yuejie Lei, Yuanmeng Yan, Zhiyuan Zeng, Keqing He, Ximing Zhang, and Weiran Xu. Hierarchical speaker-aware sequence-to-sequence model for dialogue summarization. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL*, pages 7871–7880, Online, 2020. Association for Computational Linguistics.

[Li *et al.*, 2019a] Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proc. of ACL*, pages 2190–2196, Florence, Italy, 2019. Association for Computational Linguistics.

[Li *et al.*, 2019b] Q. Li, Zeyi Wen, and B. He. Federated learning systems: Vision, hype and reality for data privacy and protection. *ArXiv*, abs/1907.09693, 2019.

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

[Liu and Lapata, 2019] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *Proc. of EMNLP*, pages 3730–3740, Hong Kong, China, 2019. Association for Computational Linguistics.

[Liu and Liu, 2010] Fei Liu and Yang Liu. Exploring speaker characteristics for meeting summarization. In *INTERSPEECH*, 2010.

[Liu and Liu, 2021] Yixin Liu and Peng Liu. Simcls: A simple framework for contrastive learning of abstractive summarization. *ArXiv*, abs/2106.01890, 2021.

[Liu et al., 2019a] Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. Automatic dialogue summary generation for customer service. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 1957–1965. ACM, 2019.

[Liu et al., 2019b] Zhengyuan Liu, A. Ng, Sheldon Lee Shao Guang, AiTi Aw, and Nancy F. Chen. Topic-aware pointer-generator networks for summarizing spoken conversations. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821, 2019.

[Liu et al., 2021a] Yixin Liu, Zi-Yi Dou, and Pengfei Liu. RefSum: Refactoring neural summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1437–1448, Online, 2021. Association for Computational Linguistics.

[Liu et al., 2021b] Zhengyuan Liu, Ke Shi, and Nancy F. Chen. Coreference-aware dialogue summarization. 2021.

[Loza et al., 2014] Vanessa Loza, Shibamouli Lahiri, Rada Mihalcea, and Po-Hsiang Lai. Building a dataset for summarization and keyword extraction from emails. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2441–2446, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).

[Malykh et al., 2020] Valentin Malykh, Konstantin Chernis, Ekaterina Artemova, and Irina Piontkovskaya. SumTitles: a summarization dataset with low extractiveness. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5718–5730, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics.

[Mehnaz et al., 2021] Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle Lee, Anish Acharya, and Rajiv Ratn Shah. Gupshup: An annotated corpus for abstractive summarization of open-domain code-switched conversations. *arXiv preprint arXiv:2104.08578*, 2021.

[Mihalcea and Tarau, 2004] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proc. of EMNLP*, pages 404–411, Barcelona, Spain, 2004. Association for Computational Linguistics.

[Misra et al., 2015] Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. Using summarization to discover argument facets in online idealogical dialog. In *Proc. of NAACL-HLT*, pages 430–440, Denver, Colorado, 2015. Association for Computational Linguistics.

[Mukherjee et al., 2020] Sudipto Mukherjee, Subhabrata Mukherjee, Marcello Hasegawa, Ahmed Hassan Awadallah, and Ryen White. Smart to-do: Automatic generation of to-do items from emails. In *Proc. of ACL*, pages 8680–8689, Online, 2020. Association for Computational Linguistics.

[Muresan et al., 2001] Smaranda Muresan, Evelyne Tzoukermann, and Judith L. Klavans. Combining linguistic and machine learning techniques for email summarization. In *Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning (ConLL)*, 2001.

[Murray et al., 2005] Gabriel Murray, S. Renals, and J. Carletta. Extractive summarization of meeting recordings. In *INTERSPEECH*, 2005.

[Murray et al., 2010] Gabriel Murray, Giuseppe Carenini, and Raymond Ng. Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of the 6th International Natural Language Generation Conference*, 2010.

[Nallapati et al., 2016] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Nallapati et al., 2017] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In Satinder P. Singh and Shaul Markovitch, editors, *Proc. of AAAI*, pages 3075–3081. AAAI Press, 2017.

[Narayan et al., 2021] Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simoes, and Ryan McDonald. Planning with entity chains for abstractive summarization. *arXiv preprint arXiv:2104.07606*, 2021.

[Nenkova and Bagga, 2004] Ani Nenkova and Amit Bagga. Facilitating email thread access by extractive summary generation. *Recent advances in natural language processing III: selected papers from RANLP*, 2003:287–294, 2004.

[Newman and Blitzer, 2003] Paula S Newman and John C Blitzer. Summarizing archived discussions: a beginning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 273–276, 2003.

[Nihei et al., 2018] Fumio Nihei, Y. Nakano, and Yutaka Takase. Fusing verbal and nonverbal information for extractive meeting summarization. *Proceedings of the Group Interaction Frontiers in Technology*, 2018.

[Oard et al.,] Douglas Oard, William Webber, David Kirsch, and Sergey Golitsynskiy. Avocado research email collection ldc2015t03.

[Paice, 1990] Chris D Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 26(1):171–186, 1990.

[Qin et al., 2021] Libo Qin, Tianbao Xie, Wanxiang Che, and T. Liu. A survey on spoken language understanding: Recent advances and new frontiers. *ArXiv*, abs/2103.03095, 2021.

[Radev et al., 2000] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*, 2000.

[Raffel *et al.*, 2020] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, W. Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683, 2020.

[Rambow *et al.*, 2004] Owen Rambow, Lokesh Shrestha, John Chen, and Christy Laurdisen. Summarizing email threads. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 105–108, Boston, Massachusetts, USA, 2004. Association for Computational Linguistics.

[Rameshkumar and Bailey, 2020] Revanth Rameshkumar and Peter Bailey. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proc. of ACL*, pages 5121–5134, Online, 2020. Association for Computational Linguistics.

[Renals, 2011] S. Renals. Automatic analysis of multiparty meetings. *Sadhana*, 36:917–932, 2011.

[Riedhammer *et al.*, 2008] K. Riedhammer, B. Favre, and Dilek Z. Hakkani-Tür. A keyphrase based approach to interactive meeting summarization. *2008 IEEE Spoken Language Technology Workshop*, pages 153–156, 2008.

[Rohde *et al.*, 2021] Tobias Rohde, Xiaoxia Wu, and Yinhan Liu. Hierarchical learning for generation with long source sequences. *ArXiv*, abs/2104.07545, 2021.

[Sandu *et al.*, 2010] Oana Sandu, Giuseppe Carenini, Gabriel Murray, and Raymond Ng. Domain adaptation to summarize human conversations. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 16–22, Uppsala, Sweden, 2010. Association for Computational Linguistics.

[See *et al.*, 2017] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proc. of ACL*, pages 1073–1083, Vancouver, Canada, 2017. Association for Computational Linguistics.

[Shang *et al.*, 2018] Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proc. of ACL*, pages 664–674, Melbourne, Australia, 2018. Association for Computational Linguistics.

[Singla *et al.*, 2017] Karan Singla, Evgeny Stepanov, Ali Orkan Bayer, Giuseppe Carenini, and Giuseppe Riccardi. Automatic community creation for abstractive spoken conversations summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 43–47, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

[Song *et al.*, 2020] Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics.

[Tarnpradab *et al.*, 2021] Sansiri Tarnpradab, Fereshteh Jafariakinabad, and Kien A Hua. Improving online forums summarization via hierarchical unified deep neural network. *arXiv preprint arXiv:2103.13587*, 2021.

[Tepper *et al.*, 2018] Naama Tepper, Anat Hashavit, Maya Barnea, Inbal Ronen, and Lior Leiba. Collabot: Personalized group chat summarization. In Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek, editors, *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 771–774. ACM, 2018.

[Ulrich *et al.*, 2008] Jan Ulrich, Gabriel Murray, and Giuseppe Carenini. A publicly available annotated corpus for supervised email summarization. In *Proc. of aaai email-2008 workshop, chicago, usa*, 2008.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017.

[Wang and Cardie, 2011] Lu Wang and Claire Cardie. Summarizing decisions in spoken meetings. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, pages 16–24, Portland, Oregon, 2011. Association for Computational Linguistics.

[Wu *et al.*, 2019] Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *Proc. of ICLR*. OpenReview.net, 2019.

[Wu *et al.*, 2021] Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. Controllable abstractive dialogue summarization with sketch supervision. *arXiv preprint arXiv:2105.14064*, 2021.

[Xi *et al.*, 2020] Xuefeng Xi, Zhou Pi, and Guodong Zhou. Global encoding for long chinese text summarization. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(6), 2020.

[Xie *et al.*, 2008] Shasha Xie, Yang Liu, and H. Lin. Evaluating the effectiveness of features and sampling in extractive meeting summarization. *2008 IEEE Spoken Language Technology Workshop*, 2008.

[Yu *et al.*, 2021] Tiezheng Yu, Zihan Liu, and Pascale Fung. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5892–5904, Online, 2021. Association for Computational Linguistics.

[Yuan and Yu, 2019] Lin Yuan and Z. Yu. Abstractive dialog summarization with semantic scaffolds. *ArXiv*, abs/1910.00825, 2019.

[Zhang and Tetreault, 2019] Rui Zhang and Joel Tetreault. This email could save your life: Introducing the task of email subject line generation. In *Proc. of ACL*, pages 446–456, Florence, Italy, 2019. Association for Computational Linguistics.

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

[Zhang *et al.*, 2020b] Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. Emailsum: Abstractive email thread summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2020.

[Zhang *et al.*, 2020c] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *Proc. of ICLR*. OpenReview.net, 2020.

[Zhang *et al.*, 2020d] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proc. of ACL*,

pages 270–278, Online, 2020. Association for Computational Linguistics.

[Zhang *et al.*, 2021] Xinyuan Zhang, Ruiyi Zhang, M. Zaheer, and Amr Ahmed. Unsupervised abstractive dialogue summarization for tete-a-tetes. In *AAAI*, 2021.

[Zhao *et al.*, 2019a] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proc. of EMNLP*, pages 563–578, Hong Kong, China, 2019. Association for Computational Linguistics.

[Zhao *et al.*, 2019b] Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, and Min Yang. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3455–3461. ACM, 2019.

[Zhao *et al.*, 2020] Lulu Zhao, Weiran Xu, and Jun Guo. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics.

[Zheng *et al.*, 2020] Chujie Zheng, Kunpeng Zhang, Harry Jiannan Wang, and Ling Fan. A two-phase approach for abstractive podcast summarization. *ArXiv*, abs/2011.08291, 2020.

[Zhong *et al.*, 2021] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QM-Sum: 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*, pages 5905–5921, Online, 2021. Association for Computational Linguistics.

[Zhou *et al.*, 2018] Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. Neural document summarization by jointly learning to score and select sentences. In *Proc. of ACL*, pages 654–663, Melbourne, Australia, 2018. Association for Computational Linguistics.

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

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

[Zou *et al.*, 2021a] Yicheng Zou, Jun Lin, Lujun Zhao, Yangyang Kang, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. Unsupervised summarization for chat logs with topic-oriented ranking and context-aware autoencoders. In *AAAI*, 2021.

[Zou *et al.*, 2021b] Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. In *AAAI*, 2021.