A Survey on Dialogue Summarization: Recent Advances and New Frontiers

Xiachong Feng , Xiaocheng Feng , Bing Qin
Harbin Institute of Technology, China
{xiachongfeng, xcfeng, bqin}@ir.hit.edu.cn

Abstract

Dialogue summarization aims to condense the original dialogue into a shorter version covering salient information, which is a crucial way to reduce dialogue data overload. Recently, the promising achievements in both dialogue systems and natural language generation techniques drastically lead this task to a new landscape, which results in significant research attentions. However, there still remains a lack of a comprehensive survey for this task. To this end, we take the first step and present a thorough review of this research field carefully and widely. In detail, we systematically organize the current works according to the characteristics of each domain, covering meeting, chat, email thread, customer service and medical dialogue. Additionally, we provide an overview of publicly available research datasets as well as organize two leaderboards under unified metrics. Furthermore, we discuss some future directions, including faithfulness, multi-modal, multi-domain and multi-lingual dialogue summarization, and give our thoughts respectively. 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 [Chen and Yang, 2020]. With the development of communication technology and the ravage of COVID-19, different types of dialogues have emerged as an important way for information exchange. Therefore, there is an urgent need for summarization techniques to save people from large amounts of dialogue data.

Conventional works mainly focus on single-participant document summarization, such as news and scientific papers [See *et al.*, 2017]. Thanks to the neural models, especially the sophisticated pre-trained language models, which have advanced these tasks significantly [Lewis *et al.*, 2020]. Despite the success of single-participant document summarization, these methods can not be easily transferred to the multi-participant dialogue summarization. Firstly, the dialogue contains multiple participants, inherent topic drifts, frequent coreferences, diverse interactive signals and domain terminologies [Feng *et al.*, 2021b]. All of these characteristics make dialogue a hard-to-model data type. Secondly, in terms of different domains, the above characteristics further pose domain-specific challenges to summarization models, e.g., *How to model long meeting transcripts* [Zhu *et al.*, 2020]. Thirdly, compared with widely used document summarization benchmarks, collecting labeled dialogue-summary paired data is highly-costing or even intractable [Chen and Yang, 2021a]. To mitigate these challenges, researchers draw on successful experiences from the study of dialogue systems and natural language generation techniques and put their efforts on solving this challenging task, which result in nearly 100 papers covering various domains being published over the past 5 years.

To review the current progress and help new researchers get into the field quickly, 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 (see §2). As the core content, we summarize existing works according to the domain of dialogue, mainly covering the meeting, chat, email thread, customer service and medical dialogue. 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 (see §3). For example, we focus on two main streams of works for chat summarization including interaction and participant modeling [Liu *et al.*, 2021; Liu and Chen, 2021]. In terms of customer service, we organize related works from two perspectives, one is inherent topic modeling [Liu *et al.*, 2019a], the other is task-oriented-specific information integration [Zhao *et al.*, 2021]. Besides, we provide an overview of publicly available research datasets (see Table 1). Especially for meeting and chat summarization, we also carefully organize leaderboards under the unified evaluation metric by collecting reported results from published literatures and re-evaluating official outputs (see Table 2 and Table 3). Based on the analyses of existing works,

we present several research directions, including faithfulness in dialogue summarization, multi-modal, multi-domain and multi-lingual dialogue summarization (see §4). All of these frontiers not only pose new research challenges but also meet actual application needs and fit in with real-world scenarios.

To sum up, our contributions are as follows:

- We are the first to present a comprehensive survey for the dialogue summarization task.
- We thoroughly summarize existing works according to different types of dialogues and carefully organize leaderboards under the unified evaluation metric.
- We discuss some new frontiers and highlight their challenges to motivate future researches.

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]. 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 [Mihalcea and Tarau, 2004]. With sophisticated neural architectures, data-driven approaches have made much progress in both two paradigms. Especially for abstractive methods, sequence-to-sequence learning combined with attention mechanism is adopted as the backbone architecture for solving this task [See *et al.*, 2017]. 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 [Lewis *et al.*, 2020].

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, bi-gram overlap and longest common sequence between the ground truth and the generated summary respectively.

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 and medical dialogue. Table 1 lists currently available datasets for these dialogue summarization researches.

Name	Domain	Language
ICSI [Janin *et al.*, 2003]	Meeting	English
AMI [Carletta *et al.*, 2005]		English
QMSum [Zhong *et al.*, 2021]		English
SAMSum [Gliwa *et al.*, 2019]	Chat	English
GupShup [Mehnaz *et al.*, 2021]		Code-Mix
CSDS [Lin *et al.*, 2021]	Customer Service	Chinese
TODSum [Zhao *et al.*, 2021]		English
TWEETSUMM [Feigenblat *et al.*, 2021]		English
CRD3 [Rameshkumar and Bailey, 2020]	TV Show	English
[Song *et al.*, 2020]	Medical	Chinese
SumTitles [Malykh *et al.*, 2020]	Movie	English
MEDIASUM [Zhu *et al.*, 2021]	Interview	English
DIALOGSUM [Chen *et al.*, 2021]	Spoken	English
EMAILSUM [Zhang *et al.*, 2021a]	Email	English
ForumSum [Khalman *et al.*, 2021]	Forum	English
ConvoSumm [Fabbri *et al.*, 2021]	Mix	English

Table 1: Major datasets for dialogue summarization.

3.1 Meeting Summarization

Meeting plays an essential part in our daily life. Especially due to the spread of COVID-19 worldwide, people are more dependent on online meetings to share information and collaborate with others. Accordingly, meeting summaries, aka meeting minutes could be of great value for both participants and non-participants to quickly grasp the main meeting 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 important utterances, such key phrases, topics and speaker characteristics. However, due to the multi-participants nature, information is scattered and incoherent in the meeting, which makes the extractive methods unsuitable for meeting summarization. Therefore, recent years witness a growing trend of abstractive meeting summarization methods [Shang *et al.*, 2018].

With the development of neural networks, many works have explored the application of deep learning in meeting the summarization task and have achieved remarkable success [Zhu *et al.*, 2020]. Although deep learning-based methods have strong modeling abilities, taking only literal 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 [Feng *et al.*, 2021b], dialogue acts [Goo and Chen, 2018] and domain terminologies [Koay *et al.*, 2020]. Besides, several strategies are carefully devised to handle long meeting transcripts, including hierarchical modeling strategy [Zhu *et al.*, 2020], sliding window strategy [Koay *et al.*, 2021], retrieve-then-summarize strategy [Zhang *et al.*, 2021c] and pre-training strategy [Zhong *et al.*, 2022].

Instead of summarizing the whole meeting, generating meeting summaries of a particular aspect, such as decisions, actions, ideas and hypotheses, could also address specific needs. Recently, Zhong *et al.* [2021] propose the query-based meeting summarization task, which aims to summarize the

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.*, 2019]	51.53	12.23	25.47	-	-	-
TopicSeg+VFOA [Li *et al.*, 2019]	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.*, 2021c]	50.91	17.75	24.59	-	-	-
DDAMS [Feng *et al.*, 2021b]	51.42	20.99	24.89	39.66	10.09	17.53
DDAMS+DDADA [Feng *et al.*, 2021b]	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
DIALOGLM [Zhong *et al.*, 2022]	53.72	19.61	51.83*	49.56	12.53	47.08*

Table 2: Leaderboard of meeting summarization on AMI [Carletta *et al.*, 2005] and ICSI [Janin *et al.*, 2003] datasets. We adopt reported results from published literatures [Feng *et al.*, 2021b] and corresponding publications. The results of Longformer [Fabbri *et al.*, 2021] are obtained by evaluating the output files provided by the author. Results with * indicate that ROUGE-L is calculated with sentence splitting.

specific part of the meeting according to the given query.

In addition to multi-party characteristics, meeting summarization has also been explored under the multi-modal setting. 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 both the extractive and abstractive multi-modal meeting summarization problem by fusing verbal and non-verbal features to enrich the representation of the utterance [Li *et al.*, 2019].

Leaderboard: To unify this research direction, we systematically present a comprehensive leaderboard for two widely used meeting summarization datasets: the AMI and ICSI datasets, using *pyrouge* package[1], as shown in Table 2.

Highlight: Meetings always involve several participants with specific roles. Thus, it is necessary to model such distinctive role characteristics. Besides, 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, which has led to people being overwhelmed by massive amounts of chat information. Such complex dialogue context poses a challenge to the new chat participant, since he or she may be unable to quickly review the main idea of the dialogue. Therefore, summarizing chats becomes a new trending direction.

[1]https://pypi.org/project/pyrouge/

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. Afterward, Chen and Yang [2020] take the first step and propose a multi-view dialogue summarizer by introducing both topic segments and conversational stages. More importantly, they conduct a comprehensive study for the challenges in this task, revealing the importance of dialogue modeling for the dialogue summarization, which points out the direction for future researchers.

Roughly speaking, the majority of current works put much emphasis on two aspects: *dialogue interaction modeling* and *dialogue participant modeling*, which are in line with the prominent characteristics of conversational data.

To model the interaction, graph modeling strategies combined with additional features are widely adopted. Zhao *et al.* [2020] utilize fine-grained topic words as bridges between utterances to construct a topic-word guided dialogue graph. Chen and Yang [2021b] consider the inter-utterance dialogue discourse structure and intra-utterance action triples to explicitly model the interaction. Feng *et al.* [2021a] view commonsense knowledge as cognitive interactive signals behind different utterances and shows the effectiveness of the integration of knowledge and heterogeneity modeling for different types of data. Liu *et al.* [2021] explicitly incorporate coreference 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.

To model the participants, Lei *et al.* [2021] implicitly model complex relationships among participants and their relative personal pronouns via speaker-aware self-attention mechanism. From another perspective, Narayan *et al.* [2021] and Liu and Chen [2021] explicitly adopt the guided summarization framework and introduce the participant information

into the coarse-to-fine generation procedure, in which the final dialogue summary is controlled by a precedent, such as a sketch or named entities.

As shown in the above works, current dialogue summarization systems usually encode the text with additional information. However, these annotations are usually obtained via open-domain toolkits, which are not suitable for dialogues, or require manual annotations, which are labor-consuming. Therefore, Feng *et al.* [2021c] present an unsupervised DialoGPT annotator, which can perform three dialogue-specific annotation tasks, including keywords extraction, redundancy detection and topic segmentation.

Despite the encouraging results reported, current models still suffer from the data-insufficient problem. Accordingly, some researchers study this task in the low-resource regime. Gunasekara *et al.* [2021] innovatively explore the summary-to-dialogue generation problem and verify the augmented dialogue-summary pairs can do good to dialogue summarization. Chen and Yang [2021a] propose three conversational data augmentation methods to enrich the data, including random swapping or deletion utterances, dialogue-acts-guided utterance insertion and conditional-generation-based utterance substitution.

Leaderboard: Previous works have already achieved remarkable success on the SAMSum dataset [Gliwa *et al.*, 2019]. 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[2]. The results are shown in Table 3.

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			
Transformer [Vaswani *et al.*, 2017]	36.62	11.18	33.06
PGN [See *et al.*, 2017]	40.08	15.28	36.63
D-HGN [Feng *et al.*, 2021a]	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.*, 2020]	39.77	16.58	38.42
UniLM [Dong *et al.*, 2019]	47.85	24.23	46.67
BART [Lewis *et al.*, 2020]	52.98	27.67	49.06
S-BART [Chen and Yang, 2021b]	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]	54.05	28.56	50.57
BART($\mathcal{D}_{\text{ALL}}$) [Feng *et al.*, 2021c]	53.70	28.79	50.81
Coref-ATTN [Liu *et al.*, 2021]	53.93	28.58	50.39
Entity-Plan [Liu and Chen, 2021]†	56.53	32.40	54.92

Table 3: Leaderboard of chat summarization on the SAMSum dataset [Gliwa *et al.*, 2019], where "R" is short for "ROUGE". We mainly adopt results from corresponding publications. Besides, the results of S-BART, MV-BART, Coref-ATTN and Entity-Plan are obtained by evaluating output files provided by the author. † indicates the model obtains these results with the help of golden summaries.

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 selecting the important parts and tend to generate hallucinations. In the future, powerful chat modeling strategies and reasoning abilities should be explored for this task, and more low-resource settings should be considered.

3.3 Email Threads Summarization

Email thread is an asynchronous multi-party communication consisting of a coherent exchange of email messages among several participants, which is widely used in the 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.

Major efforts lie on email thread summarization. Pioneer works present publicly available datasets to facilitate this task. Carenini *et al.* [2007] collect 39 email threads from Enron email dataset and annotate them with extractive summaries. They propose an email fragment quotation graph based on the occurrence of clue words and conduct extractive summarization. Notably, quotation plays an important role in the email that can directly highlight the salient part of the previous email. To enrich the annotation, Ulrich *et al.* [2008] collect 40 email threads from W3C email dataset and annotate them with both abstractive and extractive summaries along with meta sentences, subjectivity and speech acts. Loza *et al.* [2014] collect 107 email threads from Enron email dataset and annotate them with extractive and abstractive summaries combined with key phrases. Recently, Zhang *et al.* [2021a] present EMAILSUM, which contains 2549 email threads collected from Avocado Research Email Collection associated with human-written short and long abstractive summaries. This large-scale and high-quality dataset provides opportunities to data-hungry neural models.

In light of emails always being used for workflow organization and task tracking, some works explore action-focused email summarization, aka TO-DO item generation, which can help users with task management over emails. Mukherjee *et al.* [2020] propose a Smart TO-DO system, which first detects commitment sentences and then generates to-do items using sequence-to-sequence 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 understanding the fine-grained action items in the email and the coarse-grained intent of the entire email. Besides, better use of quotations can yield significant benefits.

[2]https://pypi.org/project/py-rouge/

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 the 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 intent 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 dialogue, which contains informative entities, covers various domains and involves two distinct types of participants. To integrate dialogue-specific information, Zhao *et al.* [2021] craft a new dataset annotated with dialogue state knowledge, which is helpful for tracking the fine-grained dialogue information flow and generating faithful summaries. Since participants in customer service play distinct roles, in addition to the overall summary for the whole dialogue, Zhang *et al.* [2021b] propose an unsupervised framework based on variational auto-encoder to generate summaries for the customer and the agent respectively. [Lin *et al.*, 2021] directly propose CSDS datasets annotated with role-oriented summaries to acquire different speakers' viewpoints.

Highlight: Customer service aims to address the 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. Besides, some fine-grained information also should be taken into consideration to ensure faithfulness, such as slots, states and intents.

3.5 Medical Dialogue Summarization

Medical dialogues happen between patients and doctors. During this process, doctors are required to record a 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.* [2021] 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 dialogue and negations scattered 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 in medical dialogues. Joshi *et al.* [2020] leverage the 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, the medical dialogue summary mainly describes core items and concepts in the dialogue, therefore, the summarization methods should bias towards extractive methods while keeping the 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.

4 New Frontiers

Section 3 mainly introduces prominent achievements in different domains respectively. 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 still suffer from the factual inconsistency problem, which distorts or fabricates the factual information in the article and is also known as hallucinations. Tang *et al.* [2021] systematically study the taxonomy of factuality errors for dialogue summarization, which includes the following 8 error types: *Missing Information, Redundant Information, Circumstantial Error, Wrong Reference Error, Negation Error, Object Error, Tense Error* and *Modality Error*. Specifically, the last five types of errors no-

toriously tend to appear in dialogue summaries, which largely hinder the application of dialogue summarization systems.

To remedy these issues, future works need specific designs target for the above errors. Importantly, fine-grained dialogue-specific features need to be incorporated into the summarization model, such as personal pronoun information, coreference information and tense information. On the one hand, these features can implicitly alleviate the difficulty of dialogue understanding. On the other hand, some features can directly serve as the explicitly extracted information to help final summary generation.

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 voiceprint 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.

Asynchronous multi-modal dialogues refer to different modalities that 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 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 exploring the multi-domain learning or domain adaption for dialogue summarization [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, to further utilize these datasets.

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

4.4 Multi-lingual Dialogue Summarization

With the acceleration of globalization, a dialogue involving multinational participants becomes increasingly common thanks to the sophisticated instantaneous translation system. Therefore, there is an urgent need for providing people with dialogue summaries in a preferred language. However, current works overwhelmingly focused on English, while leaving other languages under exploration. We argue that the current dilemma is mainly caused by the intractable access to available multi-lingual data resources.

Firstly, future works should devote efforts to creating a suitable testbed for multi-lingual dialogue summarization. As an initial step, Mehnaz *et al.* [2021] transform English utterances in the SAMSum dataset [Gliwa *et al.*, 2019] into Hindi-English utterances and study the chat summarization under the code-switched setting. From a higher point of view, large-scale high-quality datasets covering diverse languages should be carefully crafted. Practically speaking, on the one hand, researchers can translate one specific dataset into different languages followed by automatic and human quality checking to get aligned datasets. On the other hand, researchers can also borrow ideas from unsupervised multi-lingual learning to utilize currently available datasets in different languages. Secondly, future works should set up systematic settings for this multi-lingual research, including *one-to-one*, *one-to-many*, *many-to-one* and *many-to-many*, in which the *one-to-one* setting can be further divided into *mono-lingual* setting and *cross-lingual* setting. Thirdly, plenty of multi-lingual pre-trained language models can be explored for this task. Especially, models that have already been fine-tuned on the translation datasets may bring significant benefits

5 Conclusion

This article presents the first comprehensive survey on the progress of dialogue summarization carefully and widely. We thoroughly summarize the existing works, which cover various domains and highlight their challenges respectively. Besides, we summarize currently available datasets and organize two 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

[Carenini *et al.*, 2007] Giuseppe Carenini, Raymond T. Ng, and Xiaodong Zhou. Summarizing email conversations with clue words. In *Proc. of WWW*, 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*, 2005.

[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*, 2020.

[Chen and Yang, 2021a] Jiaao Chen and Diyi Yang. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proc. of EMNLP*, 2021.

[Chen and Yang, 2021b] Jiaao Chen and Diyi Yang. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proc. of NAACL*, 2021.

[Chen *et al.*, 2021] Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. DialogSum: A real-life scenario dialogue summarization dataset. In *Proc. of Findings of ACL-IJCNLP*, 2021.

[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 *Proc. of NeuIPS*, 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 *Proc. of the First Workshop on Natural Language Processing for Medical Conversations*, 2020.

[Fabbri *et al.*, 2021] Alexander 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. In *Proc. of the ACL-IJCNLP*, 2021.

[Feigenblat *et al.*, 2021] Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. TWEETSUMM a dialog summarization dataset for customer service. In *Proc. of Findings of EMNLP*, 2021.

[Feng *et al.*, 2021a] Xiachong Feng, Xiaocheng Feng, and Bing Qin. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. In *Proc. of CCL*, 2021.

[Feng *et al.*, 2021b] Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *Proc.of IJCAI*, 2021.

[Feng *et al.*, 2021c] Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. Language model as an annotator: Exploring DialoGPT for dialogue summarization. In *Proc. of ACL-IJCNLP*, 2021.

[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*, 2019.

[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)*, 2018.

[Gunasekara *et al.*, 2021] Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Sachindra Joshi, and David Konopnicki. Summary grounded conversation generation. In *Proc. of Findings of ACL-IJCNLP*, 2021.

[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*, 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 *Proc. of Findings of EMNLP*, 2020.

[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*, 2019.

[Khalman *et al.*, 2021] Misha Khalman, Yao Zhao, and Mohammad Saleh. ForumSum: A multi-speaker conversation summarization dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

[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 *Proc. of COLING*, 2020.

[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 *Proc. of NAACL: Student Research Workshop*, 2021.

[Krishna *et al.*, 2021] Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proc. of ACL-IJCNLP*, 2021.

[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 *Proc. of 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*, 2020.

[Li *et al.*, 2019] Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proc. of ACL*, 2019.

[Lin *et al.*, 2021] Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. CSDS: A fine-grained Chinese dataset for customer service dialogue summarization. In *Proc. of EMNLP*, 2021.

[Lin, 2004] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

[Liu and Chen, 2021] Zhengyuan Liu and Nancy Chen. Controllable neural dialogue summarization with personal named entity planning. In *Proc. of EMNLP*, 2021.

[Liu *et al.*, 2019a] Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. Automatic dialogue summary generation for customer service. In *Proc. of KDD*, 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)*, 2019.

[Liu *et al.*, 2021] Zhengyuan Liu, Ke Shi, and Nancy F. Chen. Coreference-aware dialogue summarization. In *SIGDIAL*, 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 *Proc. of LREC*, 2014.

[Malykh *et al.*, 2020] Valentin Malykh, Konstantin Chernis, Ekaterina Artemova, and Irina Piontkovskaya. SumTitles: a summarization dataset with low extractiveness. In *Proc. of COLING*, 2020.

[Mehnaz *et al.*, 2021] Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle G. Lee, Anish Acharya, and Rajiv Ratn Shah. GupShup: Summarizing open-domain code-switched conversations. In *Proc. of the EMNLP*, 2021.

[Mihalcea and Tarau, 2004] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proc. of EMNLP*, 2004.

[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*, 2020.

[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 *Proc. of AAAI*, 2017.

[Narayan *et al.*, 2021] Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. Planning with learned entity prompts for abstractive summarization. *Transactions of ACL*, 2021.

[Paice, 1990] Chris D Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 1990.

[Rameshkumar and Bailey, 2020] Revanth Rameshkumar and Peter Bailey. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proc. of ACL*, 2020.

[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*, 2017.

[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*, 2018.

[Song *et al.*, 2020] Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. Summarizing medical conversations via identifying important utterances. In *Proc. of COLING*, 2020.

[Tang *et al.*, 2021] Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. *ArXiv*, 2021.

[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 workshop*, 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 *Proc. of NeuIPS*, 2017.

[Wu *et al.*, 2021] Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. Controllable abstractive dialogue summarization with sketch supervision. In *Proc. of Findings of ACL-IJCNLP*, 2021.

[Yu *et al.*, 2021] Tiezheng Yu, Zihan Liu, and Pascale Fung. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In *Proc. of ACL*, 2021.

[Zhang *et al.*, 2020] 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*, 2020.

[Zhang *et al.*, 2021a] Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. EmailSum: Abstractive email thread summarization. In *Proc. of ACL-IJCNLP*, 2021.

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

[Zhang *et al.*, 2021c] Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir Radev. An exploratory study on long dialogue summarization: What works and what's next. In *Proc. of Findings of EMNLP*, 2021.

[Zhao *et al.*, 2020] Lulu Zhao, Weiran Xu, and Jun Guo. Improving abstractive dialogue summarization with graph structures and topic words. In *Proc. of COLING*, 2020.

[Zhao *et al.*, 2021] Lulu Zhao, Fujia Zheng, Keqing He, Weihao Zeng, Yuejie Lei, Huixing Jiang, Wei Wu, Weiran Xu, Jun Guo, and Fanyu Meng. Todsum: Task-oriented dialogue summarization with state tracking. *arXiv*, 2021.

[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. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proc. of ACL*, 2021.

[Zhong *et al.*, 2022] Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *Proc. of AAAI*, 2022.

[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 *Proc. of Findings of EMNLP*, 2020.

[Zhu *et al.*, 2021] Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proc. of NAACL*, 2021.

[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 auto-encoders. In *Proc. of 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 *Proc. of AAAI*, 2021.