Taxonomy of Abstractive Dialogue Summarization: Scenarios, Approaches and Future Directions

QI JIA, Shanghai Jiao Tong University, China
YIZHU LIU, Meituan, China
SIYU REN, Shanghai Jiao Tong University, China
KENNY Q. ZHU*, University of Texas at Arlington, United States

Abstractive dialogue summarization generates a concise and fluent summary covering the salient information in a dialogue among two or more interlocutors. It has attracted significant attention in recent years based on the massive emergence of social communication platforms and an urgent requirement for efficient dialogue information understanding and digestion. Different from news or articles in traditional document summarization, dialogues bring unique characteristics and additional challenges, including different language styles and formats, scattered information, flexible discourse structures, and unclear topic boundaries. This survey provides a comprehensive investigation of existing work for abstractive dialogue summarization from scenarios, approaches to evaluations. It categorizes the task into two broad categories according to the type of input dialogues, i.e., open-domain and task-oriented, and presents a taxonomy of existing techniques in three directions, namely, injecting dialogue features, designing auxiliary training tasks and using additional data. A list of datasets under different scenarios and widely-accepted evaluation metrics are summarized for completeness. After that, the trends of scenarios and techniques are summarized, together with deep insights into correlations between extensively exploited features and different scenarios. Based on these analyses, we recommend future directions, including more controlled and complicated scenarios, technical innovations and comparisons, publicly available datasets in special domains, etc.

CCS Concepts: • **Computing methodologies → Natural language generation**; **Discourse, dialogue and pragmatics**; • **General and reference** → *Surveys and overviews.*

Additional Key Words and Phrases: dialogue summarization, dialogue context modeling, abstractive summarization

ACM Reference Format:
Qi Jia, Yizhu Liu, Siyu Ren, and Kenny Q. Zhu. 2023. Taxonomy of Abstractive Dialogue Summarization: Scenarios, Approaches and Future Directions. *ACM Comput. Surv.* 37, 4, Article 111 (August 2023), 37 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Abstractive text summarization aims at generating a concise summary output covering key points given the source input. Prior studies mainly focus on narrative text inputs such as news stories , including CNN/DM [58] and XSum [122], and other publications, including PubMed and arXiv [34],

*Corresponding author.

Authors' addresses: Qi Jia, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, China, 200240, Jia_qi@sjtu.edu.cn; Yizhu Liu, Meituan, Shanghai, China, 200093, liuyizhu@meituan.com; Siyu Ren, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, China, 200240, roy0702@sjtu.edu.cn; Kenny Q. Zhu, University of Texas at Arlington, 500 UTA Blvd, Arlington, TX, United States, 76010, kenny.zhu@uta.edu.

0360-0300/2023/8-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

and have achieved remarkable success. As a natural way of communication, dialogues have attracted increasing attention in recent years. With the rapid growth of real-time messaging, consultation forums, and online meetings, information explosion in the form of dialogues calls for more efficient ways of searching and digesting dialogues.

Dialogue summarization targets summarizing salient information in a third party's view given utterances among two or more interlocutors. This task is not only helpful in providing a quick context to new participants of a conversation, but can also help people grasp the central ideas or search for key contents in the conversation, which promotes efficiency and productivity. It is first proposed as meeting summarization by Carletta et al. [17] and Janin et al. [60] and generally covers a number of scenarios, such as daily chat [54], medical consultation [63], customer service [201], etc. Different from document summarization where inputs are narrative texts from a third party, inputs for dialogue summarization are from multiple speakers in the first person. Dialogues are not only abundant with informal expressions and elliptical utterances [97, 186], but also full of question answerings and repeated confirmations to reach a consensus among speakers. The inherent semantic flows are complicatedly reflected by vague topic boundaries [151] and interleaved inter-utterance dependencies [1]. In a word, the information in dialogues is sparse and less structured, and the utterances are highly content-dependent, raising the difficulty for dialogue summarization.

Based on these characteristics, abstractive dialogue summarization generating fluent summaries is preferred by humans instead of the extractive one that extracts utterances. The earliest efforts approached this by transforming dialogues into word graphs and selecting the suitable paths in the graph as summary sentences by complicated rules [10, 139]. Template-based approaches [126, 143] were also adopted, which collect templates from human-written summaries and generate abstractive summaries by selecting suitable words from the dialogue to fill in the blank. However, their generated summaries lack fluency and diversity thus are far from practical use. Later, neural encoder-decoder models showed up. They projected the input into dense semantic representations and summaries with novel words were generated by sampling from the vocabulary list step-by-step until a special token representing the end was emitted. Abstractive text summarization has achieved remarkable progress based on these models tracing back from non-pretrained ones such as PGN [137], Fast-Abs [27] and HRED [138], to pretrained ones including BART [80] and Pegasus [181]. At the same time, techniques for dialogue context modeling have also evolved significantly with neural models in dialogue-related researches, such as dialogue reading comprehension [149], response selection [172] and dialogue information extraction [177]. The rapid growth of the two areas above paves the way for a recent revival of research in abstractive dialogue summarization.

Dozens of papers have been published in the area of dialogue summarization in recent years. Notably, a number of technical papers have dug into various dialogue features and datasets under different scenarios. It is time to look at what has been achieved, find potential omissions and provide a basis for future work. However, there is no comprehensive review of this field, except for Feng et al.'s recent survey [44]. Different from their paper which focuses on datasets and benchmarks targetting only a few applications, our survey aims at providing a thorough account of abstractive dialogue summarization, containing taxonomies of task formulations with different scenarios, various techniques, and evaluations covering different metrics and 35 datasets. This survey not only serves as a review of existing work and points out future directions for research but also can be a useful look-up manual for engineers when solving problems. We also hope this survey could serve as a milestone for dialogue summarization approaches mainly before the emergence of large language models (LLM), such as LLaMa [157] and ChatGPT [124], and bring inspirations for developing new techniques with LLMs.

The remainder of this review is structured as follows. Sec. 2 is the problem formulation, providing a formal task definition, unique characteristics compared to document summarization and a hierarchical classification of existing application scenarios. Sec. 3 to Sec. 6 presents a comprehensive taxonomy of dialogue summarization approaches in which current dialogue summarization techniques are mainly based on tested document summarization models and can be divided into three directions, including (1) injecting pre-processed features (Sec. 4), (2) designing self-supervised tasks (Sec. 5), and (3) using additional data (Sec. 6). A collection of proposed datasets and evaluation metrics are in Sec. 7. Based on the highly related papers, we offer deep insights on correlations between techniques and scenarios in Sec. 8.1. We further suggest several future directions, including more controlled and complicated scenarios, technical innovations and feature comparisons, open-source datasets in special domains, and benchmarks and methods for evaluation in Sec. 8.2.

2 PROBLEM FORMULATION

We formally define the abstractive dialogue summarization task with mathematical notations, and highlight the characteristics of this task by contrasting it with the well-studied document summarization problem. Finally, we present a hierarchical classification of application scenarios, demonstrating the practicality of this task.

2.1 Task Definition

A dialogue can be formalized as a sequence of T chronologically ordered turns:

$$D = \{U_1, U_2, ..., U_T\} \tag{1}$$

Each turn U_t generally consists of a speaker/role s_t and corresponding utterance $u_t = \{w_i^t|_{i=1}^{l_t}\}$. w_i^t represents the i-th token[1] in the t-th utterance, l_t is the length of u_t.

Dialogue summarization aims at generating a short but informative summary $Y = \{y_1, y_2, ..., y_n\}$ for D, where n is the number of summary tokens. Y and $\hat{Y}$ represent the reference summary and the generated summary respectively.

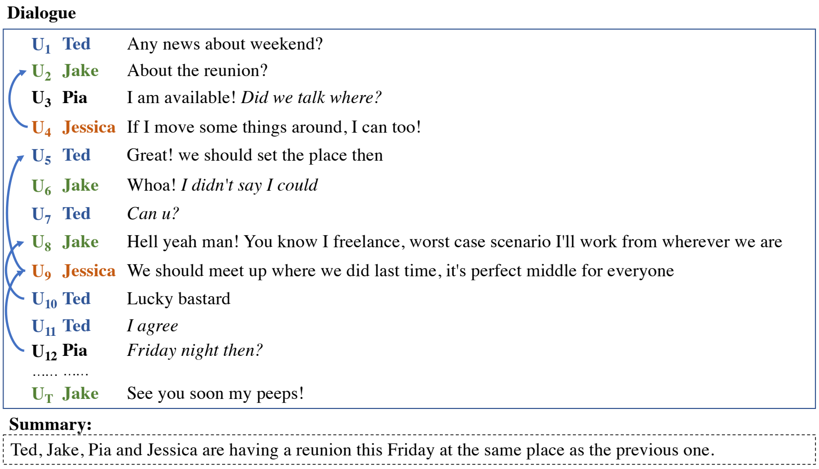
2.2 Comparisons to Document Summarization

Dialogue summarization is different from document summarization in various aspects, including language style and format, information density, discourse structure, and topic boundaries.

Word Level - Language Style and Format: Documents in previous well-researched summarization tasks are written from the third point of view, while dialogues consist of utterances expressed by different speakers in the first person. Informal and colloquial expressions are common especially for recorded dialogues from speech, such as "Whoa" in U_6 and "u" representing "you" in U_7 from Fig. 1. Moreover, pronouns are frequently used to refer to events or persons mentioned in the dialogue history. Around 72% of mentions in the conversation are anaphoras as stated in Bai et al. [9]. Meanwhile, the performance of coreference resolution models trained on normal text drops dramatically on dialogues [106]. All of these points manifest the existence of language style differences between documents and dialogues, posing a barrier in understanding the mappings between speakers and events in dialogues.

Sentence/Utterance Level - Information Density: Document sentences are more self-contained with complete SVO (subject-verb-object) structures, while elliptical utterances are ubiquitous in dialogues, including U_3, U_6, U_7, U_{11} and U_{12}. Besides, the long dialogue can be summarized into a single summary sentence in Fig. 1 as a result of back-and-forth questions and confirmations among speakers for communication purposes. Question answerings, acknowledgments, and comments [6] are frequent discourse relations among utterances to narrow down speakers' information gaps

[1] Texts are tokenized into tokens in the vocabulary as the input for neural models. Rare words may result in multiple tokens by algorithms such as Byte-Pair-Encoding. We do not strictly distinguish words and tokens in this survey.



Dialogue

U$_1$ **Ted** Any news about weekend?
U$_2$ **Jake** About the reunion?
U$_3$ **Pia** I am available! *Did we talk where?*
U$_4$ **Jessica** If I move some things around, I can too!
U$_5$ **Ted** Great! we should set the place then
U$_6$ **Jake** Whoa! *I didn't say I could*
U$_7$ **Ted** *Can u?*
U$_8$ **Jake** Hell yeah man! You know I freelance, worst case scenario I'll work from wherever we are
U$_9$ **Jessica** We should meet up where we did last time, it's perfect middle for everyone
U$_{10}$ **Ted** Lucky bastard
U$_{11}$ **Ted** *I agree*
U$_{12}$ **Pia** *Friday night then?*
....
U$_T$ **Jake** See you soon my peeps!

Summary:

Ted, Jake, Pia and Jessica are having a reunion this Friday at the same place as the previous one.

Fig. 1. An example multi-party dialogue and its summary. The arrows represent unsequential dependencies between utterances. Elliptical sentences are in italic.

and reach agreements. In this way, dialogue utterances are highly content-dependent, and the information is scattered [190], raising the difficulties for generating integral contents.

Inter-sentence/utterance Level - Discourse structure: Articles tend to be well-structured, such as general-to-specific structure or deductive order. For example, the most important information in news summarization are always at the beginning of the document, resulting in a competitive performance of the simple Lead-3 baseline [137]. However, it is not the same for dialogue summarization. Both Lead-3 and Longest-3, i.e. $\{U1, U2, U3\}$ and $\{U4, U8, U9\}$ in Fig. 1, get poor results in different dialogue scenarios [28, 54, 183]. The dependencies among utterances are interleaved, shown by arrows in Fig. 1, and discourse relations in dialogues are more flexible, even with the correction of wrong information. For example, Jake refused to be available for the reunion in U_6, but later agreed in U_8. As a result, it is more challenging to reason cross utterances for dialogue summarization than document summarization.

Passage/Session Level - Topic boundaries: Sentences under the same topic in documents are collected together in a paragraph or a section. Previous works for extractive [170] and abstractive summarization [34] both took advantage of such features and made great progress. However, a dialogue is a stream of continuous utterances without boundaries, even for hours of discussion. The same topic may be discussed repeatedly with redundancies and new information, setting up obstacles for content selection in dialogue summarization.

To better explain why abstractive approaches are more preferred than extractive ones, we list the result of the best rule-based extractive baseline, i.e., Longest-3 [54], the oracle extractive result determined by Rouge-L Recall score between each summary sentence and dialogue utterances [27], and the generation by BART fine-tuned on SAMSum dataset [54] of the dialogue in Fig. 1 as follows:

Longest-3	Jessica: If I move some things around, I can too! Jake: Hell yeah man! You know I freelance, worst case scenario I'll work from wherever we are Jessica: We should meet up where we did last time, it's perfect middle for everyone.
Oracle	Jake: Hell yeah man! You know I freelance, worst case scenario I'll work from wherever we are
BART	Ted, Pia, Jessica and Jake are going to meet up on Friday night.

We can see that the readability of generated summaries are poor for Longest-3 and Oracle due to the language style and format difference. The compression ratio of Longest-3 is apparently low while it still misses the involvement of Ted and Pia as a result of low information density of dialogues. Oracle is concise but much more information is missing. The fine-tuned BART as an

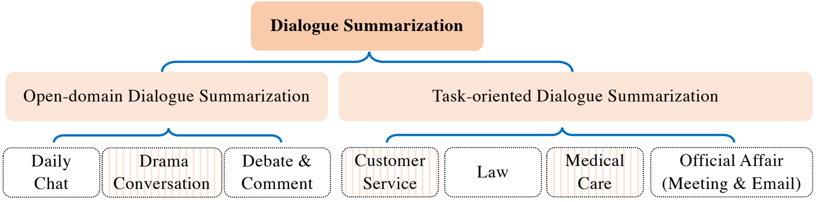

Dialogue Summarization
Open-domain Dialogue Summarization
Task-oriented Dialogue Summarization
Daily Chat
Drama Conversation
Debate & Comment
Customer Service
Law
Medical Care
Official Affair (Meeting & Email)

Fig. 2. The classification of dialogue summarization tasks with different application scenarios. Datasets proposed for evaluations under each scenario are in Sec. 7.1.

abstractive approach shows the favorable performance. In a word, dialogue summarization is a valuable research direction, where the modeling and understanding of dialogues are challenging compared with document summarization and abstractive approaches are especially preferred.

2.3 Scenarios for Dialogue Summarization

Considering the source of dialogues and the purpose of doing summarization, we divide the application scenarios into two classes: **open-domain dialogue summarization (ODS)** and **task-oriented dialogue summarization (TDS)**. This taxonomy is similar to the one of dialogue systems [20]. However, one should note that a pre-defined domain ontology is not necessarily required for TDS, which is different from that in task-oriented dialogue systems. The application scenarios investigated in previous papers are classified into these two classes as shown in Fig. 2.

Open-domain dialogue summarization is further divided into daily chat, drama conversation, and debate & comment. **Daily chat** [28, 54] refers to the dialogues happening in our daily lives, such as making appointments, discussions between friends, etc. **Drama conversation** [24, 109, 132, 197] represents dialogues from soap operas, movies or TV shows, which are dramatized or fabricated with drama scripts behind them. Dialogues in these two classes are full of person names and events, resulting in narrative summaries about "who did what". **Debate & comment** [32, 39, 114] focuses more on question answering and discussions in online forums and arguments. These dialogues emphasize opinions or solutions to the given subject or questions.

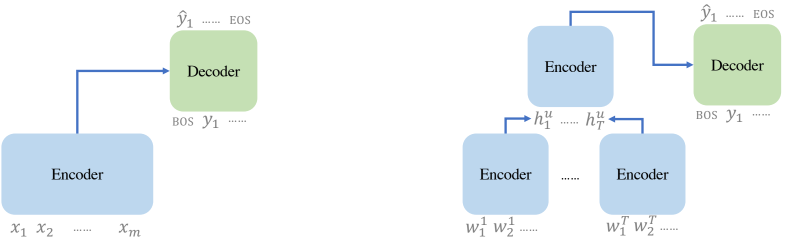
Task-oriented dialogue summarization arises from application scenarios of different domains, which includes but is not limited to customer service, law, medical care and official issue. **Customer service** [25, 42, 94, 193, 201] refers to conversations between customers and service providers. Customers start the conversation with their specific intents and agents are required to meet these requirements with the help of their in-domain databases, such as hotel reservations and express information consultation for online shopping. Dialogue summarization for this task is mainly to help service providers quickly go through solutions to users' questions for agent training and service evaluation. **Law** [38, 48, 169] is dialogues related to legal service and criminal investigations. Dialogue summarization in this scenario alleviates the recording and summarizing workload for law enforcement or legal professionals. **Medical care** [63, 105, 145, 145, 182] is dialogues between doctors and patients and medical dialogue summarization has some similarity to the research on electronic health records (EHR). Unlike the previous work focusing on mining useful information from EHR [173], summarization is to extract useful information from the doctor-patient dialogue and generate an EHR-like or fluent summary for clinical decision-making or online search. It also aims to reduce the burden of domain experts. **Official affair** [17, 60, 158, 183] is conversations between colleagues for technical or teachers and students for academic issue discussion. They can be in the format of meetings or e-mails, with summaries covering problems, solutions, and plans.

We compare and contrast ODS and TDS as follows.

- Dialogues happen between **two or more speakers** both in ODS and TDS, whereas the **interpersonal relationship** and **functional relationship** among speakers are different. Generally, speakers in ODS are friends, neighbors, lovers, family members, and so on. They are equal either in the aspect of interpersonal relationships or functional relationships. For example, one can raise a question or answer others' questions in online forums [39]. In TDS, speakers have different official roles acting for corresponding responsibilities. For example, plaintiff, defendant, witness and judge in court debates [38], project manager, marketing expert, user interface designer and industrial designer in official meetings [17] are corresponding roles. Among different dialogues, roles are the same and can be played by different speakers and a speaker's role is always unchanged for a service platform. In a word, TDS pays more attention to functional roles while ODS focuses on speakers.
- Multiple **topics** may be covered in the same dialogue session. Topics in ODS are more diverse than in TDS. The summarization models are expected to deal with unlimited open-domain topics such as chitchat, sales, education, and climate at the same time [28]. However, topics in TDS are more concentrated and need more expertise for understanding. Dialogues in TDS either focus on a single domain with more fine-grained topics, such as medical dialogues of different specialties, or several pre-defined domains, such as restaurant, hotel, and transformation reservation. Domain knowledge is significant for summarization, and it is divergent across sub-domains. For instance, expertise and medical knowledge are required in doctor-patient dialogues for generating accurate medical concepts [63] while specific knowledge bases for internal medicine and primary care are not the same.
- The input dialogue for both ODS and TDS is made up of **a stream of utterance** as defined in Equation 1. However, the **structure** of these two types of dialogues are different. Open-domain dialogues often happen casually and freely while dialogues in TDS may have some inherent working procedures or writing formats. For example, the program manager in meetings usually masters the meeting progress [198] implicitly with words such as "okay, what about ...", and communications by e-mails consist of semi-structured format including subjects, receivers, senders, and contents [183].
- **Focuses of summaries** are distinct. Summaries for ODS in recent research are more like condensed narrative paraphrasing. An example is a synopsis from the Fandom wiki maintained by fans for the Critical Role transcripts [132], helping to quickly catch up with what is going on in the long and verbose dialogues. Differently, dialogues in TDS take place with strong intentions for solving problems. Summaries for such dialogues are expected to cover the user intents and corresponding solutions, such as medical summaries for clinical decision making [63] and customer service summaries for ticket booking [193].

3 OVERVIEW OF APPROACHES

In abstractive text summarization, early researchers tried non-neural abstractive summarization methods [11], which used statistical models to recognize important words and sentences and then concatenate them into a final summary with or without pre-defined templates. The most direct way is to select a set of keywords from input [123], such as log-likelihood ratio test [87], which identified the set of words that appear in the input more often than in a background corpus. Another way is to assign weights to all words in the input, such as TF-IDF weights [12]. Word weights have also been estimated by supervised approaches with typical features, including word probability and location of occurrence [144]. Some other traditional work directly focuses on predicting sentence importance, by either emphasizing select sentences that match the template of summaries or selecting the sentences in which keywords appeared near each other. Such sentences can better convey important information and be selected as a summary [18, 92]. Researchers also



(a) Sequential Modeling (b) Hierarchical Modeling

Fig. 3. Two mainstream modeling designs for encoder-decoder summarization models.

productively explored the relationship between word and sentence importance, and tried to estimate each in either supervised or unsupervised framework [95]. Since 2015, neural-based abstractive text summarization models [134, 137] began to be widely developed, such as recurrent neural network (RNN) [121], convolutional neural network (CNN) [53] and Transformer [159] models. Benefiting from the semantic representation learned from large training data, neural-based methods outperform non-neural ones, especially in the aspect of fluency and semantic coherence.

The mainstream approaches in recent years hinge on the neural-based encoder-decoder architecture. In document/news summarization, document sentences can be concatenated into a single sequence of tokens $X = \{x_1, x_2, ..., x_m\}$ as the input to the encoder $\mathrm{Enc}(\cdot)$ which maps the tokens into contextualized hidden states $H = \{h_1, h_2, ..., h_m\}$. m represents the number of input tokens. Besides such flat and sequential modeling, hierarchical modeling is another representative design as shown in Fig. 3, which is usually favored by longer dialogues. Sentences are no more concatenated but instead modeled with hierarchical encoders. The lower layer encoder projects tokens within a sentence into hidden states. Then, the higher layer encoder takes these hidden states as sentence embeddings and projects them into global hidden representations. The decoder $\mathrm{Dec}(\cdot)$ takes all of the hidden states H and previously generated tokens as input, predicting the next token step by step in an auto-regressive way. The training objective is to minimize the negative log-likelihood L with the teacher-forcing strategy as follows:

$$\begin{aligned} &H = \mathrm{Enc}(x_1, x_2, ..., x_m) \\ &P(y_p | y_{<p}, H) = \mathrm{Softmax}(W_v \mathrm{Dec}(BOS, y_1, y_2, ..., y_{p-1}, H)) \\ &L = -\frac{1}{n} \sum_{p=1}^{n} P(y_p | y_{<p}, H) \end{aligned} \tag{2}$$

where W_v is a trainable parameter matrix mapping hidden states into a vocabulary distribution. During inference, the predicted distribution over vocabulary at step p is:

$$P(\hat{y}_p | \hat{y}_{<p}, H) = \mathrm{Softmax}(W_v \mathrm{Dec}(BOS, \hat{y}_1, \hat{y}_2, ..., \hat{y}_{p-1}, H)) \tag{3}$$

Tokens are sampled based on this distribution with generation strategies such as greedy and beam searches to produce the optimal summary. Greedy search selects the next token with the largest probability at each step and subsumes it into the current generation, while beam search expands each candidate generation with top-k possible next tokens and preserves the k-best candidate generations at each step [134]. The candidate with the highest probability is the final output. The decoding process starts with the beginning of a sentence (BOS) token and terminates when the end of a sentence (EOS) token is generated. Nowadays, pre-trained models taking advantage of the

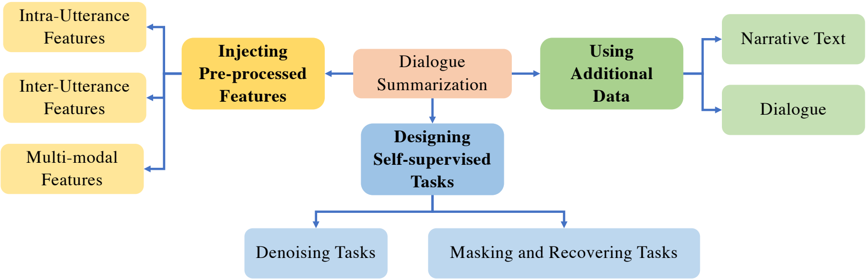


Fig. 4. The taxonomy of dialogue summarization techniques. Methods are mainly categorized into three directions with more fine-grained sub-categories under each direction.

Transformer encoder-decoder architecture with sequential modelings, such as BART and Pegasus, are the state-of-the-art abstractive text summarization techniques for document summarization.

These models also work for dialogue summarization. For sequential modeling, utterances prefixed with corresponding speakers are simply concatenated into the input sequence for a dialogue, i.e.

$$X = \{x_1, x_2, ..., x_m\} = [s_1, u_1, s_2, u_2, ..., s_T, u_T] \tag{4}$$

where $[\cdot]$ represents concatenation operation. However, such a simple operation largely ignores the flexible discourse structure and topic boundaries challenges in dialogue summarization. For hierarchical modeling, utterances $\{u^t|_{t=1}^T\}$ are passed into encoders separately, which sets a significant barrier for word-level cross-utterance understanding. Besides, models pretrained with normal text are not ideal for dialogue language understanding. To deal with these challenges, a number of techniques have emerged. This survey mainly focuses on newly introduced techniques for adapting tested abstractive document summarization models to dialogues. More detailed explanations of neural-based text summarization models and other methods please refer to other surveys [141, 150].

At a high level, recent researches tackle dialogue summarization in three directions:

- **Injecting pre-processed features** which explicitly exploits additional features in dialogue context either by human annotators or external labeling tools as part of the input.
- **Designing self-supervised tasks** which trains the model with auxiliary objectives besides the vanilla generation objective or individually for unsupervised summarization.
- **Using additional data** which includes bringing training data from other related tasks or performing data augmentation based on existing training corpus.

A number of techniques have been proposed under each direction which can be either adopted individually or combined for the targeted applications. An overall taxonomy is illustrated in Fig. 4. The following three sections present more details, accompanied by highlights of pros and cons.

4 INJECTING PRE-PROCESSED FEATURES

To pursue better dialogue understanding and reasoning, different features either designed by experts back on linguistic knowledge or engineered with observations are proposed to simulate the human comprehension process. Recognizing these features is not only independent dialogue analysis tasks but also critical enablers for downstream applications. A subset of these features has been proved helpful for dialogue summarization by extracting from D explicitly and injecting it into the vanilla model. We group different features into two sub-categories by their scopes:

- **Intra-utterance features** are features within an utterance or for an individual utterance.
- **Inter-utterance features** are features connecting or distinguishing multiple utterances.

4.1 Intra-Utterance Features

We divide the intra-utterance features into three groups: word-level, phrase-level or utterance-level.

4.1.1 Word-level. Word-level intra-utterance features include TF-IDF weights, Part-of-speech (POS) tags, and named entity tags.

The **TF-IDF weight** is a well-known statistical feature for each word, signifying its importance in the whole corpus. Term-frequency (TF) is the number of word occurrences in a dialogue or an utterance divided by the number of words. Inverse-document-frequency (IDF) refers to the logarithm of the number of dialogues or utterances divided by the number of them containing the word. Each dialogue or utterance can be represented by a vocabulary-sized TF-IDF weight vector, where each element is the product of TF and IDF. In early work, Murray et al. [118] used such utterance vectors as features for classifiers to find important utterances. This feature is still prevalent in constructing better prompts for the summary generation with large language models [129].

POS tags and **named entity tags** are linguistic labels assigned for each word. POS tags represent grammatical properties, including nouns, verbs, adjectives, etc. Named entity tags belong to pre-defined categories such as person names, organizations, and locations. Both of them are easily labeled by well-known NLP packages such as NLTK and Spacy, and can be assigned to summaries in the training set [126, 143], to generate summary templates for abstractive text summarization without neural models. Zhu et al. [198] trained two embedding matrices for both tags and concatenated them with word embeddings as part of the embedding layer for the model, i.e. $x_i^t = [e_i^t; POS_i^t; ENT_i^t]$. e_i^t, POS_i^t, and ENT_i^t are the word embedding, POS embedding, and named entity embedding for x_i^t, respectively. These features, which were also adopted by Qi et al. [130] in the same way, work for hierarchical models trained from scratch on this task and help with language understanding and entity recognition. However, the probing tests indicated that pre-trained language models have already captured both features well implicitly [155], and the two are no longer needed.

4.1.2 Phrase-level. Intra-utterance features here have key phrases/words and negation scopes.

Key phrases/words emphasize salient n-grams in the original dialogue, which can help with the information scattering challenge and lead to more informative summaries. The definition of key phrases varies. Wu et al. [168] regarded the longest common sub-sequence (LCS) between each candidate phrase, extracted from D first using a trained constituency parser, and Y as key phrases. The LCSs are concatenated into a sketch, which is prefixed to Y as a weakly supervised signal for the summary generation. Similarly, Zou et al. [201] proposed that words that appear both in D and Y are salient or informative topic words, i.e., another kind of keywords. They used an extension of the Neural Topic Model (NTM) [113] to learn the word-saliency correspondences. Then, input utterances are converted to topic representations by the saliency-aware NTM and further incorporated into Transformer Decoder layers for a better extractor-abstractor two-stage summarizer. Differently, Feng et al. [47] regarded unpredictable words by DialoGPT as keywords since they assumed that highly informative words could not be predicted. They appended all extracted keywords at the end of X as inputs to the summarization model.

The **negation scope** is also a set of consecutive words reflecting denied contents in utterances. Chen and Yang [21] pointed out that negations are challenging for dialogues. With that in mind, Khalifa et al. [65] trained a Roberta model on CD-SCO dataset [116] for negation scope prediction, which labels the beginning and end positions of sentences' negation scopes in D with designated special tokens. Unfortunately, inputting such labeled D to the model hurt the performance according to their experiment results. Negations are of great importance in task-oriented scenarios for generating accurate facts, such as realizing the patient's confirmation or negation of a symptom in a medical care conversation. Joshi et al. [63] proposed using an additional binary vector to label each x_i based on a set of manually-curated negative unigrams, and to modify the cross-attention distribution. Besides, they extended the vocabulary with a special token '[NO]' and learned when to generate it by formulating the probability distribution over extended vocabulary, similarly to See et al. [137]. The results showed reductions in coherency despite capturing negations.

4.1.3 Utterance-level. Speakers or roles, redundancies, user intents, and dialogue acts are utterance-level intra-utterance features. Domain knowledge is another kind of intra-utterance feature. It lies across phrase-level to utterance-level depending on specific circumstances.

Speaker or **role** is a naturally provided "label" for each dialogue utterance. Since the general default input to models is the concatenation of all of the utterances into a sequence of tokens, each speaker or role token s_t is encoded just like any other content token w_i^t [21, 47]. Thus, the speaker or role information is likely ignored or misunderstood, especially by language models pre-trained on common crawled texts. For a better utilization of speaker information, Lei et al. [77] introduced Speaker-Aware Self-Attention made up of Self-Self Attention and Self-Others Attention, which only considered whether utterances were from the same speaker. This structured feature is also adopted in [78]. In addition, the number of speakers is used as a feature for finding similar dialogues in the training set by Prodan and Pelican [129]. In TDS, the number of roles is always fixed in a specific scenario, although the speakers are various among dialogue sessions. [176] modified the input with template "{*speaker*} of role {*role*} said: {*utterance*}". Other previous work only focuses on modeling roles, reflecting functional information bias in utterances. The cheapest way is to represent each role with a dense vector r_t which is either obtained by randomly initialized trainable vectors [8, 38, 50, 130, 198] or a small trainable neural network [145]. This vector is further concatenated, summed up, or fused by non-linear layers with input embeddings e_i^t or utterance-level representations h_t^u in summarization models. There are also works that capture such features by different sets of model parameters for different roles [178, 187, 201]. More complicated methods that regard speakers or roles as graph nodes beyond the utterance-level are in Sec. 4.2.2.

Since dialogue utterances are mixed with backchanneling or repetitive confirmations [135], **redundancy** is also a significant feature where each utterance is either preserved or removed. Murray et al. [118] and Zechner [180] regarded utterances similar to the previous ones as redundant by calculating the cosine similarity between two sentence vectors computed using TF-IDF features. Then, the remaining utterances can be regarded as a summary. Different from previous work calculating similarities between individual utterances, Feng et al. [47] brought the context into consideration which calculated similarities on the dialogue level. Utterance representations h_t^u are collected by inputting the whole dialogue into DialoGPT [191]. Then, they assume that if adding an utterance u_{t+1} to the previous history $\{u_1, ..., u_t\}$ doesn't result in a big difference between the context representation h_t^u and h_{t+1}^u, u_{t+1} will be regarded as a redundant utterance. Such features will be added as part of the dialogue input with special tokens. Wu et al. [168] regarded non-factual utterances such as chit-chats and greetings as redundancies, and removed them by a sentence compression method with neural content selection for their summary sketch construction.

Another group of utterance-level features is matching each utterance with a label from a pre-defined multi-label set. Wu et al. [168] defined a list of interrogative pronoun category to encode the **user intent**, including *WHY*, *WHAT*, *WHERE*, *WHEN*, *CONFIRM* and *ABSTAIN*. Each utterance is labeled by a few heuristics and these user intents are combined with the keywords and redundancies mentioned above as a sketch prefixed to the summary output. This definition is different from the so-called user intent in task-oriented dialogue systems, while the latter can be used for TDS and will be discussed in domain ontologies in Sec. 4.2.2.

A more widely-accepted label set is **dialogue act**, which is defined as the functional unit used by speakers to change the context [16] and has been used for different goals [73, 125]. The whole dialogue act taxonomy, including dialogue assess, inform, offer, etc., is tailored for different scenarios. For example, only 15 kinds of dialogue are labeled in the meeting summarization corpus AMI [17] while the total number of categories is 42 [148]. Goo and Chen [55] explicitly modeled the relationships between dialogue acts and the summary by training the dialogue act labeling task and abstractive summarization task jointly. Di et al. [37] further added the dialogue act information as a contextualized weight to h_t^u. These labels are required from human annotators.

Commonsense knowledge generated by widely-used generative commonsense model PARA-COMET [49] is considered in [67]. PARA-COMET takes dialogue history with the target utterance or a summary sentence as input and outputs short phrases for each of the 5 relation types, which are strongly correlated with speakers' intentions and the hidden knowledge, such as "XINTENT" and "XREACT". The generated knowledge is concatenated with each utterance as input and is used as an additional generation target in a dual-decoder setting.

In addition, **domain knowledge** plays an important role in TDS. Koay et al. [69] showed that the existence of terms affects summarization performance substantially. Such knowledge is considered as intra-utterance features in previous work. Joshi et al. [63] leveraged a compendium of medical concepts for medical conversation summarization. They incorporated domain knowledge at the phrase level by simply encoding the presence of medical concepts, which are both in the source and the reference. The corresponding one-hot vectors affect the attention distribution by the weighted sum with contextualized hidden states H for each word only during training, like the teacher forcing strategy. Gan et al. [50] defined a number of domain aspects, and labeled text spans manually in D and S. Auxiliary classification tasks of these aspects help generate more readable summaries covering important in-domain contents. Differently, Duan et al. [38] incorporated their legal knowledge for each utterance. This is because their legal knowledge graph (LKG) depicts the legal judge requirements for different cases rather than a dictionary to look up, and each node represents a judicial factor requiring semantic analysis beyond the word level. A series of graph knowledge mining approaches were adopted to seek relevant knowledge w.r.t. each utterance u_t, and the legal knowledge embedding was added to the sentence embedding h_t^u for further encoding.

4.2 Inter-Utterance Features

As dialogue utterances are highly dependent, information transitions among utterances are of great importance for dialogue context understanding. Multiple inter-utterance features show up for more efficient and effective dialogue summarization, which can be categorized into two sub-categories:

- **Partitions** refer to extracting or segmenting the whole dialogue into relatively independent partitions. Information within each partition is more concentrated with fewer distractions for the summary generation. Meanwhile, these features reduce the requirements on GPU memory with shorter input lengths, which are especially preferred for long dialogue summarization.
- **Graphs** refer to extracting key information and relations from utterances to construct graphs, serving as a complement to the dialogue. These features are designed to help the summarization model understand the inherent dialogue structure.

4.2.1 Partitions. There are two types of partitions. One is to cut the dialogue into a sequence of K consecutive segments $\{S_k|_{k=1}^K\}$ with or without overlaps, i.e., $|D| \leq |S_1| + ... + |S_K|$, where $|\cdot|$ counts the number of utterances. Representative features under this category are as follows.

Topic transition is important for dialogues where speakers turn to focus on different topics. Consecutive utterances that focus on the same topic constitute a topic segment, which should meet three criteria[5], including being reproducible, not relying heavily on task-related knowledge, and being grounded in discourse structure. Some previous works annotate this feature when constructing datasets such as Carletta et al. [17] and Janin et al. [60]. Di et al. [37] took advantage of such labeled information during decoding. Others collected such features by rules or algorithms. Asi et al. [8] adopted the text segmentation idea from Alemi and Ginsparg [2] and broke the long dialogue into semantically coherent segments by word embeddings. Liu et al. [105] regarded different symptoms as different topics in medical dialogues and detected the boundaries by heuristics. To alleviate human annotation burdens, unsupervised topic segmentation methods are adopted. Chen and Yang [21] used the classic segmentation algorithm C99 [31] based on inter-utterance similarities, where utterance representations were encoded by Sentence-BERT [133]. Feng et al. [47] regarded sentences that are difficult to be generated based on the dialogue context to be the starting point of a new topic. Thus, sentences with the highest losses calculated based on DialoGPT are marked. However, the window size and std coefficient for C99 algorithm in Chen and Yang [21] and percentage of unpredictable utterances in Feng et al. [47] are still hyper-parameters that need assigning by humans. Among these works, some models use topic transitions as prior knowledge and input to summarisation models. They either add special tokens to dialogue inputs [21, 47], add interval segment embeddings, such as $\{t_a, t_a, t_b, t_b, t_b, t_a, ...\}$ for each utterance [130], or guide the model on learning segment-level topic representations h_k^s based on utterance representations h_t^u [194]. Others adjust their RNN-based models to predict topic segmentation first and do summarization based on the predicted segments [83, 105], either with or without using additional supervised topic labels for computing the segmentation training loss.

Multi-view [21] describes **conversation stages** [3] from a conversation progression perspective. They assumed that each dialogue contained 4 hidden stages, which were interpreted as "openings→intentions→discussions→conclusions", and annotated with an HMM conversation model. In their approach, both the preceding topic view and such stage view are labeled on dialogues with a separating token "|", encoded with two encoders sharing parameters and guided the Transformer decoder in BART with additional multi-view attention layers.

There also exists a simple **sliding-window** based approach that regards window-sized consecutive utterances as a snippet and collects snippets with different stride sizes. On the one hand, it can be used to deal with long dialogues. Sub-summaries are generated for each snippet and merged to get the final summary. Most works regarded the window size and the stride size as two constants [70, 96, 182, 189], while Liu and Chen [103] adopted a dynamic stride size which predicts the stride size by generating the last covered utterance at the end of Y'. Koay et al. [70] generated abstractive summaries for each snippet by news summarization models as a coarse stage for finding the salient information. On the other hand, pairs of (snippet, sub-summary) are augmented data for training better summarization models. By calculating Rouge scores between reference sentences and snippets, the top-scored snippet is paired with the corresponding sentence [96, 189]. Alternatively, multiple top-scored snippets can be merged as the corresponding input to the sentence [182] for the sub-summary generation. However, the gap between training and testing is that we don't know the oracle snippets since there is no reference summary during testing. Therefore, each snippet was also considered to be paired with the whole summary [182, 189], but it leads to hallucination problems. These constructed pairs can also be used with an auxiliary training objective [96], or as pseudo datasets for hierarchical summarization [2].

The other is to **cluster utterances** or **extract utterances** into a single part or multiple parts $\{P_l|_{l=1}^{K'}\}$. In this way, outlier utterances or unextracted utterances will be discarded, i.e., $|D| > |P_1| + ... + |P'_K|$. Then, the abstractive summarization model is trained between the partitions and the reference summary. The whole process can be regarded as variants under the extractor-abstractor framework for document summarization [27, 99].

[2] Hierarchical summarization means we do summarization, again and again, using the previously generated summaries as input to get more concise output. These models either share parameters [81] or not [182, 189] in each summarization loop.

Zou et al. [200] proposed to select topic utterances according to centrality and diversity[3] in an unsupervised manner. Each utterance with its surrounding utterances in a window size forms a topic segment. Zhong et al. [196] extracted relevant spans given the query with the Locator model which is initialized by Pointer Network [161] or a hierarchical ranking-based model. Cluster2Sent by Krishna et al. [71] extracted important utterances, clustered related utterances together and generated one summary sentence per cluster, resulting in semi-structured summaries suitable for clinical conversations. Banerjee et al. [10] and Shang et al. [139] followed a similar procedure, i.e., (segmentation, extraction, summarization) and (clustering, summarization) respectively. The oracle spans are required to be labeled for supervised training of extractors or classifiers for most approaches, except that Shang et al. [139] used K-means for utterance clustering in an unsupervised manner. Generally, the partitions are concatenated as the input to summarization models [196], or the generated summary of each segment is concatenated or ranked to form the final $\hat{Y}$ [10, 200].

4.2.2 Graphs.

The intuition for constructing graphs is attributed to the divergent structure between dialogues and documents introduced in Sec. 2.2. To capture the semantics among complicated and flexible utterances, a number of works constructed different types of graphs based on linguistic theories or observations and demonstrated improvements empirically. We group these graphs into three categories according to the type of nodes, i.e., being either a word, a phrase or an utterance.

Word-level graphs focus on finding the central words buried in the whole dialogue. Some works [10, 126, 127, 139] parsed utterances together with or without summary templates using the Standford or NLTK packages. Words in the same form and the same POS tag or synonyms according to WordNet [111] are regarded as a single node. The natural flow of text, parsed dependency relations and relations in WordNet are adopted to connect nodes, resulting in a directed **word graph**. It is used for unsupervised sentence compression by selecting paths covering nodes with high in-degree and out-degree without language models.

The purpose for *phrase-level graphs* is mainly to emphasize relations between important phrases. Liu et al. [106] and Liu and Chen [101] transferred document coreference resolution models [64, 75] to dialogues, applied data post-processing with human-designed rules and finally constructed **coreference graphs** for dialogues. The nodes are mainly person names and pronouns, and the edges connect nodes belonging to the same mention cluster. Based on the coreference results, Chen and Yang [23] took advantage of information extraction system [4] and constructed an **action graph** with "WHO-DOING-WHAT" triples. Zhao et al. [192] manually defined an undirected **semantic slot graph** based on NER and POS Tagging focusing on entities, verbs, and adjectives in texts, i.e., slot values. Edges in this graph represent the existence of dependency between slot values collected by a dependency parser tool. More strictly defined “domain-intent-slot-value” tuples based on structured **domain ontologies** are marked in advance [178, 193]. It is different from domain to domain, such as “food, area” slots for “restaurant” and “leaveAt, arriveBy” slot for “taxi” labeled in the MultiMOZ dataset [15]. Ontologies in the medical domains containing clinical guidelines in “subject-predicate-object” triples were introduced in Molenaar et al. [115]. Triples are extracted from D and matched with the ontology to construct a patient medical graph for report generation. Moreover, external commonsense **knowledge graphs**, such as ConceptNet [146], have been adopted to find the relations among speaker nodes, utterance nodes and knowledge nodes [43].

Utterance-level graphs considering the relationship among utterances have been explored mainly in five ways. One is **discourse graph** mainly based on the SDRT theory [7] which models the relationship between elementary discourse units (EDUs) with 16 types of relations for dialogues. Both Chen and Yang [23] and Feng et al. [46] adopted this theory and regarded each utterance as an

[3] Centrality reflects the center of utterance clusters in the representation space. Diversity emphasizes diverse topics among selected utterances.

EDU. They labeled the dialogue based on a discourse parsing model [142]. The former work used a directed discourse graph with utterances as nodes and discourse relations as edges. Differently, the latter one transformed the directed discourse graph with the Levi graph transformation where both EDUs and relations are nodes in the graph with two types of edges, including default and reverse. Self edges and global edges were also introduced to aggregate information in different levels of granularity. Ganesh and Dingliwal [51] designed a set of discourse labels themselves and trained a simple CRF-based model for discourse labeling. Unfortunately, they haven't released the details so far. **Dependency graph** can be regarded as a simplification of discourse graph since it only focuses on the "reply-to" relation among utterances. The tree structure of a conversation is a kind of it and is adopted in [176] by modifying the self-attention into thread-aware attention which considers the distance between two utterances, and also proposing a thread prediction task to predict the historical utterances in the same thread for some sampled utterances. Another one is **argument graph** [147] for identifying argumentative units, including claims and premises and constructing a structured representation. Fabbri et al. [39] did argument extraction with pretrained models [19] and connected all of the arguments into a tree structure for each conversation by relationship type classification [79]. Such a graph not only helps to reason between arguments but also eliminates unnecessary content in dialogues. Similarly, **entailment graph** [111] is used to identify important contents by entailment relations between utterances. The fifth is **topic graph**. Usually, we regard the topic structure in dialogues as a linear structure as discussed above, but it can be hierarchical with subtopics [17, 60] or non-linear structures since the same topic may be discussed back and forth [66]. Lei et al. [78] used ConceptNet to find the related words that indicate the connections among utterances under the same topic, capturing more flexible topic structures.

The graphs above are used in three ways. One is to convert the original dialogue into a narrative similar to documents by linearizing graphs and inputting to the basic summarization models [39, 51]. Second is to bring graph neural layers, such as Graph Attention Network [160] and Graph Convolutional Networks [68]. Such graph neural layer can be solely used as the encoder [43]. It can also cooperate with the Transformer-based encoder-decoder models, either based on the encoder hidden states or injected into the Transformer layer in encoder [106] or decoder [23]. The rest modify attention heads in Transformer with constructed graphs from a model pruning perspective. Liu et al. [106] replace attention heads containing the most coreference information with their coreference graph, while Liu and Chen [104] replace underused heads with a similar graph.

4.3 Multi-modal Features

Humans live and communicate in a multi-modal world. As a result, multi-modal dialogue summarization is naturally expected. Even for virtual dialogues from TV shows or movies, character actions and environments in videos are important sources for humans to generate meaningful summaries. However, due to the difficulties of collecting multi-modal data in real life and the limited multi-modal datasets, this area remains to be researched. **Prosodic features** gained attention in early speech-related works. Murray et al. [118] collected the mean and standard deviation of the fundamental frequency, energy and duration features based on speech. With the marvelous automatic speech recognition (ASR) models, most works later only focused on transcripts and ignored such multi-modal features. Besides, **visual focus of attention** (VFOA) feature from the meeting summarization scenarios has been introduced to highlight the importance of utterances [83]. It represents the interactions among speakers reflected by the focusing target that each participant looks at in every timestamp. They assumed that the longer a speaker was paid attention to by others, his or her utterance would be more important. Such orientation feature was converted into a vector by their VFOA detector framework and further concatenated to the utterance representations.

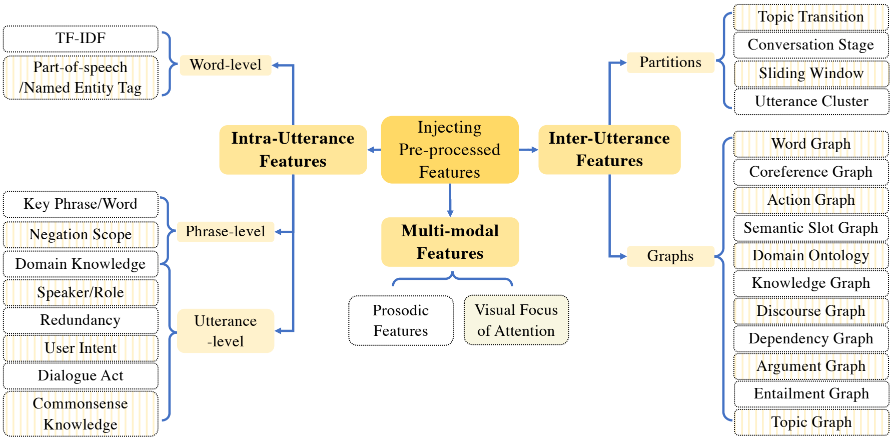


Fig. 5. A summary of all features.

4.4 Summary and Opinions

The above features are summarized in Fig. 5 and mainly injected into vanilla models in three ways:

- **Manipulating the input and output** by adding annotations or data reformulation. The former one adds additional tokens to the dialogue or summary to highlight the features, such as topic transition marks in the dialogue [21] and key phrase prefixes of the summary [168]. This is suitable for features linearly buried in the texts. For the hierarchical or more complicated structures, researchers tend to reformulate the dialogue into different segments, especially for long dialogues [10, 139, 196] or reordering utterances with graph features. For example, Fabbri et al. [39] linearized the argument graph following a depth-first approach to fine-tune sequence-to-sequence models for summarization, and Zhao et al. [193] linearized the dialogue states, i.e., slot-related labels, as a complement to D with a bi-encoder model.
- **Modifying the model architecture or hidden states** for learning inductive bias on known features. Embedding layers are always modified for word-level or phrase-level features indicating the binary or multi-class classification properties, including the POS embeddings in [198] and medical concept embedding in [63]. Modifications on self-attentions and cross-attentions are used to merge multiple features and are also preferable to graph features. For instance, Chen and Yang [21] modified the cross-attention layer for balancing and fusing hidden states of two kinds of labeled input from double encoders. Lei et al. [77] changed the self-attention layer in the encoder with two speaker-aware attentions to highlight the information flow within the same speaker or among speakers. Different graph neural layers [23, 43, 106] are also introduced for capturing graph features.
- **Adding additional training targets** means that features are regarded as a supervision output during training under multi-task learning and are ignored during inference. For example, Goo and Chen [55], Li et al. [83], Kim et al. [67] used an additional decoder for dialogue act labeling, topic segmenting and commonsense knowledge generation, respectively. Yuan and Yu [178] incorporated domain features by formulating domain classification as a multi-label binary classification problem for the whole D. All of them use utterance-level features to learn better encoder representations, which will lead to a high-quality summary.

The advantages and disadvantages of injecting pre-processed features are as follows:

✓ Injecting pre-processed features as the mainstream research direction for dialogue summarization significantly improves the results compared with the basic summarization model. Features including negation scope, speaker/role, coreference graph, action graph and semantic slot graph pay more attention to generating consistent summaries, while most of the other features help to select valuable information for summarization.

✓ Such explicitly incorporated features are more interpretable to humans and can be manipulated for more controllable summaries. Different features can be selected and combined to promote the model performance in specific application scenarios.

✓ Features collected by labelers on other dialogue understanding tasks capture the essence of these tasks and also establish connections with various aspects of dialogue analysis. Therefore, leveraging such features is a good way to alleviate the human labeling burden.

✗ Features are not transferable in different scenarios and some features are not compatible with each other, thus feature engineering is shown to be important.

✗ Labelers trained with other datasets are always out-of-domain compared to the targeting dialogue summarization scenario. Hyper-parameters introduced in labeling algorithms with these labelers need try and error for the domain transfer.

✗ Error propagation exists in these dialogue summarization approaches. Incorrect features hinder the understanding of dialogues and lead to poor summaries.

5 DESIGNING SELF-SUPERVISED TASKS

To alleviate human labor and avoid error propagation, self-supervised tasks emerged, which leverage dialogue-summary pairs without additional labels. We divide such tasks into three sub-categories:

- **Denoising tasks** are designed for eliminating noises in the input or penalizing negatives.
- **Masking and recovering tasks** mean that parts of the input are masked and the masked tokens are required to be predicted.
- **Dialogue tasks** refer to response selection and generation for better dialogue understanding.

5.1 Denoising Tasks

Denoising tasks focus on adding noises to the dialogue input or output and aims at generating concise summaries by filtering out the noisy information, resulting in more robust dialogue summarization models. Zou et al. [200] used the original dialogue as output and trained a denoising auto-encoder which is capable of doing content compression for unsupervised dialogue summarization. Noising operations, which include fragment insertion, utterance replacement, and content retention, are applied together on each sample. For a utterance u_t in D, **fragment insertion** means that randomly sampled word spans from u_t is inserted to u_t for lengthening the original sequence. **Utterance replacement** is that u_t is replaced by another utterance $u_{t'}$ in D and **content retention** means that u_t is unchanged. Chen and Yang [22] augmented dialogue data by swapping, deletion, insertion and substitution on utterance level and used the corresponding summary as the output, resulting in more various dialogue inputs for training the summarization model. **Swapping** and **deletion** aim to perturb discourse relations by randomly swapping two utterances in D or deleting some utterances. **Insertion** includes inserting repeated utterances that are chosen from D randomly and inserting utterances with specific dialogue acts such as self-talk or hedge from a pre-extracted set, mimicking interruptions in natural dialogues. **Substitution** replaces the chosen utterances in D by utterances generated with a variant of text infilling task adopted in the BART pre-training process. Only one operation is adopted to noise D at a time, and these operations pay more attention to dialogue characteristics, such as the structure and context information.

This kind of task can be extended to learn beyond the denoising ability when combined with contrastive learning or classification tasks on positive and negative data. Contrastive learning trains the model to maximize the distance between positive data and negative data for learning more informative semantic representations, which extends the classification's ability on generation tasks. Liu et al. [96] proposed coherence detection and sub-summary generation for implicitly modeling the topic change and handling information scattering problems. They cut the dialogue into snippets by sliding windows and separated the long summary into sentences as a first step. **Coherence detection** is to train the encoder to distinguish a snippet with shuffled utterances from the original ordered one. **Designated sub-summary generation** is to train the model to generate more related summaries by constructing negative samples with unpaired dialogue snippets and sub-summaries, where the positive pair is obtained by finding the snippet with the highest Rouge-2 recall for each sub-summary. Tang et al. [152] also designated summaries where negative summaries are constructed for different error types, such as swapping the nouns for wrong reference and object errors, swapping verbs for circumstance errors and tense and modality errors, etc. Positive summaries are collected by back translation. The distance of decoder representations measures the contrastive loss. They also considered the **token identification** task to identify whether two tokens belong to the same speaker based on their encoder representations. Zhao et al. [192] made improvements by **perturbing hidden representations** of the target summary for alleviating the exposure bias following Lee et al. [76], which is useful for conditional generation tasks.

5.2 Masking and Recovering Tasks

Masking and recovering tasks are commonly used in pre-training for better language modeling by recovering the original dialogue and bear some resemblance to the noising operations. The main difference is that these tasks try to recover the original text given the corrupted one. It can be divided into work-level and sentence-level by the granularity of masked contents. Word-level masks for **pronouns** [65], **entities** [65, 102], **high-content tokens** [65], **roles** [130] and **speakers** [195] are considered in previous work, for a better understanding of the complicated speaker characteristics and capturing salient information. Words masked in Khalifa et al. [65]'s work was determined by POS tagger, named entity recognition or simple TF-IDF features. Although the lexical features and statistical features have been captured by pre-trained models for different words as mentioned in Sec. 4, predicting the specific content words under these features reversely given the dialogue context is still challenging and helpful to dialogue context modeling especially with models pre-trained on general text. Utterance-level masking objective inspired by **Gap Sentence Prediction** [181] is adopted by Qi et al. [130]. Key sentence selection from dialogues is done by a graph-based sorting algorithm TextRank and Maximum Margin Relevance. Zhong et al. [195] introduced three new utterance-level tasks: **Turn splitting** is cutting a long utterance into multiple turns and adding "[MASK]" in front of each turn except the first one with the speaker. **Turn merging** is randomly merging consecutive turns into one turn and neglecting the speakers except the first one. And **turn permutation** means that utterances are randomly shuffled.

5.3 Dialogue Tasks

There are also papers incorporating well-known dialogue tasks into dialogue summarization. General **response selection** and **generation** models can be trained with unlabelled dialogues by simply regarding a selected utterance u_t as the output and the utterances before it $u_{<t}$ as the input. Negative candidates for the selection task are the utterances randomly sampled from the whole corpus. Fu et al. [48] assumed that a superior summary is a representative of the original dialogue. So, either inputting D or Y is expected to achieve similar results on other tasks. In this way, the next utterance generation and classification tasks are acted like evaluators, to give guidance on better

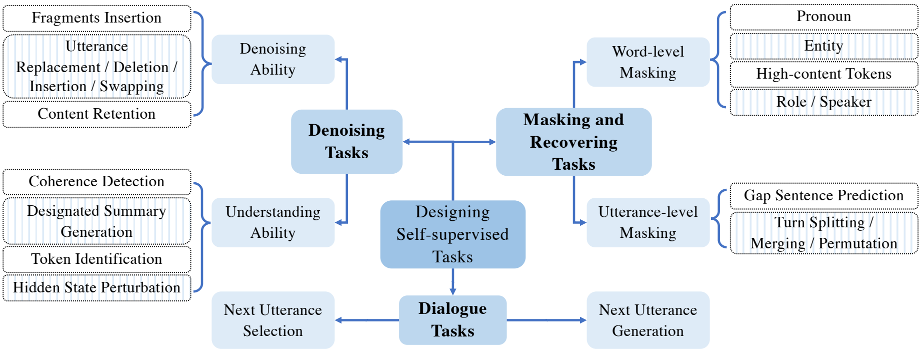


Fig. 6. A summary of self-supervised tasks.

summary generation. Feigenblat et al. [42] trained response selection models for identifying salient utterances. The intuition is that the removal of a salient utterance in dialogue context will lead to a dramatic drop in response selection, and these salient sentences are the same for summarization. Therefore, they regard the drop in probability as a saliency score to rank utterances and adopt the top 4 utterances as the extractive summary, which can be further used to enhance abstractive results by appending it at the end of the input dialogue.

5.4 Summary and Opinions

The self-supervised tasks are summarized in Fig. 6. Most of them are adopted in two ways:

- **Cooperating with the vanilla generation task under different training paradigms.** Multi-task learning refers that the losses from self-supervised tasks are weighed summed with the vanilla generation for updating [48, 192], or updated sequentially in a batch [96]. Pre-training with auxiliary tasks and then fine-tuning on dialogue summarization is also widely accepted [65, 130]. The former is usually selected when the auxiliary training tasks are close to the summarization target. The latter one is chosen for learning more general representations, and is more flexible to use additional data in Sec. 6.
- **Training an isolated model for different purposes.** The model is used as the summarization model directly [42, 200], or as a trained labeler providing information for dialogue summarization [42] with less artificial facts compared with Feng et al. [47].

The advantages and disadvantages of designing self-supervised tasks are as follows:

✓ Most self-supervised tasks take advantage of self-supervision to train the model. They don't need to go through the expensive and time-consuming annotation process for collecting high-quality labels, and avoid the domain transfer problems of transferring labelers trained on the labeled domain to the target summarization domain.

✓ Useful representations are learned with these tasks by the summarization model directly or as an initial state for the summarization model, avoiding the error propagation caused by wrong labels. Although labeling tools such as POS tagger and TextRank are adopted, these predicted labels are not used as the training target or explicitly injected into the summarization model. They are just incorporated to find more effective self-supervisions.

✓ It's a good way to make full use of dialogue-summary pairs without additional labels, or even utilize pure dialogues without summaries.

- ✗ Although designing self-supervised tasks reduces the data pre-processing complexity, it increases the training time and computing costs with additional training targets on corresponding variations of the data.
- ✗ Different self-supervised tasks are not always compatible and controllable. It is challenging to design suitable tasks and find the best combination of tasks in different scenarios.

6 USING ADDITIONAL DATA

Since dialogue summarization data is limited, researchers adopt data augmentation or borrow datasets from other tasks. We divide the data into two categories: Narrative Text and Dialogues.

6.1 Narrative Text

A number of narrative text corpora are utilized to do language modeling and learn commonsense knowledge which is shared across tasks. Since most of today's summarization models are based on pre-trained encoder-decoder models, such as BART [80], PEGASUS [181], and T5 [131], **common crawled text corpora** can be regarded as the backbone corpora of dialogue summarization. It generally includes Wikipedia, BookCorpus [199] and C4 [131]. Pre-trained large language models on top of these corpora, such as GPT-3 [14], can be directly used for dialogue summarization with prefix-tuning approaches [129]. Li et al. [82] transformed such data by dividing the sequence into two spans, selecting span pairs with higher overlaps by Rouge scores for training their model with better copying behaviors. The corresponding overlapped text generation task boosts their proposed model with the correlation copy mechanism on both document and dialogue summarization tasks.

Document summarization is the most similar task to dialogue summarization. As a result, **document summarization data** is a natural choice for learning the summarization ability. Zhang et al. [190] show that BART pre-trained with CNN/DM [58] enhances the dialogue summarization in the meeting and drama scenarios. CNN/DM, Gigaword [134], and NewsRoom [56] were all adopted to train a model from scratch by Zou et al. [202]. For taking advantage of models trained document summarization data to do zero-shot on dialogues, Ganesh and Dingliwal [51] narrowed down the format gap between documents and dialogues by restructuring dialogue with complicated heuristic rules, such as discourse labels mentioned in Sec. 4.2.2 Differently, Zhu et al. [198] shuffled sentences from multiple documents to get a simulated dialogue for pre-training, including CNN/DM, XSum [122] and NYT [136]. Similarly, Park et al. [128] simulated dialogues with three transformation functions: arranging text into dialogue format by adding "Speaker 1:", shuffling sentence order and omitting the most extractive sentences for enhancing the abstractiveness of constructed samples.

Commonsense knowledge data are also welcomed as a basis for language understanding. Khalifa et al. [65] considered three reasoning tasks, including ROC stories dataset [117] for short story ending prediction, CommonGen [85] for generative commonsense reasoning, and ConceptNet for commonsense knowledge base construction. These three tasks, together with dialogue summarization, are jointly trained and show a performance boost. Besides, MSCOCO [90] as a **short text corpus** is used in Zou et al. [202] for training the decoder with narrative text generation ability.

6.2 Dialogue

For collecting or constructing more **dialogue summarization data** without the need for human annotations, data augmentation approaches are proposed. Liu and Chen [101] and Khalifa et al. [65] augmented by replacing person names in both the dialogue and the reference summary at the same time. These augmented data are definitely well-paired and mixed with the original training data. Jia et al. [61] simply paired the whole dialogue with each summary sentence and further trained the model with a prefix-guided generation task before fine-tuning, where the first several tokens of the target sentence are provided for guiding the model on generation and learning to rephrase from dialogue to narrative text to some extent. Asi et al. [8] showed that it is possible to take advantage of giant language models such as GPT-3 [14] to collect pseudo summaries by inputting dialogues with pre-defined question hints. Liu et al. [100] collected augmented training pairs with a small seed dataset by following steps: aligning summary spans with utterances, replacing utterances by reconstruction of the masked dialogue, and filling up the masked summary given the augmented dialogue. Fang et al. [40] augmented and refined the original training pairs with an utterance rewriter [97] and a coreference resolution model [64]. Besides, using relatively large-scaled crawled dialogue summarization data as a pre-training dataset, such as MediaSum [197], for other low-resource dialogue summarization scenarios was considered by Zhu et al. [197]. For crawled data without summaries, Yang et al. [176] constructed pseudo summaries by selecting leading comments from the long forum threads on the Reddit.

Dialogue data without paired summary are also valuable. Feng et al. [46] took questions as outputs and a number of utterances after each question as inputs, regarding question generation as the pre-training objective to help identify important contents in downstream summarization. Khalifa et al. [65] adopted word-level masks on PersonaChat [184] and Reddit comments for fine-tuning. Qi et al. [130] pre-trained with dialogues from MediaSum and TV4Dialogue besides document summarization datasets used in Zhu et al. [198]. They also stitch dialogues randomly to simulate topic transitions. Zhong et al. [195] proposed a generative pre-training framework for long dialogue understanding and comprehension. DialogLM in this work is specially pretrained on dialogues from MediaSum dataset and OpenSubtitles Corpus [91]. It corrupts a window of dialogue utterances with dialogue-inspired noises, similar to the noising operations mentioned in Sec. 5. The original window-sized utterances are the recovering target based on the remaining dialogue. Such a window-based recovering task is suggested to be more suitable for dialogues considering its scattered information and highly content-dependent utterances. Besides, Bertsch et al. [13] took advantage of a self-annotated corpus based on SAMSum [54] which converts each utterance individually to a third-person rephrasing. They showed benefits on the same dataset under the zero-shot setting by pre-training with this perspective shift corpus.

Furthermore, Zou et al. [202] broke the training for dialogue summarization model into three parts, namely encoder, context encoder and decoder, to train the dialogue modeling, summary language modeling and abstractive summarization respectively. Dialogue corpus, short text corpus, and summarization corpus were all used in this work, helping to bridge the gap between out-of-domain pre-training and in-domain fine-tuning, especially for low-resource settings.

6.3 Summary and Opinions

Additional data in previous work are in Fig. 7. These data are always used in the following ways:

- **Pre-training with corresponding training objectives**. Common crawled text data, document summarization data and dialogue data are mostly used in this way [202], where the language styles or data formats are quite different from dialogue-summary pairs. It hopes to provide a better initialization state of the model for dialogue summarization. On the other hand, it is also a good way for coarse-to-fine-grained training, where pre-training is done with the noisy data by data augmentation or from other domains and fine-tuning with the oracle dialogue summarization training data [46, 197].
- **Mixing with dialogue summarization training data** and training for dialogue summarization directly. Data here are usually more similar to dialogue-summary pairs obtained by data augmentation [65, 101] or with intensive commonsense [65, 100].

The advantages and disadvantages of using additional data are as follows:

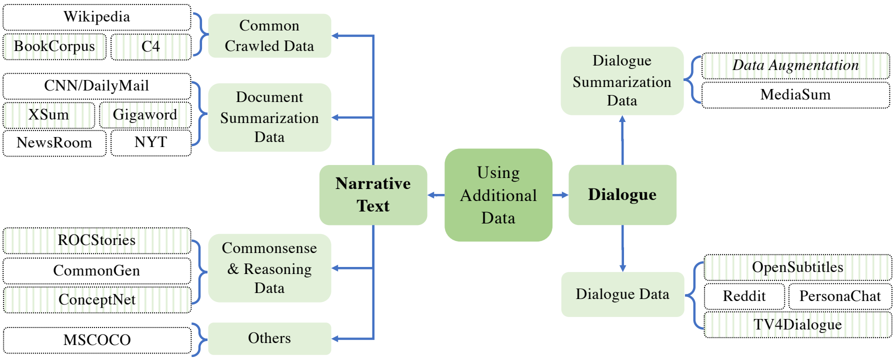


Fig. 7. A summary of additional data.

- ✓ The language understanding ability among different corpora is the same intrinsically. As a result, additional data helps dialogue summarization, especially in low-resource settings, which further alleviates the burden of summary annotation by humans.
- ✓ The intensive knowledge in specially designed corpora helps strengthen the dialogue summarization model.
- ✓ The additional unlabeled data can be trained with self-supervised tasks mentioned in Sec. 5 for better performance.
- ✗ Training with additional data makes significant improvements while requiring more time and computational resources, reflecting the data inefficiency of current models.
- ✗ Training with more data is not always effective [119, 190], especially when the divergence between the additional corpus and original dialogue summarization corpus is huge. Elaborate data augmentation approaches avoid this problem when training data is not too scarce.

7 EVALUATIONS

We present a comprehensive description of existing dialogue summarization datasets under different scenarios and introduce several widely-accepted evaluation metrics for this task.

7.1 Datasets

A great number of dialogue summarization datasets have been proposed. We categorize them according to the scenarios in Sec. 2.3.

7.1.1 Open-domain Dialogue Summarization. Open-domain dialogue summarization datasets under daily chat, drama conversation and debate&comment are as follows and summarized in Table 1.

Daily Chat Datasets: **SAMSum** [54] and **DialogSum** [28] are two large-scale real-life labeled datasets. Each dialogue in SAMSum is written by one person to simulate a real-life messenger conversations and the single reference summary is annotated by language experts. DialogSum, on the other hand, contains dialogues from existing datasets, including DailyDialog [84], DREAM [149] and MuTual [35], and English-speaking practice websites. These spoken dialogues have a more formal style than those in SAMSum, and each is accompanied by three reference summaries in the test set. Besides, **HubDial** [4] also contains dialogues covering a range of daily topics.

Drama Conversation Datasets: **CRD3** [132] is collected from a live-stream role-playing game called Dungeons and Dragons, which is more amenable to extractive approaches with low abstractiveness.

[4] https://aihub.or.kr/

MediaSum [197] includes interview transcripts from NPR and CNN and their reviews or topic descriptions are regarded as the corresponding summaries. Other two datasets are collected from a variety of movies and TV series, including **SubTitles** [109] and **SummScreen** [24]. Dialogues are corresponding transcripts, and summaries are aligned synopses or recaps written by humans.

Debate&Comment Datasets: **ADSC** [114] is a test-only dataset. It contains 45 two-party dialogues about social and political topics, each associated with 5 reference summaries. **FORUM** [154] contains human-annotated forum threads collected from tripadvisor.com and ubuntuforums.org. Three out of four sub-datasets in **ConvoSumm** [39] are similar discussions, including news article comments (**NYT**), discussion forums and debate (**Reddit**) and community question answers (**Stack**) from different sources. Each sample has a human-written reference. **CQASUMM** [32] is another community question answering dataset without back and forward discussions among speakers. Its summary aims to summarize multiple answers, close to a multi-document summarization setting.

Table 1. Open-domain dialogue summarization datasets. "Lang." and "Spk" stands for "Language" and "Speakers". "DW" and "SW" represents the average number of words in the dialogues and summaries respectively. "AVL" refers to the public availability of the dataset (Y is available, N is not available, and C is conditional). HubDial is only available for Koreans.

Name	#Samples train/val/test	#Spk	Lang.	DW	SW	AVL
Daily Chat						
SAMSum [54]	14.7k/0.8k/0.8k	≥2	English	94	25	Y
DialogSum [28]	12.5k/0.5k/0.5k	2	English	131	22	Y
HubDial	350k	≥2	Korean	-	-	C
Drama Conversation						
CRD3 [132]	26.2k/3.5k/4.5k	≥2	English	31,803	2,062	Y
MediaSum [197]	463.6k/10k/10k	≥2	English	1,554	14	Y
SumTitles [109](Subtitiles/Scripts/Gold)	132k 21k 290	≥2	English	6,406 423 395	85 55 51	Y
SummScreen [24](FD/TMS)	3,673/338/337 18,915/1,795/1,793	≥2	English	7,605 6,421	114 381	Y
Debate & Comment						
ADSC [114]	45	2	English	672	151	Y
CQASUMM [32]	100k	≥2	English	782	100	Y
FORUM [154]	689	≥2	English	825	191	Y
ConvoSumm [39](NYT/Reddit/Stack)	-/0.25k/0.25k -/0.25k/0.25k -/0.25k/0.25k	≥2	English	1,624 641 1,207	79 65 73	Y

7.1.2 Task-oriented Dialogue Summarization. Datasets here are rooted in specific domains, including customer service, law, medical care and official issue. We list them in Table 2.

Customer Service Datasets: Zou et al.[200, 201] propose two similar datasets with summaries from the agent perspective. Lin et al. [88] provides a more fine-grained dataset **CSDS** containing a user summary, an agent summary, and an overall summary based on JDDC dataset [25]. Summaries from **Didi dataset** [94] are also written from agents' points of view, in which dialogues are about transportation issues instead of pre-sale and after-sale topics in the former one. More complicated multi-domain scenarios are covered in **TWEETSUMM** [42], **MultiWOZ*** [178] and **TODSum** [193]. Dialogues from TWEETSUMM spread over a wide range of domains, including gaming, airlines, retail, and so on. MultiWOZ* and TODSum transform and annotate summaries based on the original MultiWOZ [15]. **DECODA** and **LUNA** [41] are two earlier datasets containing call centre conversations with synopses summarizing the problem of the caller and solutions.

Law Datasets: **Justice** [48] includes debates between a plaintiff and a defendant on some controversies which take place in the courtroom. The final factual statement by the judge is regarded as

Table 2. Task-oriented dialogue summarization datasets. The original text data is not accessible for PLD. DECODA, LUNA and LCSPIRT-DM have to be obtained through an application. EmailSum is not free.

Name	#Samples train/val/test	#Spk	Lang.	DW	SW	AVL
Customer Service						
Zou et al. [201]	17.0k/0.9k/0.9k	2	Chinese	1,334	55	Y
CSDS [88]	9.1k/0.8k/0.8k	2	Chinese	401	83	Y
Zou et al. [200]	-/0.5k/0.5k	2	Chinese	95	37	Y
Didi [94]	296.3k/2.9k/29.6k	2	Chinese	-	-	N
TWEETSUMM [42]	0.9k/0.1k/0.1k	2	English	245	36	Y
MultiWOZ* [178]	8.3k/1k/1k	2	English	181	92	Y
TODSum [193]	9.9k	2	English	187	45	N
DECODA [41]	-/50/100	2	French/ English	42,130 41,639	23 27	C
LUNA [41]	-/-/100	2	Italian/ English	34,913 32,502	17 15	C
Law						
Justice [48]	30k	2	Chinese	605	160	N
PLD [38]	5.5k	≥2	English	-	-	C
LCSPIRT-DM [169]	30.8/3.8k/3.8k	2	Chinese	684	75	C
Medical Care						
Joshi et al. [63]	1.4k/0.16k/0.17k	2	English	-	-	N
Song et al. [145]	36k/-/9k	2	Chinese	312	23/113	Y
Liu et al. [105]	100k/1k/0.49k	2	English	-	-	N
Zhang et al. [182]	0.9k/0.2k/0.2k	2	English	-	-	N
Official Issue (Meeting & Emails)						
AMI [17]	137	>2	English	4,757	322	Y
ICSI [60]	59	>2	English	10,189	534	Y
QMSum [196]	1.3k/2.7k/2.7k	>2	English	9070	70	Y
Kyutech [120, 174]	9	>2	Japanese	-	-	Y
BC3 [158]	30	>2	English	550	134	Y
Loza et al. [107]	107	>2	English	-	-	N
EmailSum [183]	1.8k/0.25k/0.5k	≥2	English	233	27/69	C
ConvoSumm [39](Email)	-/0.25k/0.25k	≥2	English	917	74	Y

the summary. A similar scenario is included in **PLD** [38], which is more difficult to summarize due to the unknown number of participants. There is also another version of PLD by Gan et al. [50] with fewer labeled cases. Xi et al. [169] proposed a long text summarization dataset **LCSPIRT-DM** based on police inquiry records full of questions and answers.

Medical Care Datasets: Both Joshi et al. [63] and Song et al. [145] proposed medical summarization corpora by crawling data from online health platforms and annotating coherent summaries by doctors. Song et al. [145] also proposed one-sentence summaries of medical problems uttered by patients, whereas Liu et al. [105] used simulated data with summary notes in a structured format. Zhang et al. [182] used unreleased dialogues with coherent summaries of the history of the illness.

Official Issue Datasets: **AMI** [17] and **ICSI** [60] are meeting transcripts concerning computer science-related issues in working background and research background. Both datasets are rich in human labels, including abstractive summary, topic segmentation, and so on. They are also included in **QMSum** [196] with annotations for query-based meeting summarization. **Kyutech** [174] is a similar dataset in Japanese containing multi-party conversations, where the participants pretend to be managers of a shopping mall in a virtual city and do some decision-making tasks. Their later work [120] annotated more fine-grained summaries for each topic instead of the whole conversation. In addition, official communications are also prevalent in e-mails. Ulrich et al. [158] propose the first email summarization dataset **BC3** with only 30 threads and Loza et al. [107] release 107 threads. Both of them contain extractive as well as abstractive summaries. **EmailSum** [183] has both a human-written short summary and a long summary for each e-mail thread. Besides, Email threads (**Email**) in ConvoSumm [39] have only one abstractive summary for each dialogue.

7.1.3 Summary. We make the following observations and conclusions.

- The size of dialogue summarization datasets is much smaller than document summarization datasets. Most dialogue summarization datasets have no more than $30K$ samples, while representative document summarization datasets, such as CNNDM and XSum, have more than $200K$ samples. Datasets for drama conversations are relatively larger and can be potential pre-training data for other scenarios.
- The number of interlocutors in different dialogue summarization scenarios is different. Most ODS dialogues have more than 2 speakers while most dialogues in TDS have only 2 speakers except in official meetings or e-mails.
- TDS dialogues tend to be more private. Thus, half of the TDS datasets are not publicly available, especially for Law and Medical Care scenarios.
- Datasets with more than 4,096 dialogue words, which is the upper bound of the input length of most pre-trained language models, are suitable for research on long dialogue summarization. They contain both open-domain datasets and task-oriented datasets.

7.2 Evaluation Metrics

In existing works, *automatic evaluation metrics* commonly used for summarization, such as **Rouge** [86], **BERTScore** [185] and **BARTScore** [179], are also used for dialogue summarization by comparing the generations with references. However, these widely-accepted metrics' performance may deviate from human [57], especially in the aspect of consistency. Therefore, more focused automatic and human evaluations emphasizing *information coverage* and *factual consistency* are considered.

Instead of comparing only with the whole reference summary, most researches for TDS only consider key words/phrases while ignoring other common words for measuring the **information coverage**. In other words, evaluation for TDS emphasizes the coverage of key information which are generally domain-specific terms and can be easily recognized. For example, medical concept coverage [63, 182] and critical information completeness [178] both extract essential phrases based on domain dictionaries by rules or publicly available tools. Zhao et al. [192] uses slot-filling model [26] to recognize slot values for factual completeness. Then, the accuracy or F1 scores are calculated by comparing extracted phrases or concepts from Y and Y'.

ODS pays less attention to information coverage due to the higher subjectivity on salient information selection. Instead, measuring the **factual consistency** of generations gains increasing attention. Unlike the above metrics which compare generations with the reference summary, most evaluation metrics here compare generations with the source dialogue and can be classified into reference-free evaluation metrics [98, 140]. A QA-based model [162] is borrowed by Zhao et al. [192]. It follows the idea that factually consistent summaries and documents generate the same answers to a question. NLI-based methods [110] that require the content in the summary to be fully inferred from the dialogue were adopted by Liu et al. [100]. Liu and Chen [101] automatically evaluate inconsistency issues of person names by using noised reference summaries as negative samples and training a BERT-based binary classifier. Asi et al. [8] used the FactCC metric [72] where the model was trained only with source documents with a series of rule-based transformations. Information correctness is also important for TDS. For instance, negation correctness is considered by Joshi et al. [63] with publicly available tools for recognizing negated concepts.

Meanwhile, *human evaluations* are required to complement the above metrics. Besides ranking or scoring the generated summary with an overall quality score [21], more specific aspects are usually provided to annotators. Representative ones include: **readability/fluency** [178, 192] requiring a summary to be grammatically correct and well structured, **informativeness** [42, 46, 47, 78] measuring how well the summary includes salient information, **conciseness/non-redundancy** [47, 178] pursuing a summary without redundancy, and **factualness/consistency** [46, 67, 78, 192] evaluating whether the summary is consistent with the source dialogue. There are also some typical fine-grained metrics evaluating errors in generated summaries mentioned in previous works [21, 28, 106]: **Information missing** means that content mentioned in references are missing in generated summaries, while **information redundancy** is the opposite. **Reference error** refers to wrong associations between a speaker and an action or a location. **Reasoning error** is that the model reasons incorrectly among multiple dialogue turns. Moreover, Chen and Yang [21] mentioned **improper gendered pronouns** referring to improper gendered pronouns. Tang et al. [152] proposed **circumstantial error**, **negation error**, **object error**, **tense error** and **modality error** for detailed scenarios. These error types can also be grouped into two classes, where the information missing and redundancy are for information coverage, and the rest are for factual consistency.

8 ANALYSIS AND FUTURE DIRECTIONS

We first present a statistical analysis of the papers covered in this survey. Then, some future directions are proposed inspired by our observations.

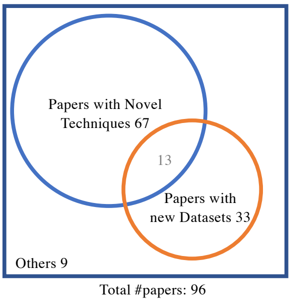


Fig. 8. Statistics of abstractive dialogue summarization papers.

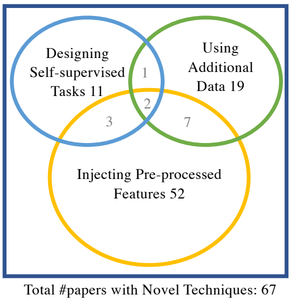


Fig. 9. Statistics of papers with technical contributions.

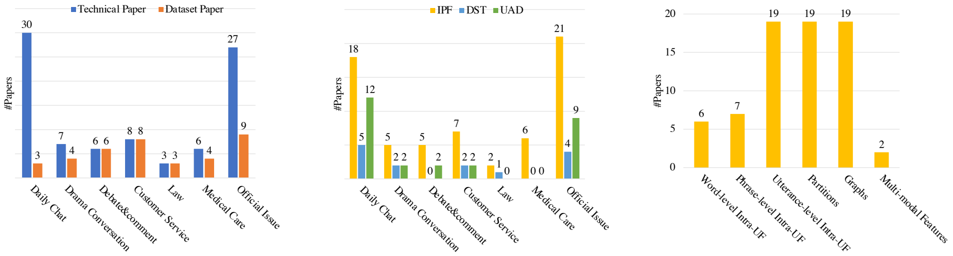
8.1 Paper Analysis

The total number of papers on abstractive dialogue summarization investigated in this survey is 96. As shown in Fig. 8, 33 of them propose new datasets and 67 make novel technical contributions. The other 9 papers are either a survey, a demo, or other strongly related papers. The overall ratio between technical papers and dataset papers (*tech-data ratio*) is around 2.03 : 1. Compared with the number of papers under different application scenarios in Fig. 10(a), we found that scenarios of daily chat and official issues receive more attentions. However, the other scenarios are less explored, with much lower tech-data ratios ranging from 1.0 to 1.75. There is no significant difference in the number of datasets between well-researched domains and the others. However, the release time and availability of different datasets vary. AMI and ICSI are well-known meeting summarization datasets released in the early stage of the 20th century, while most other datasets have been proposed in recent years. Datasets for daily chat are all publicly available, while datasets for medical care and laws are not accessible to the majority of researchers. It's a good sign that high-quality corpora, such as AMI and SAMSum, lead to a prosperous of techniques for dialogue summarization, but also raise a worry about the generalization ability of current techniques because of their over-reliance on specific datasets which may lead to over-fitting.

Table 3. Existing work on injecting pre-processed features for different scenarios. The taxonomy of features and dialogue summarization scenarios are in the columns and rows respectively. A work may appear multiple times since it experimented with datasets under various scenarios or utilized features in different groups.

Features / Scenarios	Intra-Utterance Features			Inter-Utterance Features		Multi-modal Features
	Word level	Phrase level	Utterance level	Partitions	Graphs	
Open-domain Dialogue Sumamrization						
Daily Chat	[129]	[47][65] [127][168]	[8][47][67] [77][78][129] [168][180]	[8][21] [47][96]	[23][43][78] [101][106][127] [192][104]	-
Drama Conversation	-	[127]	-	[81][96][189]	[127][192]	-
Debate & Comment	-	[127]	[176]	-	[23][39] [43][127][176]	-
Task-oriented Dialogue Sumamrization						
Customer Service	-	[201]	[8][176][178] [187][201]	[8][200]	[176][178][193]	-
Law	-	[50]	[38][50]	-	-	-
Medical Care	-	[63]	[145]	[71][105][182]	[115]	-
Official Issue (Meeting&Email)	[118][126][130] [143][198]	[47][127]	[37][47][55] [118][130][176] [198]	[10][37][47] [70][83][103] [130][139][189] [194][196]	[10][46][51] [111][126][127] [139][176]	[83][118]

The distribution of technical papers in each of the three research directions is shown in Fig. 9. While 11 and 19 papers focus on designing self-supervised tasks and using additional data, respectively, more than 77% of the entire body of works targets the injection of pre-processed features. The trends of paper account for different techniques across scenarios that are similar to each other according to Fig. 10(b). The number of papers using features under different categories is shown in Fig. 10(c), and we go for a deep insight into correlations between features and applications scenarios by categorizing papers according to features and their tested scenarios in Table 3.



(a) (b) (c)

Fig. 10. (a) The number of technical papers and dataset papers under different scenarios. (b) The number of technical papers dividing by directions under different scenarios. IPF, DST and UAD are short for the three directions. (c) The number of technical papers under different features. Intra-UF is intra-utterance features.

We make following observations:

- Scenarios of Official Issue and Daily Chat attracted the most attentions while other scenarios lack research as mentioned before.
- Utterance-level intra-utterance features and inter-utterance features are widely exploited, indicating that modeling utterance-level or beyond utterance-level features is more effective for dialogue understanding. Among them, speaker/role information and topic transitions are

two features which work well under both ODS and TDS scenarios. There is also a lack of attention on multi-modal features, possibly due to the scarcity of multi-modal datasets.
- Word-level and phrase-level intra-utterance features are no longer required with the wide adoption of pre-trained language models, except in integrating domain dictionaries in TDS. These features, especially keywords, are preferred to use as nodes for further constructing graphs, which helps capture the global information flows for both ODS and TDS.
- Partitions are extremely effective for TDS where dialogues are usually long with inherent semantic transitions, such as agendas for meetings and domain shifts in customer service. Identifying these transitions achieves a high degree of consensus among annotators. In contrast, semantic flows in ODS are often interleaved in a complex fashion, which can be better represented as graphs, such as discourse graphs and topic graphs.

8.2 Future Directions

We discuss some possible future directions and organize them into three dimensions: *task scenarios*, *approaches* and *evaluations*.

8.2.1 More Complicated and Controllable Scenarios. Newly explored scenarios such as multi-lingual, multi-modal, multi-session and personalized dialogue summarization are worth researching.

Multilingual dialogue summarization is a rising topic. It considers multiple languages existing in the dialogue and summary on three levels of granularity. The most fine-grained one considers interactions between peers who are fluent in multiple languages resulting in the intra-utterance multilingual phenomenon is called “code-mixing” strictly [112]. Second, dialogues happening among multinational participants where they use their mother tongue to communicate lead to the inter-utterance multilingual phenomenon is called “code-switching” [112], i.e., mix-lingual in [45]. Third, summarizing a monolingual dialogue in a different language is called “cross-lingual” in Wang et al. [164]. Different multilingual datasets have been constructed for these settings based on the existing datasets [28, 54, 196, 197] by human annotations [30, 112, 164] or machine translation [45]. Preliminary studies in these papers show the potential of end-to-end multilingual models, such as mBART [153], in this task and their weaknesses in low-resource languages, poor domain transfer ability [164] and performance drops when processing multiple languages with a single model [45]. Chen et al. [29, 30] proposed the cross-lingual conversation summarization challenge, paving the way for the prosperity of research in this direction. Our survey focuses on approaches for monolingual dialogue summarization, which we expect to provide a backbone for this raising area.

Multi-modal dialogue summarization refers to dialogues occurring in multi-modal settings, which are rich in non-verbal information that often complements the verbal part and therefore contributes to summary contents. Some early work did research on speech dialogue summarization. However, most of them only extract audio features from speech and text features from ASR transcripts independently to produce extractive summaries. There is also work on video summarization [59] focusing on highlighting critical clips while a textual summary is not considered. Fusing the synchronous and asynchronous information among modalities is challenging. AMI and ICSI are still valuable resources for research on multi-modal dialogue summarization.

Multi-session dialogue summarization is required when conversations occur multiple times among the same group of speakers. The Information mentioned in previous sessions becomes their consensus and may not be explained again in the current session. The summary generated merely from the current session is unable to recover such information and may lead to implausible reasoning. A similar multi-session task has been proposed by Xu et al. [171]. This setting also has some correlations with life-long learning [93]. Such multi-session dialogues exists in ODS datasets, such as SubTitles [109] and SummScreen [24]. However, current approaches generally break down the long dialogue and summary into shorter chunks. For task-oriented scenarios, it is also common in real life. For example, the customer may repeatedly ask for help from the agent with the same issue that hasn't been solved before. An updated summary covering the questions and answers can remind participants of the long dialogue history and therefore facilitate the negotiation process.

Recent work mainly focuses on summarizing the dialogue content but ignores the speaker-related or reader-related information. **personalized dialogue summarization** can be understood in two ways. On the one hand, a personalized dialogue refers to the consideration of personas for interlocutors in dialogues. For example, the character role-playing information is indispensable information for generating summaries given dialogue from CRD3 [132]. On the other hand, it refers to generating different dialogue summaries for different readers or speakers. Tepper et al. [156] is a demo paper raising the requirements for personalized chat summarization. They did the first trial on this task considering the personalized topics of interests and social ties during the selection of dialogue segments to be summarized. Some task-oriented datasets, such as CSDS [88], contain summaries from both the user and agent aspect are similar to the problem here. Recent work from Lin et al. [89] solved this problem by adding the cross attention interaction and the decoder self-attention interaction to interactively acquire other roles' critical information. This work is designed only for scenarios with two roles. Scenarios with a variety number of speakers and summary readers from different social groups pose more challenges, raising an expectation for related datasets and approaches, which is a possible interdisciplinary research orientation.

8.2.2 Innovations in Approach. Approach innovations include four parts: feature analysis, person-related features, generalizable and non-labored techniques, and the robustness of models.

From Sec. 8.1, although tens of papers introduce different features for dialogue summarization, there is still a lot of work to do. Comprehensive experiments to **compare the features** and their combinations upon the same benchmark are expected, for features both in the same category or across categories. One can consider unifying the definition of similar features, e.g., different classification criteria of discourse relations or graphs emphasizing phrase-level semantic flows. These analyses would help design features for new applications and interpret dialogue models.

More **person-related features** can be incorporated, such as speaker personalities [188] and emotions [108]. A speaker's background can help understand the underlying motivation and select the content to be summarized, especially for personalized dialogue summarization. A plug-and-play mechanism on top of the decoder for persona-controlled summarization may be a solution [36].

Generalizable and non-labored techniques have attracted increasing attention on other dialogue modeling tasks, such as multi-turn response selection [172] and dialogue generation [188]. These works proposed different self-supervised training tasks, largely relieving human labor. However, dialogue summarization approaches overwhelmingly rely on injecting pre-processed features, which are mostly labor-intensive and has poor generalization ability among scenarios. Recently, large language models (LLMs) with tons of billions of parameters like GPT-3 [14] and LLaMA [157] have demonstrated drastically lifted text generation ability compared to previous pre-trained language models. To accomplish summarization task, LLMs are typically prompted with instructions like "*Summarize the above article:*" or chain-of-thought [166, 167] methods that elicit LLMs to extract various features, e.g., events, that are helpful to compose the final summary. Compared with traditional methods, LLM-based methods largely alleviate the tedious human labor and can be more generalizable due to the removal of unintended annotation artifacts. Nevertheless, approaches previously applied to small pre-trained language models may also provide inspirations and be adapted to augment LLMs for better dialogue summarization performance.

Approaches nowadays are mostly built on the pre-trained language models, which are sensitive to trivial changes [165, 175]. Nevertheless, the **robustness of models** hasn't been widely-investigated in dialogue summarization. The only work from Jia et al. [62] proposed that switching an ungrounded speaker name shouldn't influence the models' generation. According to their experiments with BART fine-tuned on SAMSum, such changes can lead to dramatically different summaries with information divergence and various reasoning results. This may further result in unintended ethical issues by showing discrimination against specific groups of names. Thus, analysis and improvements of models' robustness are in urgent need for practical applications.

8.2.3 Datasets and Evaluation Metrics. Expectations on datasets and evaluation metrics for dialogue summarization are as follows.

Sec. 8.1 shows that **high-quality datasets** expedite the research. Besides the expectations on benchmark datasets for the above emerging scenarios, datasets for task-oriented dialogue summarization with privacy issues are also sought after. They can be in small sizes with real cases after anonymization or can be collected by selecting drama conversations in specific scenarios and annotated with domain experts.

Evaluation metrics are significant which guides the improvement directions for upcoming models. However, widely used evaluation metrics in Sec. 7.2 are all borrowed from document summarization tasks and their effectiveness is unverified. Recent work from Gao and Wan [52] re-evaluated 18 evaluation metrics and did a unified human evaluation with 14 dialogue summarization models on SAMSum dataset. Their results not only show the inconsistent performances of metrics between document summarization and dialogue summarization, and none of them excel in all dimensions for dialogue summarization, but also raise a warning on rethinking whether recently proposed complex models and fancy techniques truly improve the backbone language model. Considering that human evaluation results are difficult to reproduce due to variations of annotator background and unpredictable situations in the annotation progress [33], automatic metrics specially designed for dialogue summarization are urgently needed.

Factual errors caused by the mismatch between speakers and events are common as a result of complicated discourse relations among utterances in dialogues. Tang et al. [152] introduced a taxonomy of factual errors for abstractive summarization and did human evaluation based on this categorization. Liu and Chen [101] made the first attempt by inputting the dialogue and summary together into a BERT-based classifier and claimed high accuracy on their own held-out data. Wang et al. [163] classified factual errors in a similar way to Tang et al. [152] and propose a model-level evaluation schema for discriminating better summarization models, which is different from the widely-accepted sample-level evaluation schema that scores generated summaries. They evaluated the model by calculating the generation probability of faithful and unfaithful summaries collected by rule-based transformations. The generalization ability for this work is doubtful, since a similar work, FactCC [72], which is a metric trained based on rule-based synthetic datasets shows a poor generalization ability [74]. With the strong generation ability of current LLMs, there's another doubt that whether the previous taxonomy of error types and evaluation metrics are still suitable. In a word, both **meta-evaluation benchmarks** and **evaluation methods** call for innovations.

9 CONCLUSION

Dialogue summarization is receiving increasing demands in recent years for releasing the burden of manual summarization and achieving efficient dialogue information digestion. It is a cross-research direction of dialogue understanding and text summarization. Abstractive summarization is a natural choice for dialogue summarization due to the characteristics of dialogues, including information sparsity, context-dependency, and the format discrepancy between utterances in first person and the summary from the third point of view. With the success of neural-based models especially pre-trained language models, the quality of generated abstractive summaries appears to be promising for real applications. This survey summarizes a wide range of papers on the subject. In particular, it presents a taxonomy for task scenarios, made up of two broad categories, i.e., open-domain dialogue summarization and task-oriented dialogue summarization. A great many techniques developed in different approaches are categorized into three directions, including injecting pre-processed features, designing self-supervised tasks, and using additional data. We also collect a number of evaluation benchmarks proposed so far and provide a deep analysis with valuable future directions. This survey is a comprehensive checkpoint of dialogue summarization research thus far and is expected to inspire the researchers to rethink this task and search for new opportunities, especially with current LLMs. It is also a useful guide for engineers looking for practical solutions.

REFERENCES

[1] Stergos D. Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *EMNLP*. 928–937.

[2] Alexander A Alemi and Paul Ginsparg. 2015. Text segmentation based on semantic word embeddings. *arXiv preprint arXiv:1503.05543* (2015).

[3] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Trans. Assoc. Comput. Linguistics* 4 (2016), 463–476.

[4] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging Linguistic Structure For Open Domain Information Extraction. In *ACL, Volume 1: Long Papers*. 344–354.

[5] Jaime Arguello and Carolyn Rosé. 2006. Topic-segmentation of dialogue. In *Proceedings of the Analyzing Conversations in Text and Speech*. 42–49.

[6] Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos D. Afantenos. 2016. Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus. In *LREC*.

[7] Nicholas Asher and Alex Lascarides. 2005. *Logics of Conversation.* Cambridge University Press.

[8] Abedelkadir Asi, Song Wang, Roy Eisenstadt, Dean Geckt, Yarin Kuper, Yi Mao, and Royi Ronen. 2022. An End-to-End Dialogue Summarization System for Sales Calls. *arXiv preprint arXiv:2204.12951* (2022).

[9] Jiaxin Bai, Hongming Zhang, Yangqiu Song, and Kun Xu. 2021. Joint Coreference Resolution and Character Linking for Multiparty Conversation. In *EACL*. 539–548.

[10] Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015. Generating Abstractive Summaries from Meeting Transcripts. In *ACM DocEng*. 51–60.

[11] Michele Banko, Vibhu O. Mittal, and Michael J. Witbrock. 2000. Headline Generation Based on Statistical Translation. In *ACL*. 318–325.

[12] Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *ACL-HLT*. 481–490.

[13] Amanda Bertsch, Graham Neubig, and Matthew R.Gormley. 2022. He Said, She Said: Style Transfer for Shifting the Perspective of Dialogues. In *Findings of EMNLP*.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS* 33 (2020), 1877–1901.

[15] Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *EMNLP*. 5016–5026.

[16] Harry Bunt. 1994. Context and dialogue control. *Think Quarterly* 3, 1 (1994), 19–31.

[17] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI Meeting Corpus: A Pre-announcement. In *MLMI (Lecture Notes in Computer Science, Vol. 3869)*. Springer, 28–39.

[18] Asli Celikyilmaz and Dilek Hakkani-Tur. 2010. A hybrid hierarchical model for multi-document summarization. In *ACL*. 815–824.

[19] Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions. In *EMNLP-IJCNLP*. 2933–2943.

[20] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *SIGKDD Explor.* 19, 2 (2017), 25–35.

[21] Jiaao Chen and Diyi Yang. 2020. Multi-View Sequence-to-Sequence Models with Conversational Structure for Abstractive Dialogue Summarization. In *EMNLP*. 4106–4118.

[22] Jiaao Chen and Diyi Yang. 2021. Simple Conversational Data Augmentation for Semi-supervised Abstractive Dialogue Summarization. In *EMNLP*. 6605–6616.

[23] Jiaao Chen and Diyi Yang. 2021. Structure-Aware Abstractive Conversation Summarization via Discourse and Action Graphs. In *NAACL-HLT*. 1380–1391.

[24] Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2021. SummScreen: A Dataset for Abstractive Screenplay Summarization. (2021). arXiv:arXiv:2104.07091

[25] Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The JDDC Corpus: A Large-Scale Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service. In *LREC*. 459–466.

[26] Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. (2019). arXiv:arXiv:1902.10909

[27] Yen-Chun Chen and Mohit Bansal. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *ACL, Volume 1: Long Papers*. 675–686.

[28] Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. In *Findings of ACL/IJCNLP*. 5062–5074.

[29] Yulong Chen, Huajian Zhang, Yijie Zhou, Xuefeng Bai, Yueguan Wang, Ming Zhong, Jianhao Yan, Yafu Li, Judy Li, Xianchao Zhu, and Yue Zhang. 2023. Revisiting Cross-Lingual Summarization: A Corpus-based Study and A New Benchmark with Improved Annotation. In *ACL (Volume 1: Long Papers)*. 9332–9351.

[30] Yulong Chen, Ming Zhong, Xuefeng Bai, Naihao Deng, Jing Li, Xianchao Zhu, and Yue Zhang. 2022. The Cross-lingual Conversation Summarization Challenge. *arXiv preprint arXiv:2205.00379* (2022).

[31] Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *ANLP*. 26–33.

[32] Tanya Chowdhury and Tanmoy Chakraborty. 2019. CQASUMM: Building References for Community Question Answering Summarization Corpora. In *ACM COMAD/CODS*. 18–26.

[33] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *ACL/IJCNLP, Volume 1: Long Papers*. 7282–7296.

[34] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *NAACL-HLT, Volume 2 (Short Papers)*. 615–621.

[35] Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A Dataset for Multi-Turn Dialogue Reasoning. In *ACL*. 1406–1416.

[36] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *ICLR*.

[37] Jiasheng Di, Xiao Wei, and Zhenyu Zhang. 2020. How to Interact and Change? Abstractive Dialogue Summarization with Dialogue Act Weight and Topic Change Info. In *KSEM, Part II (Lecture Notes in Computer Science, Vol. 12275)*. 238–249.

[38] Xinyu Duan, Yating Zhang, Lin Yuan, Xin Zhou, Xiaozhong Liu, Tianyi Wang, Ruocheng Wang, Qiong Zhang, Changlong Sun, and Fei Wu. 2019. Legal Summarization for Multi-role Debate Dialogue via Controversy Focus Mining and Multi-task Learning. In *CIKM*. 1361–1370.

[39] Alexander R. Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir R. Radev. 2021. ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining. In *ACL/IJCNLP, Volume 1: Long Papers*. 6866–6880.

[40] Yue Fang, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Bo Long, Yanyan Lan, and Yanquan Zhou. 2022. From spoken dialogue to formal summary: An utterance rewriting for dialogue summarization. In *NAACL*. 3859–3869.

[41] Benoit Favre, Evgeny Stepanov, Jérémy Trione, Frédéric Béchet, and Giuseppe Riccardi. 2015. Call centre conversation summarization: A pilot task at multiling 2015. In *SIGdial*. 232–236.

[42] Guy Feigenblat, R. Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. TWEETSUMM - A Dialog Summarization Dataset for Customer Service. In *Findings of EMNLP*. 245–260.

[43] Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. Incorporating Commonsense Knowledge into Abstractive Dialogue Summarization via Heterogeneous Graph Networks. In *CCL (Lecture Notes in Computer Science, Vol. 12869)*. 127–142.

[44] Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. (2021). arXiv:arXiv:2107.03175

[45] Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. MSAMSum: Towards Benchmarking Multi-lingual Dialogue Summarization. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*. 1–12.

[46] Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021. Dialogue Discourse-Aware Graph Model and Data Augmentation for Meeting Summarization. In *IJCAI*. 3808–3814.

[47] Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization. In *ACL/IJCNLP, Volume 1: Long Papers*. 1479–1491.

[48] Xiyan Fu, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Changlong Sun, and Zhenglu Yang. 2021. RepSum: Unsupervised Dialogue Summarization based on Replacement Strategy. In *ACL/IJCNLP, Volume 1: Long Papers*. 6042–6051.

[49] Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. 2021. Paragraph-level commonsense transformers with recurrent memory. In *AAAI*, Vol. 35. 12857–12865.

[50] Leilei Gan, Yating Zhang, Kun Kuang, Lin Yuan, Shuo Li, Changlong Sun, Xiaozhong Liu, and Fei Wu. 2021. Dialogue Inspectional Summarization with Factual Inconsistency Awareness. (2021). arXiv:arXiv:2111.03284

[51] Prakhar Ganesh and Saket Dingliwal. 2019. Restructuring Conversations using Discourse Relations for Zero-shot Abstractive Dialogue Summarization. (2019). arXiv:arXiv:1902.01615

[52] Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. In *NAACL-HLT*. 5693–5709.

[53] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML*. PMLR, 1243–1252.

[54] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. 70–79.

[55] Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive Dialogue Summarization with Sentence-Gated Modeling Optimized by Dialogue Acts. In *IEEE SLT*. 735–742.

[56] Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *NAACL-HLT, Volume 1 (Long Papers)*. 708–719.

[57] Michael Hanna and Ondřej Bojar. 2021. A Fine-Grained Analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*. 507–517.

[58] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *NeurIPS*. 1693–1701.

[59] Tanveer Hussain, Khan Muhammad, Weiping Ding, Jaime Lloret, Sung Wook Baik, and Victor Hugo C. de Albuquerque. 2021. A comprehensive survey of multi-view video summarization. *Pattern Recognit.* 109 (2021), 107567.

[60] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI Meeting Corpus. In *ICASSP*. 364–367.

[61] Qi Jia, Yizhu Liu, Haifeng Tang, and Kenny Zhu. 2022. Post-Training Dialogue Summarization using Pseudo-Paraphrasing. In *Findings of NAACL*. 1660–1669.

[62] Qi Jia, Haifeng Tang, and Kenny Zhu. 2023. Reducing Sensitivity on Speaker Names for Text Generation from Dialogues. In *Findings of ACL*. 2058–2073.

[63] Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. Summarize: Global Summarization of Medical Dialogue by Exploiting Local Structures. In *Findings of EMNLP*. 3755–3763.

[64] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Trans. Assoc. Comput. Linguistics* 8 (2020), 64–77.

[65] Muhammad Khalifa, Miguel Ballesteros, and Kathleen R. McKeown. 2021. A Bag of Tricks for Dialogue Summarization. In *EMNLP*. 8014–8022.

[66] Seokhwan Kim. 2019. Dynamic memory networks for dialogue topic tracking. (2019).

[67] Seungone Kim, Se June Joo, Hyungjoo Chae, Chaehyeong Kim, Seung won Hwang, and Jinyoung Yeo. 2022. Mind the Gap! Injecting Commonsense Knowledge for Abstractive Dialogue Summarization. In *COLING*. 6285–6300.

[68] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR, Conference Track Proceedings*.

[69] Jia Jin Koay, Alexander Roustai, Xiaojin Dai, Dillon Burns, Alec Kerrigan, and Fei Liu. 2020. How Domain Terminology Affects Meeting Summarization Performance. In *COLING*. 5689–5695.

[70] Jia Jin Koay, Alexander Roustai, Xiaojin Dai, and Fei Liu. 2021. A Sliding-Window Approach to Automatic Creation of Meeting Minutes. In *NAACL-HLT*. 68–75.

[71] Kundan Krishna, Sopan Khosla, Jeffrey P. Bigham, and Zachary C. Lipton. 2021. Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques. In *ACL/IJCNLP, Volume 1: Long Papers*. 4958–4972.

[72] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *EMNLP*. 9332–9346.

[73] Harshit Kumar, Arvind Agarwal, and Sachindra Joshi. 2018. Dialogue-act-driven Conversation Model : An Experimental Study. In *COLING*. 1246–1256.

[74] Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *TACL* 10 (2022), 163–177.

[75] Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In *NAACL-HLT, Volume 2 (Short Papers)*. 687–692.

[76] Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2021. Contrastive Learning with Adversarial Perturbations for Conditional Text Generation. In *ICLR*.

[77] Yuejie Lei, Yuanmeng Yan, Zhiyuan Zeng, Keqing He, Ximing Zhang, and Weiran Xu. 2021. Hierarchical Speaker-Aware Sequence-to-Sequence Model for Dialogue Summarization. In *ICASSP*. 7823–7827.

[78] Yuejie Lei, Fujia Zheng, Yuanmeng Yan, Keqing He, and Weiran Xu. 2021. A Finer-grain Universal Dialogue Semantic Structures based Model For Abstractive Dialogue Summarization. In *Findings of EMNLP*. 1354–1364.

[79] Mirko Lenz, Premtim Sahitaj, Sean Kallenberg, Christopher Coors, Lorik Dumani, Ralf Schenkel, and Ralph Bergmann. 2020. Towards an Argument Mining Pipeline Transforming Texts to Argument Graphs. In *COMMA (Frontiers in Artificial Intelligence and Applications, Vol. 326)*. 263–270.

[80] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*. 7871–7880.

[81] Daniel Li, Thomas Chen, Albert Tung, and Lydia B. Chilton. 2021. Hierarchical Summarization for Longform Spoken Dialog. In *ACM UIST*. 582–597.

[82] Haoran Li, Song Xu, Peng Yuan, Yujia Wang, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. Learn to Copy from the Copying History: Correlational Copy Network for Abstractive Summarization. In *EMNLP*. 4091–4101.

[83] Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization. In *ACL, Volume 1: Long Papers*. 2190–2196.

[84] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP, Volume 1: Long Papers*. 986–995.

[85] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In *Findings of EMNLP*. 1823–1840.

[86] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[87] Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *COLING volume 1*.

[88] Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. CSDS: A Fine-Grained Chinese Dataset for Customer Service Dialogue Summarization. In *EMNLP*. 4436–4451.

[89] Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other Roles Matter! Enhancing Role-Oriented Dialogue Summarization via Role Interactions. In *ACL*. 2545–2558.

[90] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 8693)*. 740–755.

[91] Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *LREC*.

[92] Marina Litvak, Mark Last, and Menahem Friedman. 2010. A new approach to improving multilingual summarization using a genetic algorithm. In *ACL*. 927–936.

[93] Bing Liu and Sahisnu Mazumder. 2021. Lifelong and Continual Learning Dialogue Systems: Learning during Conversation. In *AAAI*. 15058–15063.

[94] Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019. Automatic Dialogue Summary Generation for Customer Service. In *ACM SIGKDD*. 1957–1965.

[95] Fei Liu, Feifan Liu, and Yang Liu. 2010. A supervised framework for keyword extraction from meeting transcripts. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 3, 538–548.

[96] Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021. Topic-Aware Contrastive Learning for Abstractive Dialogue Summarization. In *EMNLP*. 1229–1243.

[97] Qian Liu, Bei Chen, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang. 2020. Incomplete Utterance Rewriting as Semantic Segmentation. In *EMNLP*. 2846–2857.

[98] Yizhu Liu, Qi Jia, and Kenny Zhu. 2022. Reference-free Summarization Evaluation via Semantic Correlation and Compression Ratio. In *NAACL*. 2109–2115.

[99] Yizhu Liu, Qi Jia, and Kenny Q. Zhu. 2021. Keyword-aware Abstractive Summarization by Extracting Set-level Intermediate Summaries. In *WWW*. 3042–3054.

[100] Yongtai Liu, Joshua Maynez, Gonçalo Simões, and Shashi Narayan. 2022. Data Augmentation for Low-Resource Dialogue Summarization. In *Findings of NAACL*. 703–710.

[101] Zhengyuan Liu and Nancy Chen. 2021. Controllable Neural Dialogue Summarization with Personal Named Entity Planning. In *EMNLP*. 92–106.

[102] Zhengyuan Liu and Nancy Chen. 2022. Entity-based De-noising Modeling for Controllable Dialogue Summarization. In *SIGdial*. 407–418.

[103] Zhengyuan Liu and Nancy F. Chen. 2022. Dynamic Sliding Window Modeling for Abstractive Meeting Summarization. In *Interspeech 2022*. 5150–5154.

[104] Zhengyuan Liu and Nancy F Chen. 2023. Picking the Underused Heads: A Network Pruning Perspective of Attention Head Selection for Fusing Dialogue Coreference Information. In *ICASSP*. 1–5.

[105] Zhengyuan Liu, Angela Ng, Sheldon Lee Shao Guang, Ai Ti Aw, and Nancy F. Chen. 2019. Topic-Aware Pointer-Generator Networks for Summarizing Spoken Conversations. In *IEEE ASRU*. 814–821.

[106] Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. Coreference-Aware Dialogue Summarization. In *SIGdial*. 509–519.

[107] Vanessa Loza, Shibamouli Lahiri, Rada Mihalcea, and Po-Hsiang Lai. 2014. Building a Dataset for Summarization and Keyword Extraction from Emails. In *LREC*. 2441–2446.

[108] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *AAAI*. 6818–6825.

[109] Valentin Malykh, Konstantin Chernis, Ekaterina Artemova, and Irina Piontkovskaya. 2020. SumTitles: a Summarization Dataset with Low Extractiveness. In *COLING*. 5718–5730.

[110] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *ACL*. 1906–1919.

[111] Yashar Mehdad, Giuseppe Carenini, Frank Wm. Tompa, and Raymond T. Ng. 2013. Abstractive Meeting Summarization with Entailment and Fusion. In *ENLG*. 136–146.

[112] Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle Lee, Anish Acharya, and Rajiv Ratn Shah. 2021. GupShup: An Annotated Corpus for Abstractive Summarization of Open-Domain Code-Switched Conversations. (2021). arXiv:arXiv:2104.08578

[113] Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering Discrete Latent Topics with Neural Variational Inference. In *ICML (Proceedings of Machine Learning Research, Vol. 70)*. 2410–2419.

[114] Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn A. Walker. 2015. Using Summarization to Discover Argument Facets in Online Idealogical Dialog. In *NAACL-HLT*. 430–440.

[115] Sabine Molenaar, Lientje Maas, Verónica Burriel, Fabiano Dalpiaz, and Sjaak Brinkkemper. 2020. Medical Dialogue Summarization for Automated Reporting in Healthcare. In *Advanced Information Systems Engineering Workshops - CAiSE 2020 International Workshops (Lecture Notes in Business Information Processing, Vol. 382)*. 76–88.

[116] Roser Morante and Eduardo Blanco. 2012. *SEM 2012 Shared Task: Resolving the Scope and Focus of Negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*. 265–274.

[117] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *NAACL-HLT*. 839–849.

[118] Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. In *INTERSPEECH*. 593–596.

[119] Varun Nair, Namit Katariya, Xavier Amatriain, Ilya Valmianski, and Anitha Kannan. 2021. Adding more data does not always help: A study in medical conversation summarization with PEGASUS. (2021). arXiv:arXiv:2111.07564

[120] Yuri Nakayama, Tsukasa Shiota, and Kazutaka Shimada. 2021. Corpus construction for topic-based summarization of multi-party conversation. In *International Conference on Asian Language Processing (IALP)*. IEEE, 229–234.

[121] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *SIGNLL*. 280–290.

[122] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *EMNLP*. 1797–1807.

[123] Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005* 101 (2005).

[124] OpenAI. 2022. Online ChatGPT: Optimizing Language Models for Dialogue. *OpenAI Blog* (2022). https://online-chatgpt.com/

[125] Shereen Oraby, Pritam Gundecha, Jalal Mahmud, Mansurul Bhuiyan, and Rama Akkiraju. 2017. " How May I Help You?" Modeling Twitter Customer Service Conversations Using Fine-Grained Dialogue Acts. In *Proceedings of the 22nd international conference on intelligent user interfaces*. 343–355.

[126] Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond T. Ng. 2014. A Template-based Abstractive Meeting Summarization: Leveraging Summary and Source Text Relationships. In *INLG*. 45–53.

[127] Seongmin Park and Jihwa Lee. 2022. Unsupervised Abstractive Dialogue Summarization with Word Graphs and POV Conversion. In *Proceedings of the 2nd Workshop on Deriving Insights from User-Generated Text.* 1–9.

[128] Seongmin Park, Dongchan Shin, and Jihwa Lee. 2022. Leveraging Non-dialogue Summaries for Dialogue Summarization. In *Proceedings of the First Workshop On Transcript Understanding.* 1–7.

[129] George Prodan and Elena Pelican. 2021. Prompt scoring system for dialogue summarization using GPT-3. (2021).

[130] MengNan Qi, Hao Liu, Yuzhuo Fu, and Ting Liu. 2021. Improving Abstractive Dialogue Summarization with Hierarchical Pretraining and Topic Segment. In *Findings of EMNLP.* 1121–1130.

[131] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67.

[132] Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with Dialogue: A Critical Role Dungeons and Dragons Dataset. In *ACL.* 5121–5134.

[133] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP.* 3980–3990.

[134] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *EMNLP.* 379–389.

[135] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. (1978), 7–55.

[136] Evan Sandhaus. 2008. The New York Times annotated corpus. In *Linguistic Data Consortium.*

[137] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL, Volume 1: Long Papers.* 1073–1083.

[138] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI.* 3776–3784.

[139] Guokan Shang, Wensi Ding, Zekun Zhang, Antoine J.-P. Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised Abstractive Meeting Summarization with Multi-Sentence Compression and Budgeted Submodular Maximization. In *ACL,Volume 1: Long Papers.* 664–674.

[140] Liqun Shao, Hao Zhang, Ming Jia, and Jie Wang. 2017. Efficient and effective single-document summarizations and a word-embedding measurement of quality. *arXiv preprint arXiv:1710.00284* (2017).

[141] Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy. 2021. Neural Abstractive Text Summarization with Sequence-to-Sequence Models. *ACM Trans. Data Sci.* 2, 1 (2021), 1:1–1:37.

[142] Zhouxing Shi and Minlie Huang. 2019. A Deep Sequential Model for Discourse Parsing on Multi-Party Dialogues. In *AAAI.* 7007–7014.

[143] Karan Singla, Evgeny A. Stepanov, Ali Orkan Bayer, Giuseppe Carenini, and Giuseppe Riccardi. 2017. Automatic Community Creation for Abstractive Spoken Conversations Summarization. In *Proceedings of the Workshop on New Frontiers in Summarization, NFiS@EMNLP'17.* 43–47.

[144] Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Large-margin learning of submodular summarization models. In *EACL.* 224–233.

[145] Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing Medical Conversations via Identifying Important Utterances. In *COLING.* 717–729.

[146] Robyn Speer and Catherine Havasi. 2012. Representing General Relational Knowledge in ConceptNet 5. In *LREC.* 3679–3686.

[147] Manfred Stede, Stergos D. Afantenos, Andreas Peldszus, Nicholas Asher, and Jérémy Perret. 2016. Parallel Discourse Annotations on a Corpus of Short Texts. In *LREC.*

[148] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *CoRR* cs.CL/0006023 (2000).

[149] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A Challenge Dataset and Models for Dialogue-Based Reading Comprehension. *Trans. Assoc. Comput. Linguistics* 7 (2019), 217–231.

[150] Ayesha Ayub Syed, Ford Lumban Gaol, and Tokuro Matsuo. 2021. A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization. *IEEE Access* 9 (2021), 13248–13265.

[151] Ryuichi Takanobu, Minlie Huang, Zhongzhou Zhao, Feng-Lin Li, Haiqing Chen, Xiaoyan Zhu, and Liqiang Nie. 2018. A Weakly Supervised Method for Topic Segmentation and Labeling in Goal-oriented Dialogues via Reinforcement Learning. In *IJCAI.* 4403–4410.

[152] Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. *arXiv preprint arXiv:2112.08713* (2021).

[153] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of ACL-IJCNLP*. 3450–3466.

[154] Sansiri Tarnpradab, Fei Liu, and Kien A Hua. 2017. Toward extractive summarization of online forum discussions via hierarchical attention networks. In *The Thirtieth International Flairs Conference*.

[155] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.

[156] Naama Tepper, Anat Hashavit, Maya Barnea, Inbal Ronen, and Lior Leiba. 2018. Collabot: Personalized Group Chat Summarization. In *ACM WSDM*. 771–774.

[157] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman , Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

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

[159] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* 30 (2017).

[160] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR, Conference Track Proceedings*.

[161] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer Networks. In *NeurIPS*. 2692–2700.

[162] Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *ACL*. 5008–5020.

[163] Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022. Analyzing and Evaluating Faithfulness in Dialogue Summarization. In *EMNLP*. 4897–4908.

[164] Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. Clidsum: A benchmark dataset for cross-lingual dialogue summarization. *arXiv preprint arXiv:2202.05599* (2022).

[165] Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis. In *NAACL*. 3071–3081.

[166] Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method. In *ACL (Volume 1: Long Papers)*. 8640–8665.

[167] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[168] Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable Abstractive Dialogue Summarization with Sketch Supervision. In *Findings of ACL/IJCNLP*. 5108–5122.

[169] Xue-Feng Xi, Zhou Pi, and Guodong Zhou. 2020. Global Encoding for Long Chinese Text Summarization. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* 19, 6 (2020), 84:1–84:17.

[170] Wen Xiao and Giuseppe Carenini. 2019. Extractive Summarization of Long Documents by Combining Global and Local Context. In *EMNLP-IJCNLP*. 3009–3019.

[171] Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. (2021). arXiv:arXiv:2107.07567

[172] Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021. Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues. In *AAAI*. 14158–14166.

[173] Pranjul Yadav, Michael S. Steinbach, Vipin Kumar, and György J. Simon. 2018. Mining Electronic Health Records (EHRs): A Survey. *ACM Comput. Surv.* 50, 6 (2018), 85:1–85:40.

[174] Takashi Yamamura, Kazutaka Shimada, and Shintaro Kawahara. 2016. The Kyutech corpus and topic segmentation using a combined method. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*. 95–104.

[175] Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. 2022. On the Robustness of Reading Comprehension Models to Entity Renaming. In *NAACL*. 508–520.

[176] Ze Yang, Liran Wang, Zhoujin Tian, Wei Wu, and Zhoujun Li. 2022. TANet: Thread-Aware Pretraining for Abstractive Conversational Summarization. *arXiv preprint arXiv:2204.04504* (2022).

[177] Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-Based Relation Extraction. In *ACL*. 4927–4940.

[178] Lin Yuan and Zhou Yu. 2019. Abstractive Dialog Summarization with Semantic Scaffolds. (2019). arXiv:arXiv:1910.00825

[179] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *NeurIPS* 34 (2021), 27263–27277.

[180] Klaus Zechner. 2002. Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres. *Comput. Linguistics* 28, 4 (2002), 447–485.

[181] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *ICML (Proceedings of Machine Learning Research, Vol. 119)*. 11328–11339.

[182] Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging Pretrained Models for Automatic Summarization of Doctor-Patient Conversations. In *Findings of EMNLP*. 3693–3712.

[183] Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. 2021. EmailSum: Abstractive Email Thread Summarization. In *ACL/IJCNLP, Volume 1: Long Papers*. 6895–6909.

[184] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *ACL, Volume 1: Long Papers*. 2204–2213.

[185] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

[186] Xiyuan Zhang, Chengxi Li, Dian Yu, Samuel Davidson, and Zhou Yu. 2020. Filling Conversation Ellipsis for Better Social Dialog Understanding. In *AAAI*. 9587–9595.

[187] Xinyuan Zhang, Ruiyi Zhang, Manzil Zaheer, and Amr Ahmed. 2021. Unsupervised Abstractive Dialogue Summarization for Tete-a-Tetes. In *AAAI*. 14489–14497.

[188] Yizhe Zhang, Xiang Gao, Sungjin Lee, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Consistent dialogue generation with self-supervised feature learning. (2019). arXiv:arXiv:1903.05759

[189] Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed H Awadallah, Dragomir Radev, and Rui Zhang. 2021. Summˆ N: A Multi-Stage Summarization Framework for Long Input Dialogues and Documents. (2021). arXiv:arXiv:2110.10150

[190] Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir R. Radev. 2021. An Exploratory Study on Long Dialogue Summarization: What Works and What's Next. In *Findings of EMNLP*. 4426–4433.

[191] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *ACL*. 270–278.

[192] Lulu Zhao, Weihao Zeng, Weiran Xu, and Jun Guo. 2021. Give the Truth: Incorporate Semantic Slot into Abstractive Dialogue Summarization. In *Findings of EMNLP*. 2435–2446.

[193] Lulu Zhao, Fujia Zheng, Keqing He, Weihao Zeng, Yuejie Lei, Huixing Jiang, Wei Wu, Weiran Xu, Jun Guo, and Fanyu Meng. 2021. TODSum: Task-Oriented Dialogue Summarization with State Tracking. (2021). arXiv:arXiv:2110.12680

[194] Jiyuan Zheng, Zhou Zhao, Zehan Song, Min Yang, Jun Xiao, and Xiaohui Yan. 2020. Abstractive meeting summarization by hierarchical adaptive segmental network learning with multiple revising steps. *Neurocomputing* 378 (2020), 179–188.

[195] Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization. (2021). arXiv:arXiv:2109.02492

[196] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir R. Radev. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *NAACL-HLT*. 5905–5921.

[197] Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A Large-scale Media Interview Dataset for Dialogue Summarization. In *NAACL-HLT*. 5927–5934.

[198] Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining. In *Findings of EMNLP*. 194–203.

[199] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *IEEE ICCV*. 19–27.

[200] Yicheng Zou, Jun Lin, Lujun Zhao, Yangyang Kang, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Unsupervised Summarization for Chat Logs with Topic-Oriented Ranking and Context-Aware Auto-Encoders. In *AAAI*. 14674–14682.

[201] Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Topic-Oriented Spoken Dialogue Summarization for Customer Service with Saliency-Aware Topic Modeling. In *AAAI*. 14665–14673.

[202] Yicheng Zou, Bolin Zhu, Xingwu Hu, Tao Gui, and Qi Zhang. 2021. Low-Resource Dialogue Summarization with Domain-Agnostic Multi-Source Pretraining. In *EMNLP*. 80–91.

Received 30 January 2022; revised 20 January 2023; accepted 23 August 2023