Are We Summarizing the Right Way? A Survey of Dialogue Summarization Data Sets

Anonymous EMNLP submission

Abstract

Dialogue summarization is a long-standing task in the NLP field, and several datasets with dialogues and associated human-written summaries of different styles exist. However, it is unclear for which type of dialogue which type of summary is most appropriate. For this reason, we apply a linguistic model of dialogue types to derive matching summary items and NLP tasks. This allows us to map existing dialogue summarization datasets into this model and identify gaps and potential directions for future work. As part of this process, we also provide an extensive overview of existing dialogue summarization datasets.

1 Introduction

Dialogue summarization is a long-standing task in the NLP field and has recently gained traction through the emergence of novel datasets (Gliwa et al., 2019; Zhu et al., 2021) and community efforts like the AutoMin[1] shared task or the SummDial@SIGDial 2021 special session[2]. Dialogues can take a wide variety of forms ranging from formal interviews on a specific topic to political debates to informal conversations over the telephone.[3] Therefore, the question what a suitable, appropriate summary of this data type emerges. While abstractive and extractive summaries have emerged as the de-facto standard for summarization of continuous text (either single or multiple documents), the situation is less clear for dialogue summarization: What is a ”proper” summary for the different types of dialogues that exist?

There exist several dialogue corpora with associated human-written summaries. These summaries differ significantly in type, style, and focus, depending on the instructions that were given to the human annotators. It is usually not clear *why* the particular summary type was chosen for the dialogue corpus at hand. In fact, we are not aware of any well-founded theory that answers this questions.

To close this gap, we leverage the well-established linguistic model of dialogue types by Walton and Krabbe (1995) to identify suitable summary items for the different types of dialogues. This results in a combination of linguistically defined dialogue types, their features, and the suitable summary items. We then place into this matrix all existing dialogue datasets with summaries that we are aware of. This allows us to map the available resources and to identify gaps, which opens up directions for future work.

More precisely, this work presents four contributions:

1. A concise presentation of the linguistically grounded classification of dialogue types by Walton and Krabbe (Section 2)
2. A mapping from dialogue types to potential summary items and associated NLP tasks (Table 2). This indicates which summaries would be appropriate for which dialogue type.
3. An overview of all existing data sets for dialogue summarization that we are aware of (Section 3), which will be useful for researchers in the field even independent from the linguistic model.
4. A mapping from the existing data sets to the linguistic model, and an analysis of potential resource gaps (Section 4).

We also present the overview of existing dialogue summarization data sets in a comprehensive tabular overview in Table A in the Appendix.

2 Dialogue Types in Linguistics

The analysis of dialogues within linguistics is mainly investigated in the fields of conversation

[1] https://elitr.github.io/automatic-minuting/index.html

[2] https://elitr.github.io/automatic-minuting/summdial.html

[3] In this paper we focus on *spontaneous spoken dialogues*, leaving out written dialogues such as Twitter discussion, scripted dialogues, which occur for example in movies, and the summarization of material spoken by single persons only.

analysis and pragmatics. A large body of work investigates speech acts (Searle, 1969; Grice, 1975, inter alia), i.e. dialogues are decomposed into individual turns and their communicative intents are analysed. A smaller body of work focuses on establishing a typology of dialogues. Among these works, Walton and Krabbe (1995) is a well-established model that is often cited and discussed.

2.1 The Walton & Krabbe Model

Walton and Krabbe developed a model of dialogues types and their features which is often picked up in subsequent work in various fields. Table 1 (upper part) shows the model. It features six basic dialogue types: *Persuasion*, *Negotiation*, *Inquiry*, *Deliberation*, *Information-seeking*, and *Eristics*. There are three additional mixed types: *Debate* (*Persuasion* and *Eristics*), *Committee meeting* (mainly *Deliberation*), and *Socratic Dialogue* (mainly *Persuasion*).[4]

Recently, Macagno and Bigi (2018) showed how Walton and Krabbe's model is connected to theories of speech acts, dialect acts, and pragmatic acts and concepts such as "communicative intentions". Walton and Krabbe's dialogue types were explored by research in multi-agent communication in computer science. For example, Reed (1998) applies the model to derive dialogue frames to describe multi-agent interactions.

A related approach to dialogue type categorization is presented in Franke (2010, 2011). The approach develops a taxonomy of (minimal) dialogues. Minimal dialogues are sequences of speech acts in a dialogue that have ended in a conclusion or decision. A naturally-occurring dialogue is then modelled as a sequence of these minimal dialogues.

It is noteworthy that naturally-occurring dialogues are seen as a mixture of multiple dialogue types in both aforementioned models. Still, we find that most of the dialogue corpora we examine in Section 3 can be assigned to one (or two) main dialogue type(s) under the Walton and Krabbe model.

We choose the Walton and Krabbe model as the basis of our analysis of resources in the dialogue summarization space as it is generally the most established one and has been shown to extend well into other domains. Furthermore, the features that Walton and Krabbe attribute to the dialogue types enable us to infer desiderata for the type of summary that suits the dialogue type. For example, if the the main goal of a negotiation is "making a deal", then a suitable summary would present the deal that resulted from the negotiation. Similarly, if the main goal of a debate is "accommodating conflicting points of views", then a suitable summary would list these points of view, and by extension, attribute them to the speakers participating in the debate, and, going further, provide insight into the reasoning of the speakers etc. Finally, the generation of the desirable summary types can then be decomposed naturally into well-established NLP tasks such as topic detection, argument mining, and stance detection, etc.

In summary, the Walton and Krabbe model and its features provide a structured perspective on dialogues that lets us identify suitable connections between dialogue types and summary items, and enables us to pin-point NLP tasks that are applicable for accomplishing such summaries.

2.2 Mapping Dialogue Types to Summary Items

Having selected the model of dialogue types by Walton and Krabbe (1995) as the lens through which we wish to explore the resources in the dialogue summarization domain, we first infer desirable properties of summaries for each of the dialogue types. For this purpose, we examine the dialogue types' features (primarily: *Initial situation* and *Main goal*; secondarily: *Participant's aim* and *Side benefits*) to derive items that an optimal summary would contain in this view. To link the desirable summary items to specific NLP tasks, we note down NLP targets that need to be identified and extracted to enable a summarization system to produce the summary items in its outputs.

The lower part of Table 1 presents the result of this process.[5] The summary items are ordered by importance in relation to our prioritization of the dialogue type features (i.e. Main goals are more important than Side benefits). We exemplify our mapping based on the Persuasion dialogue type: The main goal of Persuasion dialogues is to revolve a conflict between multiple speakers. Each participant wants to persuade the others. For a summary,

[4] We omit the mixed dialogue types in Table 1 for brevity, as they are combinations of the other types.

[5] To encourage different takes in this mapping process, the authors of this paper individually performed the task of mapping dialogue types to summary items and NLP tasks and then held a discussion to harmonize the mappings. Overall, the mappings of the authors overlapped to a large extent and complemented each other, i.e., no conflicting points or disagreement emerged

	Persuasion	**Negotiation**	**Inquiry**	**Deliberation**	**Information-seeking**	**Eristics**
Initial situation	Conflicting points of view (POVs)	Conflict of interests & need for cooperation	General ignorance	Need for action	Personal Ignorance	Conflict & antagonism
Main goal	Resolution of such conflicts by verbal means	Making a deal	Growth of knowledge & agreement	Reach a decision	Spreading knowledge and revealing positions	Reaching a (provisional) accommodation in a relationship
Participants' aim	Persuade the other(s)	Get the best out of it for oneself	Find a "proof" or destroy one	Influence out-come	Gain, pass on, show, or hide personal knowledge	Strike the other party & win in the eyes of onlookers
Side benefits	Develop and reveal positions, Build up confidence, Influence onlookers, Add to prestige	Agreement, Build up confidence, Reveal position Influence onlookers, Add to prestige	Add to prestige, Gain experience, Raise funds	Agreement, Develop & reveal positions, Add to prestige, Vent emotions	Agreement, Develop & reveal positions, Add to prestige, Vent emotions	Agreement, Develop & reveal positions, Gain experience, Amusement, Add to prestige, Vent emotions
Summary items	POVs, Resolutions, Disagreements, Positions, Arguments, Winners/Losers, Controversies	Final deal, Initial interests, Winners/Losers, Evolution of deal, Arguments	Initial inquiry, Gained/new knowledge, Reached agreement, (Line of) arguments, Mentioned facts	Decision, Initial need for action, Positions of speakers, Evolution of decision, Winners/Losers, Emotions	Initial problem, Solution, Positions, Emotions	Initial conflict, Resolution/agreement, Winners/Losers, Arguments, Emotions
NLP targets	Topics, Stances, Decisions, Arguments, Emotions, Sentiment	Decisions, Stances, Topic tracking, arguments	Topics, Knowledge, Decisions, Arguments, Keyfacts	Decisions, Topics, Stances, Arguments, Topic tracking, Emotions	Topics, Action items, Decisions, Stances, Emotions	Topics, Action items, Decisions, Arguments, Emotions

Table 1: Categorization of dialogue types (columns) and their features (rows) according to Walton and Krabbe (1995), and their mapping to our proposed summary items (sorted by importance) and the applicable NLP tasks' target information.

we are mainly interested in the different conflicting points of views (POV) and the resolution of the disagreement. However, the arguments used to resolve the conflict, and the final "winner" are also of interest. For each of these summery items, a corresponding NLP task can be used to extract a specific item. For instance, to extract the different POVs, stance detection can be applied. To extract the arguments used to persuade others, argument detection is applicable etc. That is, summaries of a dialogue under a given dialogue type would ideally include these targets explicitly in a structured manner to facilitate the creation and evaluation of automatic summarization systems.

The list of all NLP targets emerging in the mapping are: *Topics (tracking)*, *Decisions/Action items*, *Arguments*, *Emotions/Sentiment*, *Stances*, *Keyfacts*, and *Knowledge*. We will apply this inventory of NLP targets in Section 4 to map out existing resources and investigate which summary items have been explored for which dialogue types.

3 Data Sets – An Overview

We next provide an overview of existing dialogue summarization datasets. The overview is complemented by Table A in the Appendix which offers a compact and comprehensive outline of the data sets including descriptions, sizes, covered languages, and available summary types. We divide the datasets into the domains that they cover (*Meetings, Broadcast Interviews, Customer and Patient Support, Spontaneous Conversation*) and discuss applicable dialogue types.

Dialogues can be either spoken or written. While several corpora of written or more formal dialogues and their summarization have emerged recently (Gliwa et al., 2019; Chen et al., 2021, inter alia), we here focus on corpora for summarization of (transcripts of) spoken dialogues, which is considerably different than summarization of text, as described for example in Gurevych and Strube (2004).

Work on summarizing spoken dialogues (i.e. involving more than one speakerstarted in the late 1990s and early 2000s (see for example (Zechner and Waibel, 2000b,a)). These already covered a great variety of different types of dialogues, such as TV discussions (NewsHour, CNN CrossFire), phone calls (CALLHOME, CALLFRIEND) and meetings. An overview of these early approaches into summarizing dialogues can be found in Zechner (2002).

At the same time, the VERBMOBIL project, which focused on negotiations dialogues, also worked on summarising these (Reithinger et al., 2000; Alexandersson et al., 2000).[6]

3.1 Meetings

The topic of summarizing meetings gained considerable attraction with extensive work on the ICSI-Corpus (Morgan et al., 2001) and the AMI-Corpus (Murray et al., 2007, e.g.). Murray et al. (2005) presented work on manually summarizing the ICSI meetings, where annotators were instructed to "construct a textual summary [. . .] aimed at someone who is interested in the research being carried out". Four headlines or questions served as guidelines: 1) Why are they meeting and what do they talk about? 2) Decisions made by the group, 3) progress and achievements and 4) problems described. Liu and Liu (2008) extended this work by creating more human summaries and evaluating the summaries based on a questionnaire to be filled out by humans. Other work looked in more detail into how to detect and summarize action items, their descriptions and their appropriate time frames (Purver et al., 2007, e.g.).

The AMI corpus was also extensively studied in the context of summarization. However, while the ICSI corpus contains actual meetings of the participating research groups, which had a varied number of participants, the AMI corpus contains meetings of four persons with different roles in a product design scenario, which was not a natural scenario for the participants. Additionally, the topic is always the same, whereas the ICSI corpus has a wide variety of topics that were discussed in the meetings, including for example chit-chat among team members waiting for everyone to arrive. Summaries for the AMI corpus were created in an abstractive way, based on dialogue acts supporting the information in the summaries (Murray et al., 2007).

Fernández et al. (2008) aimed at identifying "decision-making sub-dialogues" in the AMI meeting data. The authors state that a decision subdialogue consists of three components: a) an issue raised, b) proposals are considered and c) the decision. To that end, they annotate dialogue acts in the data that represent either the issue, or parts of the resolution and the decision.

Similar to the development in the text summarization domain, the dialogue summarization domain moved to using queries to represent the information need of a specific user (Mehdad et al., 2014). Unfortunately, there was not data created for this scenario and the qualitative evaluation was performed on a small subset of the data.

Wang and Cardie (2012) and Wang and Cardie (2013) also work on summarizing meetings, but rather than aiming for a generic summary, they present work on summarizing focused summaries, that are based on specific aspects of a meeting, such as decisions, action items etc.

Following in the footsteps of the AMI corpus Yamamura et al. (2016) present a similar dataset for the Japanese language named "Kyutech Corpus", which also includes reference summaries created in the same fashion as the reference summaries for the AMI corpus.

More recently, Zhong et al. (2021) used queries to represent information need when accessing the ICSI and AMI corpora.

Another type of meeting dialogues occur in the political domain. Political debates from the UK's House of Commons have been used by Vilares and He (2017). The authors aim to produce summaries which give a brief overview on the main viewpoints exchanged and perspectives expressed, which puts it in the area of stance classification and argument mining.

Committee meetings form the Welsh and Canadian Parliament are used by Zhong et al. (2021). Their aim is to create informative summaries based on two types of queries: General queries and specific queries, which included discussion points, opinions, ideas etc. In the discussions elements relevant to the queries have been annotated, as well as informative summaries created.

[6] Note that in the following we do not present all existing work in the domain of dialogue summarization, but focus on those that present representative research results or annotations.

Dialogue Types The discussed corpora in the meeting domain mainly cover project, team, and committee meetings. Given the Initial situation settings of *Need for action, conflict of interest & need for cooperation*, and the Main goals *Reach a decision, Making a deal*, we assign this domain to the dialogue types *Deliberation* and *Negotiation*.

3.2 Broadcast Interviews

TV discussions were already studied in the early phases of speech summarization. More recent work is presented by Zhu et al. (2021) based on NPR and CNN interviews. Reference summaries are based on the descriptions of the interviews and the list of topics discussed.

Podcasts are another form of exchange, that can be an interview, but it can also be a discussion. Clifton et al. (2020) present a data set of Podcasts used for summarization. Reference summaries are based on creator-generated descriptions, which are most likely rather indicative than informative. Using generic summarization algorithms, summaries are created automatically and evaluated manually.

Dialogue Types While the formats covered in the corpora in this domain are rather open by nature, we map it to the dialogue types *Information-seeking*, e.g. interviews with an experts where ignorance (Initial situation) is remedied by the expert's knowledge (Main goal), and *Debate*, where the Initial situation is the presence of conflicting views that are accommodated and discussed in front of an audience (Main goal).

3.3 Customer and Patient Support

Early work in dialogue summarization also includes call-center dialogues. Higashinaka et al. (2010) present work in this direction, which is unfortunately not based on actual call-center dialogues, but rather on recordings of people who were assigned various roles. Tamura et al. (2011) improved on this by using actual call center data. As the logs available for each dialogue were deemed unsuitable for summarization, two types of summaries were created: 1) Indicative summaries, for agents or managers to grasp the gist of the calls and 2) Informative summaries, that contain the content and allow managers to get necessary details of the calls.

Favre et al. (2015) also present work on summarizing call center dialogues. The aim is to create synopses of the calls, which contain the problem and the suggested solution. As opposed to most other work presented, the data set covered not only English, but French (Decoda Corpus) and Italian (Luna Corpus). Based on the same data sets Danieli et al. (2016) looks into analysing the behavior shown in the conversation, which is an important aspect for quality assurance supervisors.

Liu et al. (2019) present work on the DiDi-corpus, which contains dialogues from customer service centers and summaries created by the respective agents. Their aim is to identify key-point sequences in the dialogues, to which end they devise a tagging system with 51 labels, ranging from "Question Description" to "Solution".

Zhao et al. (2020) present work on the Automobile Master Corpus, which contains data from a customer question and answer scenario. It is unclear what the summaries are aimed at, so we have to assume that they are generic summaries.

Various data sets have been used for summarization that come from the medical domain. Acharya et al. (2019) present work on a data set where patients with a specific condition are interviewed. As the data contains actual interviews it cannot be shared. The summaries created aim to include sentences that motivate patients to get better.

Joshi et al. (2020) and Yim and Yetisgen (2021) work on a data set of medical interviews where reference summaries are created by medical doctors, instructing them to summarize as they would for a "clinical note by including all the relevant information". A specific focus was put on negative utterances such as "does not have symptom X".

Dialogue Types This domain clearly evolves around the need for specific information exchange (Initial setting) and passing knowledge between the speakers (Main goal). We thus assign it the *Information-seeking* dialogue type.

3.4 Spontaneous Conversations

Spontaneous or rather informal conversations were already part of the early work presented by Zechner and Waibel (2000b) and Zechner and Waibel (2000a), which looked at the CALLHOME and CALLFRIEND data, which consists of telephone conversations.

A similar setting is the basis for the Switchboard Corpus, which also contains telephone conversations on specific topics. Gurevych and Strube (2004) required annotators to "select the most important utterances" in a selection of dialogues and formed two types of gold standard: One based on all three annotators and one based on annotations by at least two annotators.

A more recent type of informal dialogues has been presented by Rameshkumar and Bailey (2020) which contains dialogues in the context of pen and paper role-playing games (CD3 data set). Summaries are provided through a wiki and are produced by fans of the associated show.

Dialogue Types This domain is difficult to assert in terms of dialogue types as the features Initial situation and Main goal are not clearly identifiable. While speakers were given a specific topic for a conversation in most cases, they were not specifically instructed to converse in a predefined manner. We can hence only speculate on the dialogue types mirrored in these conversations; the conversations would have to be examined individually to determine a sequence of matching dialogue types, which is infeasible in our study.

4 Mapping Data Sets to Summary Items

Given the overview of dialogue summarization resources and their mapping to dialogue types under the linguistic model in the previous section, and the summary items assigned to the dialogue types in Section 2.2, we are now able to tabulate the corpora and the summary items[7] to see what areas in this space are covered and where there are opportunities for future work.

Specifically, we tabulate corpora and the NLP targets that are mapped to the summary items and insert the paper references that cover the summary item for a given corpus. We perform this mapping under the requirement that a resource explicitly annotates a given NLP target in a structured manner. That is, while a general, abstractive, manual summary of a meeting might include e.g. action items or decisions, they might not be marked explicitly as such in the summaries or the underlying transcripts. In such a setting, the resource would not enable the creation of summarization systems that explicitly extract e.g. action items.[8]

Table 2 shows the result of this mapping. A quick glance reveals that only a small portion of the potential NLP targets are explicitly annotated in the summarization resources. The table also shows where efforts to create resources have been focused in the dialogue summarization space: The corpora in all domains mainly offer topics-related summaries. The meetings domain is an exception, where considerable effort has been put into annotating decisions and action items.

5 Discussion

Early approaches to create resources for dialogue summarization in the 2000's were based on spontaneous conversations. Such dialogues are difficult to map to the Walton and Krabbe types, as the features instantiations, such as *Initial Situation* or *Main Goal*, are hard to determine. The diversity of these conversations also makes it difficult to define clear guidelines for creating summaries: Annotators were mostly guided by a somewhat under-specified *relevancy* criterion and were given a length constraint. In regards to the covered summary items, such extractive summaries might contain e.g. decisions and stances etc., however, they are not marked or labeled in the extracted dialogue segments explicitly.

In the Meetings domain, summarization efforts became more specific and a substantial body of work looked into decisions and action items, which resulted in structured datasets for these summary items. For other summary items that the dialogue types Negotiation and Deliberation yield, such as Stances and Arguments, no structured resources exist, however.

Available summaries in the Broadcast domain consist of content description by the authors/creators of the content, i.e. they were not created by researchers for the purpose of dialogue summarization. The descriptions thus rather follow the (potentially commercially-motivated) goal of raising interest in a audience, rather than providing an informative or indicative summary. The communicative intent of such descriptions can therefore be considered to be substantially different from that of research-oriented summarization datasets. Naturally, such content descriptions do not explicitly make available any specific summary items.

In the customer and patient support domain, summarization efforts also leveraged readily available resources such as synopses of call logs or doctor's

[7] We omit the Knowledge summary item, since no resource covers it. However, Knowledge Discovery might be an interesting task in Inquiry dialogues.

[8] However, it is not always straight-forward to apply the NLP targets to resources. For example, in the QMSum corpus, the most important topics are summarized for all dialogues, but the queries that cover decisions are not guaranteed to be present for all dialogues and are not explicitly labeled as being related to decisions.

Corpus	**Topics**	**Decisions / Action items**	**Arguments**	**Emotions / Sentiment**	**Stances**	**Keyfacts**
Meetings Corpora. Dialogue types: Negotiation, Deliberation						
VerbMobil		Reithinger et al. (2000); Alexandersson et al. (2000)				
ICSI	Murray et al. (2005); Wang and Cardie (2013)	Murray et al. (2005); Purver et al. (2007); Wang and Cardie (2013)				
AMI	Murray et al. (2007)	Fernández et al. (2008); Wang and Cardie (2012, 2013)				
Kyutech		Yamamura et al. (2016)				
QMSum	Zhong et al. (2021)	Zhong et al. (2021)			Zhong et al. (2021)	
Broadcast Corpora. Dialogue types: Information-seeking, Debate						
MediaSum	Zhu et al. (2021)					
Spotify Podcasts	Clifton et al. (2020)					
Customer & Patient Support Corpora. Dialogue types: Information-seeking						
DiDi	Liu et al. (2019)	Liu et al. (2019)				
Call center I	Higashinaka et al. (2010)					
Call center II	Tamura et al. (2011)					
CCCS	Favre et al. (2015)			Favre et al. (2015)		
Telemedicine	Joshi et al. (2020)					
Clinical Encounter Visits	Yim and Yetisgen (2021)					
Spontaneous Conversation Corpora. Dialogue types: N/A						
Callhome corpus	Zechner and Waibel (2000b)					
(televison shows)	Zechner and Waibel (2000a)					
Switchboard	Gurevych and Strube (2004)					
CRD3	Rameshkumar and Bailey (2020)					

Table 2: Mapping of resource papers to corpora and NLP targets that they cover.

notes as the summarization targets. Here, the goal of summarization efforts can be mainly described as automating the task of manually producing such notes or synopses. Hence, many linguistically motivated summary items that our approach yields for the Information-seeking dialogue type may simply not apply to the particular use cases that are covered by the existing resources, and are thus not marked explicitly as such.

6 Conclusion

We have provided an overview of existing corpora in the domain of spoken dialogue summarization. We found that topic-related extractive or abstractive summaries are predominant, and are often guided by high-level criteria, i.e. summary guidelines ask for content of "high relevancy" to be included without further specifications.

Furthermore, we have applied a linguistically motivated view on dialogues to the available corpora that yields more specific summary items, such as arguments, stances, or emotions. We found that such specific items are scarcely available in a structured manner in existing corpora. As there are several resources available for e.g. argument mining (Lawrence and Reed, 2020) and stance detection (Küçük and Can, 2020) in dialogues, a potential direction for future work could be an effort to bring together such resources.

While our model-driven view on the dialogue summarization space might be insightful and fruitful for future research, it should not be understood in a normative way: it is not intended to point out that certain directions are misguided. For instance, although our mapping does not yield Emotion as a summary item for Negotiation dialogues, there might be relevant use cases for this line of inquiry. Neither does the approach have any claim to completeness in terms of meeting the information need of different users. In this regard, query-based approaches seem to hold a large potential to cover a wide variety of information needs (Zhong et al., 2021). However, since summary items are seamlessly embedded in the natural-language responses in such settings, it is uncertain how well query-based methods are able to generate on-the-fly responses for realistic queries like "what are the action items assigned to me and by when do I have to complete them?". Answering such information needs robustly seems to necessitate that the underlying information is extracted in a structured manner (Purver et al., 2007, e.g.) to be able to generate an appropriate and complete response.

Overall, our analysis indicates that the question of what are appropriate summaries of dialogues is a challenging one, and we have presented a view that offers some answers. While emerging query-based approaches seem to be a fruitful direction due to their potential to cover a high variety of information needs, we believe that linguistic considerations, as those outlined in this work, can also be leveraged to support resource creation efforts in the dialogue summarization space in future work.

References

Sabita Acharya, Barbara Di Eugenio, Andrew Boyd, Richard Cameron, Karen Dunn Lopez, Pamela Martyn-Nemeth, Debaleena Chattopadhyay, Pantea Habibi, Carolyn Dickens, Haleh Vatani, and Amer Ardati. 2019. A quantitative analysis of patients' narratives of heart failure. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 232–238, Stockholm, Sweden. Association for Computational Linguistics.

Jan Alexandersson, Peter Poller, Michael Kipp, and Ralf Engel. 2000. Multilingual summary generation in a speech-to-speech translation system for multilingual dialogues. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 148–155, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Morena Danieli, Balamurali A R, Evgeny Stepanov, Benoit Favre, Frederic Bechet, and Giuseppe Riccardi. 2016. Summarizing behaviours: An experiment on the annotation of call-centre conversations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4430–4433, Portorož, Slovenia. European Language Resources Association (ELRA).

Benoit Favre, Evgeny Stepanov, Jérémy Trione, Frédéric Béchet, and Giuseppe Riccardi. 2015. Call centre conversation summarization: A pilot task at multiling 2015. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 232–236, Prague, Czech Republic. Association for Computational Linguistics.

Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. 2008. Modelling and detecting decisions in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 156–163.

Wilhelm Franke. 2010. Elementare Dialogstrukturen. In *Elementare Dialogstrukturen*. Max Niemeyer Verlag.

Wilhelm Franke. 2011. Taxonomie der Dialogtypen. In *Sprachtheorie, Pragmatik, Interdisziplinäres*, pages 213–222. Max Niemeyer Verlag.

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*, pages 70–79.

H Paul Grice. 1975. Speech acts. *Syntax and semantics*, 3:41–58.

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

Ryuichiro Higashinaka, Yasuhiro Minami, Hitoshi Nishikawa, Kohji Dohsaka, Toyomi Meguro, Satoshi Takahashi, and Genichiro Kikui. 2010. Learning to model domain-specific utterance sequences for extractive summarization of contact center dialogues. In *Coling 2010: Posters*, pages 400–408, Beijing, China. Coling 2010 Organizing Committee.

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

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1957–1965.

Feifan Liu and Yang Liu. 2008. Correlation between ROUGE and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, Short Papers*, pages 201–204, Columbus, Ohio. Association for Computational Linguistics.

Fabrizio Macagno and Sarah Bigi. 2018. Types of dialogue and pragmatic ambiguity. In *Argumentation and Language—Linguistic, Cognitive and Discursive Explorations*, pages 191–218. Springer.

Yashar Mehdad, Giuseppe Carenini, and Raymond T. Ng. 2014. Abstractive summarization of spoken and written conversations based on phrasal queries. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1220–1230, Baltimore, Maryland. Association for Computational Linguistics.

Nelson Morgan, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Adam Janin, Thilo Pfau, Elizabeth Shriberg, and Andreas Stolcke. 2001. The meeting project at ICSI. In *Proceedings of the First International Conference on Human Language Technology Research*.

Gabriel Murray, Pei-Yun Hsueh, Simon Tucker, Jonathan Kilgour, Jean Carletta, Johanna D. Moore, and Steve Renals. 2007. Automatic segmentation and summarization of meeting speech. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 9–10, Rochester, New York, USA. Association for Computational Linguistics.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2005. Evaluating automatic summaries of meeting recordings. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 33–40, Ann Arbor, Michigan. Association for Computational Linguistics.

Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 18–25, Antwerp, Belgium. Association for Computational Linguistics.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134, Online. Association for Computational Linguistics.

C Reed. 1998. Dialogue frames in agent communication. In *Proceedings of the 3rd International Conference on Multi Agent Systems*, page 246.

Norbert Reithinger, Michael Kipp, Ralf Engel, and Jan Alexandersson. 2000. Summarizing multilingual spoken negotiation dialogues. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 310–317, Hong Kong. Association for Computational Linguistics.

John R Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Akihiro Tamura, Kai Ishikawa, Masahiro Saikou, and Masaaki Tsuchida. 2011. Extractive summarization method for contact center dialogues based on call logs. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 500–508, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

David Vilares and Yulan He. 2017. Detecting perspectives in political debates. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1582, Copenhagen, Denmark. Association for Computational Linguistics.

Douglas N Walton and Erik CW Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press.

Lu Wang and Claire Cardie. 2012. Focused meeting summarization via unsupervised relation extraction. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 304–313, Seoul, South Korea. Association for Computational Linguistics.

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405, Sofia, Bulgaria. Association for Computational Linguistics.

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)*, pages 95–104, Osaka, Japan. The COLING 2016 Organizing Committee.

Wen-wai Yim and Meliha Yetisgen. 2021. Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20, Online. Association for Computational Linguistics.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

Klaus Zechner and Alex Waibel. 2000a. DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Klaus Zechner and Alex Waibel. 2000b. Minimizing word error rate in textual summaries of spoken language. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

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

A Appendix

Meetings Corpora. Dialogue types: Negotiation, Deliberation

CORPUS	*DESCRIPTION*	*LANG*	*SUMMARY CONTENTS*
VerbMobil	Negotiations in the domains of scheduling, travel planning, and hotel reservations	DE, EN, JP	- Agreements on locations, dates, hotels, trains (Reithinger et al., 2000). - Agreements on scheduling, accommodation, traveling, entertainment. (Alexandersson et al., 2000).
ICSI Corpus	Informal, natural, and even impromptu meetings at ICSI. 38 meetings for a total of 39 hours, transcribed about 12 hours. 237 participants, 49 unique speakers.	EN	- Summaries answering the following questions: Why are they meeting and what do they talk about? Decisions made by the group? Progress and achievements? Problems described (Murray et al., 2005). - Find dialogue acts that relate to action items (descriptions, time frames, owners, agreements) (Purver et al., 2007). - Abstract summarizing each important output for every meeting. Decision and problem summaries are annotated (Wang and Cardie, 2013).
AMI Corpus	100 hours of meeting recordings.	EN	- Ranking the dialogue acts in terms of being extract-worthy (Murray et al., 2007). - Classify utterances related to decisions: issue (I), resolution (R), and agreement (A). Two authors annotated 9 and 10 dialogues each (Fernández et al., 2008). - An abstract summarizing each decision; dialogue acts that support each decision are annotated (Wang and Cardie, 2012). - Abstract summarizing each important output for every meeting. Decision and problem summaries are annotated (Wang and Cardie, 2013).
Kyutech Corpus	A decision-making task in a virtual shopping mall in a virtual city. 9 conversations.	JP	- Abstractive manual summaries as in the AMI corpus (Yamamura et al., 2016).
QMSum	AMI, ICSI, and 25 committee meetings of the Welsh Parliament and 11 from the Parliament of Canada	EN	- Select and summarize relevant spans of meetings in response to a query (Zhong et al., 2021).
AutoMin	Technical meetings and parliamentary proceedings.	EN, CZ	- Meeting minutes (paper in print; `https://elitr.github.io/automatic-minuting/index.html`)
Broadcast Corpora. Dialogue types: Information-seeking, Debate			
MediaSum	Interview transcripts from NPR and CNN. 49.4K NPR transcripts and 414.2K from CNN.	EN	- Topic descriptions as summaries (Zhu et al., 2021).
Spotify Podcast Dataset	100,000 podcast episodes, comprising $\sim$ 60,000 hours of speech.	EN	- Creator-generated descriptions as reference summaries (Clifton et al., 2020).
Customer & Patient Support Corpora. Dialogue types: Information-seeking			
DiDi	Logs in the DiDi (mobile transportation platform) customer service center.	EN	- Abstractive summaries written by agents. $\sim$300k pairs of dialogues and summaries. "Key point sequences", i.e. a set of 51 a set action/decision items are also annotated (Liu et al., 2019).
Call center I	Simulated contact center dialogues in six domains. 15–20 tasks per domain. $\sim$700 dialogues.	JP	- Scenario texts used as reference data (Higashinaka et al., 2010).
Call center II	4,596 call logs from a Japanese contact center.	JP	- 1. Indicative Summary: Extract utterances to grasp the gist of calls. 2. Informative Summary: Utterances to grasp the details of calls (Tamura et al., 2011).
CCCS	Conversations from the Decoda and Luna corpora of French and Italian call centre recordings. Recordings duration from a few to 15 minutes. 100 conversations in EN, FR each, translated to EN.	FR, IT, EN	- Abstractive summaries about the main events of the conversations, such as the objective of the caller, whether and how it was solved by the agent, and the attitude of both parties. Synopses written by quality assurance experts from call centres (Favre et al., 2015).

Telemedicine	25,000 conversations from a telemedicine platform.	EN	- Medical doctors were asked to summarize the sections of 3000 snippets as they would for a typical clinical note by including all the relevant information (Joshi et al., 2020).
Clinical Encounter Visits	Audio and clinical notes from clinical encounter visits from 500 visits and 13 providers.	EN	- Clinical notes as summary of the patient visit (Yim and Yetisgen, 2021).
Spontaneous Conversation Corpora. Dialogue types: N/A			
Callhome corpus	Spontaneous telephone conversations.	EN, ES	- For 9 English and 14 Spanish dialogues, the most relevant turns were marked (Zechner and Waibel, 2000b).
Televison shows	Four audio excerpts from four television shows.	EN	- Most relevant, meaningful, concise, and informative phrases (Zechner and Waibel, 2000a).
Switchboard	Telephone conversations of at least 10 minutes duration on a given topic. ~2000 turns.	EN	- 10% of all utterances in the dialogue marked as being relevant (Gurevych and Strube, 2004).
DialogSum	Combination of English learner corpora and dialogue understanding datasets. 13,460 dialogues.	EN	- (1) convey the most salient information; (2) be brief (no longer than 20% of the conversation); (3) preserve important named entities within the conversation; (4) be written from an observer perspective; (5) be written in formal language (Chen et al., 2021).
CRD3	Transcripts of Dungeons and Dragons role-playing game. 398,682 turns.	EN	- Multiple summaries available, e.g. an abstract of the resulting plot/narrative of a game. Includes abstractive summaries collected from the Fandom wiki (Rameshkumar and Bailey, 2020).

Table 3: Overview of existing dialogue summarization datasets. The last column lists papers that provide manually created summaries for a given corpus.