Abstractive Meeting Summarization: A Survey

Virgile Rennard[1,2]*, **Guokan Shang**[1]*, **Julie Hunter**[1]†, **Michalis Vazirgiannis**[2]
[1]Linagora, France [2]École Polytechnique, France
virgile@rennard.org guokan.shang@polytechnique.edu
jhunter@linagora.com mvazirg@lix.polytechnique.fr

Abstract

A system that could reliably identify and sum up the most important points of a conversation would be valuable in a wide variety of real-world contexts, from business meetings to medical consultations to customer service calls. Recent advances in deep learning, and especially the invention of encoder-decoder architectures, has significantly improved language generation systems, opening the door to improved forms of *abstractive* summarization—a form of summarization particularly well-suited for multi-party conversation. In this paper, we provide an overview of the challenges raised by the task of abstractive meeting summarization and of the data sets, models, and evaluation metrics that have been used to tackle the problems.

1 Introduction

Being a primary and inevitable means of information exchange at the workplace, a vast amount of time and organizational resources are allocated to meetings (Mroz et al., 2018). Rogelberg et al. (2007) reported that on average, American employees and managers put 6 and 23 hours per week, respectively, into meetings. The rise in videoconferencing linked to the COVID-19 pandemic has only made the situation more severe (Kost, 2020): People are having longer and more frequent meetings, leading to increased fatigue and less time to digest the information exchanged (Fauville et al., 2021). In this context, developing a system that could reliably identify key information from a meeting or meeting transcript and use it to produce a condensed and easily digestible summary—as well as, perhaps, a set of meeting minutes that details decisions and action items—is becoming more of a priority than ever before (Edmunds and Morris, 2000; Elciyar, 2021).

Research on automatic summarization dates as far back as the 1950s, with systems aiming to generate abstracts from scientific literature (Luhn, 1958). Since then, approaches to summarization have developed along two main lines (Gambhir and Gupta, 2017). *Extractive* summarization, in which a system creates a summary by directly lifting important sentences from source documents and concatenating them without any modification to the original sentences, dominated earlier work due to its simplicity. The advent of neural encoder-decoder architectures (Sutskever et al., 2014; Vaswani et al., 2017), however, has opened the door to *abstractive* summarization (See et al., 2017; Lewis et al., 2020), which draws upon deep representations of word or sentence meaning to generate well-written, novel sentences that consolidate and concisely paraphrase information that might be distributed among numerous clauses or sentences in the original text or transcript.

While the majority of research on summarization has been conducted on formal written documents (news, scientific articles, etc.), in this paper, we highlight work on meeting summarization, a subdomain of automatic summarization that stands to benefit particularly well from advances in abstractive methods. Meeting summarization poses challenges not encountered by traditional text summarization (Kryscinski et al., 2019). Some are related to limitations on the tools needed for developing models—namely, training data, model architectures, and evaluation metrics—but others stem from the very nature of the linguistic interactions involved in meetings or multi-party conversation more generally. These challenges suggest that producing commercial-level, automatic tools for abstractive meeting summarization cannot be accomplished simply by generalizing or fine-tuning models trained on text or even certain forms of dialogue, but will require developing radically new approaches tailored to the meeting domain (Zechner, 2002).

*Primary, equal contribution; †significant contributions.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 861–884, 2023. https://doi.org/10.1162/tacl_a_00578
Action Editor: Fei Liu. Submission batch: 9/2022; Revision batch: 2/2023; Published 7/2023.

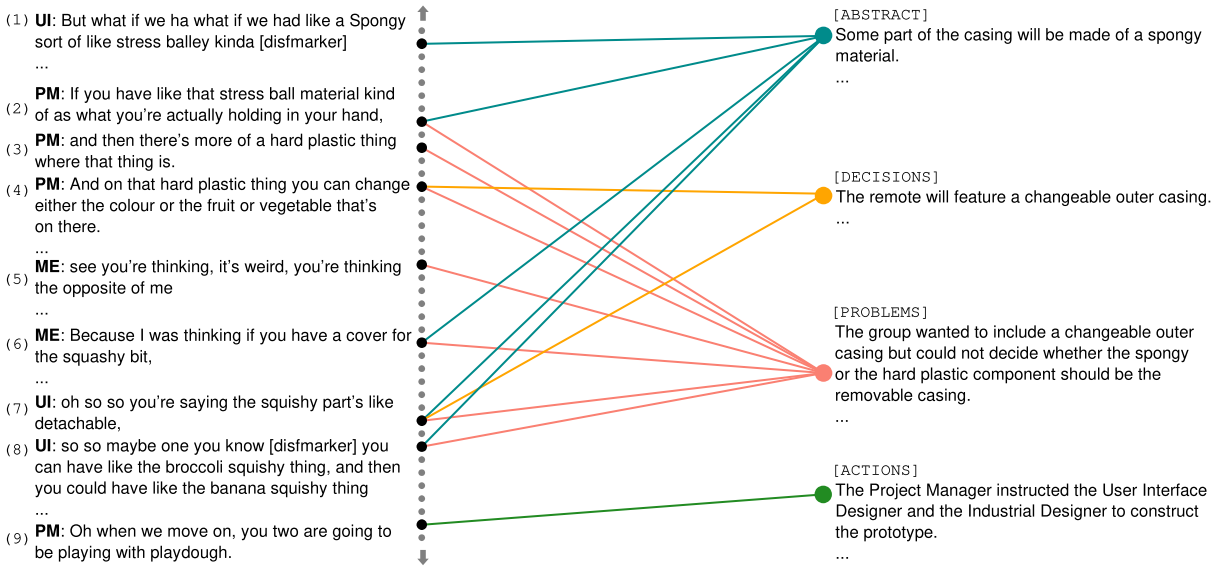

Figure 1: Example (Shang et al., 2020a) of ground truth human annotations from the ES2011c AMI meeting. Successive gray nodes on the left denote utterances recorded in the transcript, the black nodes correspond to a subset of utterances that annotators judged important (summary-worthy). A single sentence from each section of the abstractive summary is shown on the right. The colored arcs link a given sentence from the abstract summary to a single *abstractive* community; that is, the set of clauses from the extractive summary that is summarized by the sentence from the abstractive summary.

We begin in Section 2 with an overview of the particular challenges raised by meeting-style interactions and meeting summaries. Sections 3 through 5 introduce the different tools that have been used to address these challenges, focusing on data sets, summary evaluation, and summarization systems, respectively (Shang, 2021). As we will see, each of these areas introduces complications of its own and we will review the pros and cons of the different approaches. We conclude with a discussion of promising directions for future research.

2 The Challenges of Meeting-style Speech and Summaries

Figure 1 provides a glimpse of the idiosyncratic nature of meeting-style exchanges as well as certain features typical of meeting summaries. The examples are taken from the AMI corpus (Mccowan et al., 2005), a corpus of interactions in which four participants play the roles of project manager (PM), marketing expert (ME), user interface designer (UI), and industrial designer (ID) within a fictitious electronics company. The meetings are centered around a design team whose task is to develop a new television remote control from inception to market, through individual work and a series of group meetings (see Section 3 for discussion). Figure 1 shows excerpts of the human-made extractive (left column) and abstractive (right column) summaries of meeting ES2011c. The colored lines relate each abstractive sentence to the set of extractive sentences—the *abstractive community*—that annotators judged as supporting it.

The Nature of Meeting-style Speech. Conversations in meetings are unedited, often unprepared exercises in real-time collaboration, which tends to yield transcripts with low information density and a fair amount of ‘‘noise’’ stemming from choppy, repetitive, and meandering language. Even though (1)–(9) in Figure 1 are taken from an extractive summary, and thus provide a ‘‘cleaned-up’’ selection of utterances from ES2011c, they illustrate some of the disfluent language—e.g., false starts and repetitions (‘‘what if we ha what if we had’’) and incomplete utterances (‘‘stress balley kinda [disfmarker][1]’’)—and long-winded speech (‘‘so so maybe one you know [disfmarker] you can have like the broccoli squishy thing’’) common in conversational speech.

[1] ‘‘[disfmarker]’’ is short for ‘‘disfluency marker’’ and covers sounds that were not transcribable.

Meeting transcripts also tend to be significantly longer than documents for text summarization, and likewise for their summaries. On average, one AMI transcript contains 4,757 tokens and its summary has 322, while an article from the CNN/DailyMail dataset (Hermann et al., 2015) has an average of 781 tokens and its summary has an average of 56.

Moreover, in contrast to other types of conversation, such as customer service calls and medical appointments, meetings often involve more than two speakers. This increases the number of (potentially conflicting) perspectives to consider (Figure 1, (5)), and also the variety of speech styles and potential for overlapping speech. Finally, it complicates the task of speaker and addressee identification and thus, the conversion of first and second person pronouns into third person identifiers, as is needed for summarization.

The Preference for Abstractive Summarization. Early work on meeting summarization focused on extractive methods (Tur et al., 2009, 2010; Riedhammer et al., 2008; Garg et al., 2009; Tixier et al., 2017). However, while such approaches might be suitable for traditional documents—the LEAD-3 baseline, which takes the first three sentences of a document as its summary, is a good indicator of basic performance (Dohare et al., 2017), for example—humans tend to prefer abstractive summarization for conversation (Murray et al., 2010). (1) to (9) in Figure 1, for instance, provide an extractive alternative to the abstractive sentences on the right, but they are arguably less digestible for a third-party reader than the abstractive alternatives with their concise, third-person perspective on events. Abstractive summarization, however, is inherently more difficult than extraction as it involves content synthesis and language generation in addition to the selection of important material.

Heterogeneous Meeting Formats. While some meetings, as in Figure 1, are more focused on problem solving, other meetings might be more about simply sharing information or brainstorming new ideas. In other words, different meetings may focus on different kinds of information (Nedoluzhko and Bojar, 2019), and there is thus no one-size-fits-all approach to summarization. Detailed minutes (Fernández et al., 2008; Bui et al. 2009; Wang and Cardie, 2011, 2012, 2013) that provide an outline of proposed ideas, supporting arguments, and decisions might be more appropriate for decision-making meetings, while topic summaries might be sufficient for information-sharing meetings, and template-based summaries (Oya et al., 2014), better suited for well-framed meetings with clear agendas. Such diversity means that there may be a need for a variety of automatic systems to allow for tailoring summaries to particular meeting formats.

Subjectivity. Even if we restrict our attention to a single meeting and specify its summary format, there may be a fair amount of subjectivity in the summary content and style, which complicates the task of summary evaluation (see Section 4). Not only are abstractive systems free to reformulate the same content in their own words or style, but also, what counts as summary-worthy can vary from one annotator to the next. In fact, even the same person can produce different results when asked to summarize the same content on two separate occasions (Rath et al., 1961).

3 Meeting Data Sets

Supervised learning, which remains the standard for meeting summarization, generally requires large amounts of training data and yields language and domain-dependent models. Given the idiosyncratic nature of meetings and their summaries, this means that meeting summarization requires meeting-specific data sets. In this section, we first discuss what is required to make an appropriate data set and then introduce extant meeting corpora.

3.1 Data Set Conception and Creation

A data set appropriate for meeting summarization would need to contain spontaneous, spoken conversations in order to capture the unedited, disfluent speech common to meetings. The conversations would need to be relatively long, covering multiple points or topics, like a standard meeting. A good portion of the corpus would need to contain more than two speakers, as many meetings do, because conversational dynamics are affected by participants vying for the floor.

Structured meetings in which participants adhere to clear agendas are easier to summarize, even for humans. At the same time, real-life meetings are not always so well-behaved. Striking a

balance between interactions organized enough for summarization to be feasible while remaining fluid enough to be realistic is important for a model meant to generalize to a real-life setting. This might mean ensuring that participants do not already know each other, in order to minimize small-talk and dependence on background information inaccessible to summarization models. Or it might mean enforcing very clear goals to encourage participants to stay on topic.

Which meeting and summary formats should be covered by the corpus will also need to be determined. And a further consideration is that participants should not broach topics that are too private to be shared outside of their context.

Other concerns arise once the data have been recorded. Automatic speech recognition (ASR) systems can produce transcription errors, for example, that are compounded as we progress through the summarization pipeline, making it risky to skip the laborious task of manual correction. Disfluencies often need to be removed through heavy preprocessing. And in many cases, it is desirable to add additional forms of annotation, such as dialogue act labeling (see Section 5.1.1), which generally require intense effort from trained annotators.

3.2 Extant Data Sets

The significant cost and effort required for producing corpora of meetings and associated summaries or minutes, together with concerns about the privacy of meeting content, mean that there are very few such data sets available. We know of only three for English—AMI (Mccowan et al., 2005), ICSI (Janin et al., 2003), and ELITR (Nedoluzhko et al., 2022)—which together offer around 280 hours of meetings. And only ELITR, which contains roughly 50 hours of meetings in Czech, contains data for a language other than English. We note that all three corpora provide gold transcripts that have been either fully human-produced or human-corrected based on ASR-output to avoid compounded errors from ASR transcripts.

The AMI Corpus (Mccowan et al., 2005), produced as part of the Augmented Multi-party Interaction project, contains 137 scenario-driven meetings (∼65 hours) ranging from 15 to 45 minutes each. As explained in Section 2, corpus participants play the roles of employees in a fictitious electronics company. The role-playing approach helped to produce a data set with well-structured meetings and a standardized summarization pipeline. It also served to minimize concerns about privacy. Note that even though the scenario is artificial, how the participants decide to carry it out is not scripted and so the resulting interactions are spontaneous. That said, the heavily designed nature of the scenarios and the fact that participants in general did not know each other before participating in the corpus arguably led to overly ‘‘well-behaved’’ interactions that risk complicating generalization to real-world contexts. There is also a heavy emphasis on multi-modal interaction (slide presentations, interactions with prototypes) that adds an extra level of complication to the summarization task.

The ICSI Corpus (Janin et al., 2003) consists of 75 naturally occurring meetings (∼72 hours) recorded at the International Computer Science Institute. In each (weekly) meeting of around one hour, members from research groups (including undergraduate and graduate students and professors) discuss specialized and technical topics such as natural language processing, neural theories of language, and ICSI corpus related issues. There are six participants on average per meeting. Because the ICSI meetings contain real interactions with people who know each other and already have projects underway, the interactions are inevitably closer to real-life meetings and have clear goals. On the other hand, they contain considerable technical vocabulary and cover specialized subjects for which participants pull from shared background information that may be inaccessible to summarization algorithms.

The ELITR Corpus (Nedoluzhko et al., 2022) is a recently released corpus of transcripts for 113 technical project meetings in English, and 53 in Czech, totaling over 160 hours of meeting content. Like ICSI, ELITR contains natural, work-based meetings and so the interactions have similar advantages and drawbacks. We note that unlike ICSI and AMI, the original audio recordings for ELITR are not released and sections of certain meetings are censored due to privacy concerns.

Both the AMI and ICSI corpora offer multiple levels of annotation including topic segmentation (see Section 5.2.1) and dialogue act labeling (Section 5.1.1) in addition to extractive summaries, abstractive summaries, and abstractive

communities (see Figure 1). Abstractive summaries in both corpora follow the same structure with four parts: Abstract (high-level description), Decision, Problems, and Actions. Annotators were allowed to make summaries of up to 200 words for each category, but were not obligated to provide summaries for all categories unless they felt it was motivated. Annotators for the ELITR corpus, by contrast, were not provided with a structure for producing minutes. The results can thus vary widely from one annotator to the next, making the corpus a potentially valuable resource for studying subjectivity in minute and summary creation, though this has not yet been explored. We publicly release a preprocessed version of the AMI-ICSI[2] and ELITR[3] corpora including the aforementioned annotations to foster research on this topic.

4 Evaluation Methods

As noted in Section 2, the subjectivity of meeting summaries, stemming in part from the preference for abstractive summarization, complicates the evaluation task. Evaluation requires both verifying that the content in a system-based summary follows from the original transcript and measuring the overlap with a gold-standard summary. However, given the expressive freedom encouraged by abstractive approaches, checking for semantic entailment and overlap of summary-worthy content requires deep semantic understanding, not just recognition that the same words are used.

Unfortunately, the ROUGE metric (Lin, 2004), which remains the standard for both meeting and general text summarization, scores system-produced summaries based purely on surface lexicographic matches with a (usually single) gold summary, making it not ideal for assessing abstractive summaries. If we take the abstractive summary from Figure 1 (''Some part of the casing will be made of a spongy material'') as a gold example, ROUGE would assign a higher score to a system that produces ''Some part of the casing will be made of broccoli'' than one that output ''A portion of the outer layer will be constructed from a sponge-like material,'' even though the latter is a perfect reformulation of the gold summary, while the former says something very different (and false).

A reasonable way to try to improve over ROUGE would be to take advantage of massive, pretrained, contextual word embeddings and a notion of lexical similarity rather than strict lexical overlap. Some recent efforts pursue this direction (Sai et al., 2022), including BERTScore (Zhang et al., 2020a) and MoverScore (Zhao et al., 2019a), which aim to measure the semantic distance between the contextualized mapping of a generated summary and the reference, or BARTScore (Yuan et al., 2021), which calculates the log-likelihood of a summary to have been generated, motivated by the fact that a good summary should have a high probability of being generated from a source text. Building upon these methods, DATScore and FrugalScore (Eddine et al., 2022; Kamal Eddine et al., 2022) incorporate *data augmentation* and *knowledge distillation* techniques to further improve performance and overcome their drawbacks.

Despite their rapid development, recent studies on *meta-evaluation* of these metrics show mixed results. Peyrard (2019) and Bhandari et al. (2020a) compare their performance in the context of document summarization and show that metrics strongly disagree in ranking summaries from any narrow scoring range, e.g., there is no consensus among metrics regarding which summary is better than another in the high scoring range in which modern systems now operate. Bhandari et al. (2020b) argue that there is no one-size-fits-all metric that correlates better with human judgement than the others, and that can outperform others on all datasets.

Clearly, evaluation is itself a very challenging task. And we note that none of these metrics even touches on another central challenge for summary evaluation, namely that of factual consistency. When we summarize a meeting or detail decisions and actions items in our own words, it is important to get the facts straight. Otherwise, the resulting summaries are not reliable for an end user. While we do not know of current work that focuses on evaluation of factuality explicitly for the meeting domain, the study of factual consistency in summarization more generally is a budding research area that we discuss in Section 6.

In the absence of a clear winner for summary evaluation metrics, none of the alternatives has yet to be widely adopted. In our comparison of different summarization systems in Section 5, we

[2] https://github.com/guokan-shang/ami-and-icsi-corpora.

[3] https://github.com/guokan-shang/elitr-minuting-corpus.

therefore stick with ROUGE, which offers the additional advantage of conceptual simplicity and lower computational cost compared to metrics based on contextual embeddings and pretrained language models.

5 Systems for Meeting Summarization

Jones (1999) describes a general summarization pipeline consisting of three stages:

(I) Interpretation: mapping the input text to a *source representation* that adds additional information to the source document that is useful for interpreting its content.

(T) Transformation: transforming the source representation to a *summary representation* based on which the final summary will be produced.

(G) Generation: generating a summary text from the summary representation.

In the course of our literature review, we noted that when extant work on meeting summarization attempts to address one of the challenges described in Section 2, it can be seen as focusing on one of the stages of Jones' pipeline. We therefore adopt this three-stage schema as the backbone for a taxonomy of summarization systems that we lay out in the remainder of this section.

5.1 Interpretation

Consider the following (fabricated) exchange that we might imagine taking place in the AMI meeting from Figure 1:

(1) a. PM: So what color should the removable cover be?

b. ID: I think we should offer a few options. What about raspberry, lime, and blueberry and then black for those who don't like color?

c. UI: Sounds good to me.

d. ME: Me too.

e. PM: OK. Let's go with that.

There is a very clear decision that has been made in (1), but what information allows us to recognize this decision so easily? The acknowledgment in (1-e) explicitly confirms the decision and thus plays a crucial role in inferring that *a* decision has been made, but it does not tell us *what* decision has been made.

In fact, none of the utterances (1-a)–(1-e) alone allow us to infer the decision. We must rather understand that (1-b) provides an *answer to the question* asked in (1-a) and that (1-e) is *an acknowledgement of the suggestion* in (1-b) and of *the positive answers* in (1-c) and (1-d). (If instead of agreeing in (1-c) and (1-d), the UI and ME had presented and defended an alternative proposal, we might have understood (1-e) as an acknowledgement of *their* suggestion rather than of (1-b).)

Because how an utterance contributes to the larger conversational context is often crucial for understanding the conversation as a whole, some summarization approaches, which we review in Section 5.1.1, enrich meeting transcripts with explicit representations of those contributions. Other summarization accounts exploit other types of information relevant to discourse interpretation. HMNet (Zhu et al., 2020) and DDAMS (Feng et al., 2020), for instance, use information about speakers and turns (''who said what''), drawing on the fact that speaker information can help to convert (frequent) occurrences of first and second person pronouns into third person pronouns, as is necessary for summarization (Luo et al., 2009). Further interpretation-focused methods are studied in Section 5.1.2, in which we take a look at accounts that have opted to augment transcripts with *non-linguistic* information of multi-modal nature, such as information about the eye-gaze of conversational participants.

5.1.1 Discursive Information

While most work on abstractive meeting summarization implicitly assumes that conversation is merely a linear sequence of utterances without semantic relations between them, example (1) underscores the importance of semantic relations for conversational understanding. Drawing on similar insights, certain recent approaches exploit independent theories of **discourse structure**, such as Rhetorical Structure Theory (RST; Mann and Thompson, 1987) and Segmented Discourse Representation Theory (SDRT; Asher, 1993; Lascarides and Asher, 2008), to improve summarization. Accounts like RST and SDRT maintain that in a coherent conversation, each (roughly) clause-level unit should be semantically related to some other part of the conversation via a *discourse relation* such as Question-Answer Pair (QAP), Acknowledgement (Ack), Explanation,

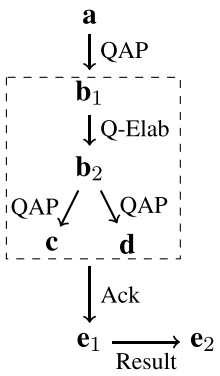

Figure 2: Discourse graph for example (1). Node b_1 represents the first sentence of (1-b), b_2, the second, and similarly for e_1 and e_2. The dashed box indicates that the PM's ''OK'' in (1-e) acknowledges and accepts the entire exchange from (1-b) to (1-d).

Contrast, etc., to reflect its contribution to the larger discourse. Each coherent discourse can thus be represented as a weakly-connected graph whose edges are labeled with discourse relations. Figure 2 illustrates a possible SDRT graph for example (1).

To the best of our knowledge, Feng et al. (2020) is the first work to exploit discourse graphs to generate abstractive meeting summaries. They employ a sequential discourse parser (Shi and Huang, 2019) trained on the STAC corpus of multi-party chats (Asher et al., 2016) to automatically obtain discourse graphs for the AMI and ICSI meeting corpora. Levi graph transformation (Gross and Yellen, 2003) is then used to turn graph edges labeled with discourse relation types into vertices. Their graph-to-sequence model consists of a graph convolutional network encoder (Schlichtkrull et al., 2018) that takes a meeting discourse graph as input and a PGN decoder (See et al., 2017) to generate the final summary. A dialogue-aware data augmentation strategy for constructing pseudo-summaries is introduced to pretrain the model.

An alternative approach to discourse interpretation that developed largely independently of RST and SDRT is **dialogue act** classification. Detailing the differences between the approaches is out of the scope of this paper, but in a nutshell, dialogue acts provide a shallower notion of discourse structure (Jurafsky et al., 1997) in that they do not entail a full graph structure over a conversation. On the other hand, systems for dialogue act labeling, such as DAMSL (Allen and Core, 1997; Core and Allen, 1997) or DiAML (Bunt et al., 2010, 2012), place more emphasis on interactive acts such as *stalling* to hold the floor, *assessing* other discourse moves, *suggesting*, or *informing*.

Both the AMI and ICSI corpora provide gold dialogue act labels, and Goo and Chen (2018) use the labels from the AMI corpus to develop a sentence-gated mechanism that jointly models the relationships between dialogue acts and topic summaries to improve abstractive meeting summarization. In particular, they show that dialogue acts of the type *inform* are more closely linked to summary-worthy material than acts such as *stall* or *assess*. Using LSTM, their model consists of three components enhanced with various attention mechanisms: an utterance encoder, a dialogue act labeler, and a summary decoder.

Dialogue acts have also been used to good effect for summarizing decisions. Fernández et al. (2008) and Bui et al. (2009) first identify relevant areas of decision-related discussion in meetings and classify them as decision-related dialogue acts including the *issue* under discussion, its *resolution*, or *agreement* with the proposed resolution. Then, key fragments of the decision related utterances are retained to form an extractive decision summary. Similar ideas can also be found in the literature on detecting and summarizing action-item-specific dialogue acts in meetings (Purver et al., 2006, 2007).

5.1.2 Multimodality

Meetings tend to take place in shared visual contexts, allowing important information to be conveyed through gestures, facial expressions, head poses, eye gaze, and so on. It is thus reasonable to think that harnessing this information could lead to improved performance over a summarization system that draws on speech transcripts alone. Indeed, Li et al. (2019) has shown that the Visual Focus Of Attention (VFOA), which is estimated based on a participant's head orientation and eye gaze, can help a multi-modal summarization system determine salient utterances, thereby improving the quality of abstractive meeting summaries. In this work, VFOA is integrated with attention mechanisms in a PGN text decoder and topic segmenter trained in a multi-task learning setting.

In the context of extractive meeting summarization, Erol et al. (2003) exploits multimodality to select segments of meeting video recordings in order to facilitate browsing of the videos for

		AMI			ICSI		
		R-1	R-2	R-L	R-1	R-2	R-L
I	Sentence-Gated (Goo and Chen, 2018)	49.29	19.31	24.82	39.37	9.57	17.17
	DDAMS + DDADA (Feng et al., 2020)	53.15	22.32	25.67	40.41	11.02	19.18
	BART + Discourse (Ganesh and Dingliwal, 2019)	35.41	7.24	–	31.84	6.19	–
	TopicSeg + VFOA (Li et al., 2019)	53.29	13.51	26.90	–	–	–
T	Topic + ILP (Banerjee et al., 2015)	–	4.80	–	–	–	–
	Community + MSCG (Mehdad et al., 2013)	32.30	4.80	–	–	–	–
	UNS (Shang et al., 2018)	37.86	7.84	13.72	31.73	5.14	14.50
	PreSeg + POV (Park and Lee, 2022)	33.66	6.85	14.17	27.80	4.56	11.77
	Template-Based (Oya et al., 2014)	31.50	6.70	–	–	–	–
	SOAP-Cluster2Sent T5 + HLSTM (Krishna et al., 2021)	50.52	17.56	24.89	–	–	–
G	SUMM^N (Zhang et al., 2022)	53.44	20.30	51.39*	45.57	11.49	43.32*
	Longformer-BART-arg (Fabbri et al., 2021)	55.27	20.89	24.94	44.51	11.80	19.19
	HMNet (Zhu et al., 2020)	53.02	18.57	24.00	46.28	10.60	18.54
	HAT-CNNDM (Rohde et al., 2021)	52.27	20.15	50.57*	43.98	10.83	41.36*
	HAS-RL (Zhao et al., 2019b)	48.64	17.45	–	–	–	–
	DialogLM (Zhong et al., 2022)	54.49	20.03	51.92*	49.25	12.31	46 .80*

Table 1: Abstractive meeting summarization benchmarks on the AMI and ICSI corpora (some results are taken from Feng et al., 2021a). The * indicates *summary-level* (with sentence split) ROUGE-L score (Lin, 2004).

important information. Along with TF-IDF text analysis of the speech transcript, their system attends to certain visual motion activities (e.g., someone entering the meeting room or standing up to make a presentation) and to acoustic features (e.g., speech volume and degree of interaction) to enhance detection of summary-worthy events. Similar ideas can be found in Xie and Liu (2010), Nihei et al. (2016, 2018), and Nihei and Nakano (2019).

Of course, enhancing meeting transcripts with information to improve interpretation comes at a cost: Most annotations are performed manually and require linguistic expertise to do well, and annotations produced automatically, as in Shi and Huang (2019) or Shang et al. (2020b), can add uncertainties to the data. It is perhaps thus unsurprising that the last couple of years have seen a shift away from Interpretation-focused methods presented in this section towards Generation methods, as shown in Table 1. Still, certain approaches showed promising results, especially Feng et al. (2020) and Li et al. (2019), and, arguably, Interpretation approaches offer other advantages that should encourage us to explore them further. Discourse graphs, for example, can provide input to graph-based networks, helping to bypass limits on input document length imposed by some Generation approaches (see Section 5.3). And a rich source representation can arguably make a single transcript more valuable for training, partially offsetting data scarcity described in Section 3.

5.2 Transformation

Similar in spirit to many standard multi-step, text summarization techniques (Salton et al., 1997) that begin by grouping sentences according to semantic similarity, this category of work focuses on transforming meeting transcripts into intermediate representations that make them easier to summarize. Utterances are first grouped according to various criteria, such as whether they share a topic, contribute to the same field in a template, or respond to the same query. The motivating idea is that first segmenting a transcript into chunks will help to streamline summary generation, as each chunk will be focused on a different aspect of the meeting. These approaches often add an extractive summarization component as well that is applied either before or after chunking to filter out non summary-worthy utterances.

As explained in Section 2, different meetings require different types of summaries; patient doctor consultations, for example, require SOAP notes to be produced to summarize the consultation point by point (Krishna et al., 2021). Having an initial segmentation creates versatile representations that can be used by summarization systems to output domain specific summaries that follow templates or topics. Additionally, while the lengthy nature of meetings is one of the biggest challenges of summarization, segmenting a meeting into multiple, topic focused sub-parts creates units small enough to be easily processed by systems that are computationally expensive.

5.2.1 Topic Segments

Topic segmentation involves dividing text into topically coherent segments of sentences. With the help of a topic segmenter, such as LCSeg (Galley et al., 2003), the work of Banerjee et al. (2015) first separates a meeting transcript into several topic segments. The dependency parse trees of all of the sentences in a topic segment are then merged into a global dependency graph from which the most informative and well-formed sub-graph (i.e., sentence) is selected. Finally, the selected sentences are combined to generate a summary. In the larger dialogue domain, topic segments are also useful for summarizing nurse-to-patient dialogues (Liu et al., 2019) and unstructured daily chats (Chen and Yang, 2020).

5.2.2 Abstractive Communities

Introduced in Murray et al. (2012), *abstractive community detection* is a sub-task of abstractive meeting summarization that seeks to identify the clusters of extractive sentences, or abstractive communities, that together support an individual sentence from an abstractive summary (see Figure 1). Murray et al. (2012) train a logistic regression classifier with handcrafted features to predict if two utterances belong to the same community, then build an utterance graph whose edges represent the binary predictions of the classifier, and finally apply an overlapping community detection algorithm to the graph. Shang et al. (2020a) approach this task by applying Fuzzy c-Means algorithm (Bezdek et al., 1984) on the utterance embeddings learned with a Siamese or triplet network (Chopra et al., 2005; Hoffer and Ailon, 2015).

Mehdad et al. (2013) are the first to propose an abstractive meeting summarization system built on communities. They first extend the method of Murray et al. (2012) by using entailment relations to eliminate less informative utterances from each detected community. Then, the utterances of the same community are represented by a multi-sentence compression graph (MSCG; Filippova, 2010), an unsupervised NLG component to compress a cluster of related, overlapping sentences. The best compression path is extracted as the abstractive summary sentence of the given utterance community. Shang et al. (2018) and Park and Lee (2022) follow the work of Mehdad et al. (2013), introducing various novel components to improve the original summarization process. Abstractive communities of specific types (e.g., `decisions`) have also proven useful in summarizing decisions (Wang and Cardie, 2011, 2012).

5.2.3 Template-related Clusters

In some cases, a meeting summary or the sentences therein are expected to follow a specific template. Oya et al. (2014) introduces the first template-based abstractive meeting summarization system. Using a clustering algorithm and the MSCG, it creates a set of sentence templates generalized from human-authored abstractive sentences, e.g., ‘‘[speaker] discussed [act] and [content]’’. For a given meeting to be summarized, the system first segments the transcription based on topics, then finds the best sentence templates for each segment, and fills the templates to create summaries. The work of Krishna et al. (2021) leverages template structure at the summary-level, e.g., the four parts of an AMI summary. Using a multi-label classifier, the system first extracts noteworthy utterances, while also predicting the section(s) for which they are relevant, i.e., clustering utterances with respect to the in-template sections. Finally, it generates the final summary in a section-by-section way, using only each section’s predicted noteworthy utterances. This approach is also tested on clinical summary generation of doctor-patient conversations.

5.2.4 Query-related Clusters

To deal with the problem of heterogeneous meeting formats as well as the subjectivity inherent in summary-production, Mehdad et al. (2014) proposed a query-based meeting summarization

system that allows for the creation of different summaries for different user interests expressed via queries. Given a phrasal query, the system first extracts the utterances that are most relevant to the keywords of the query and of the main meeting content. Then, these utterances are compressed into summary sentences with an MSCG. Since there were no human-written query-based summaries for AMI at the time, this system was evaluated manually. The QMSum dataset (Zhong et al., 2021) provides multiple query-summary pairs for each meeting in the AMI and ICSI corpora, covering both general and specific points people might be interest in. They also propose a system with two distinct modules: a locator based on pointer network (Vinyals et al., 2015) and a neural seq2seq summarizer. The former is used to locate the query-related utterances in the meeting transcripts, and the latter is meant to summarize selected utterances into summaries.

While versatile in the tasks they could potentially be used to solve, Transformation-focused systems tend to perform worse than Interpretation and Generation-focused systems (cf. Table 1). This can be attributed to the relative age of those methods, as well as their conceptual simplicity that focus on extractive and unsupervised summarization (Shang et al., 2018; Krishna et al., 2021). The spontaneous nature of meetings can lead to meandering structures and information dispersed over a variety of (not necessarily sequential) turns: Participants have side conversations, they forget things and come back, they get interrupted, and so on. All of this makes meeting transcripts harder to accurately segment than traditional documents (Xing and Carenini, 2021). Finding ways to pre-process a text with an eye to creating high-quality topic-based segmentations to feed to Interpretation or Generation-focused systems could be a way to improve the state of the art.

5.3 Generation

Recently, massive pretrained language models based on transformers (Vaswani et al., 2017), such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), and many others (Keskar et al., 2019; Brown et al., 2020; Kamal Eddine et al., 2021), have brought significant improvement to a variety of NLG tasks. Despite their success in dealing with documents, however, these models are inadequate for meeting summarization, if applied off-the-shelf. As explained in Section 2, meeting-style speech has an idiosyncratic nature unlike that of the text data on which pretrained language models are trained. In addition, pretrained language models impose harsh limits on input length that fall far short of the length of an average meeting.

5.3.1 Long Input Processing

A straightforward solution to the length problem is to segment a long document into smaller segments to be processed. Koay et al. (2021) separate ICSI meetings with a sliding window, and then apply BART on each segment to produce smaller summaries, which are concatenated into an overall extractive summary. Zhang et al. (2022) propose SUMM^N, a multi-stage split-then-summarize framework. Within each stage, it first splits the source input into sufficiently short segments. Coarse summaries are then generated for each segment and then concatenated as the input to the next stage. This process is conducted repetitively until a final, fine-grained abstractive meeting summary is produced.

While such segmenting approaches address the length problem, they can lose important cross-partition information (Beltagy et al., 2020), a risk that has led researchers to seek more sophisticated solutions to the length problem.

Long-sequence Transformers. Multiple variants of adapting transformer-based approaches to address the lengthy input problem exist in the literature (Dai et al., 2019; Beltagy et al., 2020; Martins et al., 2022). Longformer (Beltagy et al., 2020), for example, introduces a multi-layer self-attention operation that scales linearly with sequence length, enabling it to process long meeting transcriptions. Although these models are not initially proposed for abstractive meeting summarization, recent benchmarks show their promise (Fabbri et al., 2021) in comparison with certain dedicated systems.

Hierarchical Transformers. Some systems leverage transformers in a hierarchical manner, breaking down a long meeting transcript to multiple relatively shorter sequences of different levels, mirroring the underlying hierarchical structure of text (Yang et al., 2016), i.e., words combine into an utterance, and utterances form

a transcription. The HMNet model, proposed by Zhu et al. (2020), follows a two-level structure. First, each utterance in the meeting is separately encoded by the same word-level transformer encoder, resulting in a sequence of utterance vectors. That sequence is then processed by the turn-level encoder. The transformer decoder makes use of both levels of representation via cross-attention layers. Rohde et al. (2021) propose Hierachical Attention Transformer (HAT). Utterances are first prepended with a special BOS token. Then, after obtaining token-level embeddings with a standard transformer encoder, the BOS token embeddings are fed into an extra layer, yielding sentence-level representations. Finally, the decoder leverages the outputs at both levels to produce a final summary. Similarly, the hierarchical encoder of Zhao et al. (2019b) consists of three levels, sequentially encoding word, utterance, and topic segment embeddings.

5.3.2 Domain Adaptation

Gururangan et al. (2020) have shown that instead of directly using an off-the-shelf language model pretrained on a massive, heterogeneous, and broad-coverage corpus, it is helpful in performance gains to conduct domain-adaptive and task-adaptive pretraining.

Zhong et al. (2022) present the DialogLM model, along with a dialogue-dedicated, window-based denoising pretraining approach. Windows of consecutive utterances are first selected from the conversation, which is then disrupted with arbitrary dialogue-related noises, e.g., speaker mask, turn splitting, turn merging, text infilling, and turn permutation. In the end, the model is trained with the objective of reconstructing the original window based on the perturbed one and the remaining conversation. This approach allows a model to effectively learn dynamic dialogue structures within and surrounding the window. Zou et al. (2021) propose a multi-source pretraining paradigm for low-resource dialogue summarization. They first pretrain the encoder and the decoder separately on the in-domain data, to model the dialogue and summary language respectively, and then pretrain the complete encoder-decoder on the out-of-domain abstractive summarization data using adversarial critics.

Generation-focused systems are based on pretrained language models, which excel at abstractive summarization, and use their extensive pre-training to offset the lack of task specific data. Domain adaptation also helps these models to understand the idiosyncratic style of multi-party speech, albeit to a lesser extent than Interpretation-focused systems. Generation-focused systems, however, lack the flexibility that Interpretation-focused systems have in their output, and when we consider how competitive scores are for Interpretation-focused systems despite their having considerably less training data than Generation-based systems, it is arguable that the latter lack the powerful understanding in the initial representation exhibited by the former. In the following section, we will talk about how some of these shortcomings can be addressed in future work.

6 Future Directions

We conclude by laying out some directions that we think are worth investigation in future work.

Multi-task Learning. Given the variety of annotations that have led to improvements in abstractive meeting summarization, including dialogue act classification, discourse parsing, community detection, multimodal information, and so on, we might expect to see more systems combining these tasks with meeting summarization. Studies on multi-task learning have shown that solving different tasks in a simultaneous fashion often improves learning efficiency and performance, compared to models trained separately (Caruana, 1997), potentially reducing the amount of data needed to solve an individual task. Future work could follow up on previous attempts in this direction (Li et al., 2019; Feng et al., 2021b; Lee et al., 2021).

Factual Consistency. Information inconsistency is a common problem of summarization systems (Kryscinski et al., 2019, 2020), and it is reported that nearly 30% of summaries generated by neural seq2seq models suffer from fact fabrication (Cao et al., 2018). This problem is not only present in the generated summaries, but also in the training data (Guo et al., 2022). For the dialogue domain, Tang et al. (2022) show that most of the factual errors are related to dialogue flow modeling, informal interactions between speakers, and complex coreference resolution.

This suggests future research to focus on dealing with hallucinated content in generated dialogue summaries.

Moreover, specific attention could be addressed to the development of an evaluation metric that estimates factual consistency between a summary and its source, specifically designed for the dialogue domain given its special characteristics (Zechner and Waibel, 2000). Although many metrics have been put in place (Huang et al., 2021), they are inadequate for meeting summarization, as they usually work at the sentence level, which is acceptable for shorter documents but hardly for longer meeting transcripts.

Weak to No Supervision. Current work on meeting summarization focuses mainly on supervised approaches, which makes trained models inevitably language-dependent, and even corpus-dependent. One way to get around the data scarcity problem might be to move to meeting summarization techniques with weaker supervision, such as in a purely unsupervised fashion (Shang et al., 2018; Park and Lee, 2022), or even in a zero-shot way (Ganesh and Dingliwal, 2019). Future approaches could leverage auto-encoders and utterance wide attention to identify the importance of different parts of a meeting. Weak supervision and semi-supervised approaches have yet to be explored in detail.

Spoken Language Models. Over the past few years, pretrained language models have evolved from emerging to mainstream NLP technology. As mentioned in Subsection 5.3, models pretrained on free-form text cannot be directly transferred to treat dialogue transcriptions. Future research could follow the work of, notably, DialoGPT (Zhang et al., 2020b) and DialogLM (Zhong et al., 2022), to develop better language models dedicated to spoken language, which will potentially offer large gains in task performance for abstractive meeting summarization.

Commonsense Incorporation. Some facts are naturally considered as commonsense knowledge for human beings, but are far from obvious for machines, e.g., ‘‘a piano is played by pressing keys’’. The integration of such *a priori* knowledge, usually represented by graphs, into language models would improve their interpretation and reasoning potential (Ilievski et al., 2021), thereby further combining the strength of Interpretation and Generation-focused systems. Commonsense knowledge has already been applied to general dialogue summarization with success (Zhou et al., 2018; Xiachong et al., 2021). For meetings, we believe that its incorporation, much like explicitly flagging jargon terms (Koay et al., 2020), would provide further information that is likely not included in the context. Indeed, considering that meetings take place in professional environments, some background concepts are taken as given by the participants, and integrating this knowledge into models would undoubtedly facilitate dialogue understanding and summary generation.

Prompting Paradigm. The very recent success of zero- and few-shot learning with models like GPT-3 (Brown et al., 2020), Gopher (Rae et al., 2021), and PaLM (Chowdhery et al., 2022) provides a novel paradigm in NLP research. In contrast to the traditional approach of fine-tuning pre-trained language models for tasks using task-specific data sets, this approach yields high performance on a variety of NLP tasks without updating parameters. All you have to do is to give the models task-specific instructions (namely, prompts) represented in natural language, possibly in conjunction with a few demonstrative examples in the context. For example, a prompt for obtaining the summary of an *article* given a budget constraint of N sentences can be formulated as:

‘‘Article: {{article}}’’

‘‘Summarize the above article in N sentences.’’

Recent work shows that these models can be further enhanced by instruction tuning (e.g., T0 [Sanh et al., 2021], FLAN [Wei et al., 2021], and BLOOMZ [Scao et al., 2022]) and learning from human feedback (e.g., InstructGPT/ChatGPT [Ouyang et al., 2022]), in order to be better aligned with inference time usage.

Experiments on the traditional summarization datasets (e.g., CNN/DailyMail) show that these models can produce summaries nearly as good as reference summaries (Stiennon et al., 2020), and humans overwhelmingly prefer the summaries produced under the novel prompting paradigm over the classical pretraining-finetuning one (Goyal et al., 2022). To the best of our knowledge, there has been no attempt to apply it to meeting summarization, but this is an exciting avenue for future research.

Moreover, Goyal et al. (2022) pointed out that this fundamental paradigm shift brings new and serious challenges for evaluation: both existing reference-based and reference-free automatic metrics cannot reliably evaluate zero-shot summaries. Because prompt-based models are not restrictively trained to emulate gold-standard summaries and their style, reference-based metrics that compute lexical or semantic similarity (e.g., ROUGE and BERTScore) give false low scores when comparing generated summaries against available gold summaries. Reference-free metrics (e.g., BLANC [Vasilyev et al., 2020] and FactCC [Kryscinski et al., 2020]) that only rely on the input document also fail, as reference-free metrics roughly score summaries inversely related to abstractiveness, while prompt-based models often generate more abstractive summaries. Therefore, new metrics adapting to this paradigm shift that can evaluate zero-shot summaries are urgently needed.

In the end, as mentioned in Section 2, there is arguably no single ‘‘correct’’ summary for a given input meeting. Different users may choose to focus on different information, choosing different keywords and mentioning different facts. With prompt-based models, the task can be comfortably formulated by creating optimized prompts to meet different needs. Therefore, a study about such system, as well as a comparison with existing query-based meeting summarization systems (see Section 5.2.4) for the same purpose could be an interesting research topic.

7 Conclusion

In this work, we presented a comprehensive overview of the state-of-the-art in abstractive meeting summarization. We discussed the different challenges that research has faced, and we proposed a taxonomy that followed the three-step summarization pipeline, Interpretation-Transformation-Generation, presented by Jones (1999). We further compared the results obtained by previous work and finally laid out promising directions for future research.

Acknowledgments

This work was supported by the SUMM-RE project (ANR-20-CE23-0017). We thank the anonymous reviewers for their feedback.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers. https://www.cs.rochester.edu/research/cisd/resources/damsl

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Dordrecht, Boston, and London: Kluwer. https://doi.org/10.1007/978-94-011-1715-9

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: The STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015. Generating abstractive summaries from meeting transcripts. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, pages 51–60. https://doi.org/10.1145/2682571.2797061

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

James C. Bezdek, Robert Ehrlich, and William Full. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2–3):191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, and Pengfei Liu. 2020a. Metrics also disagree in the low scoring range: Revisiting summarization evaluation metrics. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5702–5711, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.501

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020b. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 9347–9359, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.751

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Trung Bui, Matthew Frampton, John Dowding, and Stanley Peters. 2009. Extracting decisions from multi-party dialogue using directed graphical models and semantic similarity. In *Proceedings of the SIGDIAL 2009 Conference*, pages 235–243, London, UK. Association for Computational Linguistics. https://doi.org/10.3115/1708376.1708410

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta. European Language Resources Association (ELRA).

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75. https://doi.org/10.1023/A:1007379606734

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.336

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Mark G. Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme.

In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, volume 56, pages 28–35. Boston, MA.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Shibhansh Dohare, Harish Karnick, and Vivek Gupta. 2017. Text summarization using abstract meaning representation. *arXiv preprint arXiv:1706.01678*.

Moussa Kamal Eddine, Guokan Shang, and Michalis Vazirgiannis. 2022. Datscore: Evaluating translation with data augmented translations. *EACL 2023 Findings*.

Angela Edmunds and Anne Morris. 2000. The problem of information overload in business organisations: A review of the literature. *International Journal of Information Management*, 20(1):17–28. https://doi.org/10.1016/S0268-4012(99)00051-1

Kemal Elciyar. 2021. Overloading in lockdown: Effects of social, information and communication overloads in covid-19 days. *İnönü Üniversitesi İletişim Fakültesi Elektronik Dergisi (İNİF E-Dergi)*, 6(1):329–342. https://doi.org/10.47107/inifedergi.872896

Berna Erol, Dar-Shyang Lee, and Jonathan Hull. 2003. Multimodal summarization of meeting recordings. In *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, volume 3, pages III–25. IEEE. https://doi.org/10.1109/ICME.2003.1221239

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.535

Geraldine Fauville, Mufan Luo, Anna C. M. Queiroz, Jeremy N. Bailenson, and Jeff Hancock. 2021. Zoom exhaustion & fatigue scale. *Computers in Human Behavior Reports*, 4:100119. https://doi.org/10.1016/j.chbr.2021.100119

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021a. A survey on dialogue summarization: Recent advances and new frontiers. In *Proceeding of the 31th International Joint Conference on Artificial Intelligence (IJCAI 2022)*. https://doi.org/10.24963/ijcai.2022/764

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2020. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*. https://doi.org/10.24963/ijcai.2021/524

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021b. Language model as an annotator: Exploring DialoGPT for dialogue summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1479–1491, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.117

Raquel Fernández, Matthew Frampton, John Dowding, Anish Adukuzhiyil, Patrick Ehlen, and Stanley Peters. 2008. Identifying relevant phrases to summarize decisions in spoken meetings. In *Ninth Annual Conference of the International Speech Communication Association*. https://doi.org/10.21437/Interspeech.2008-17

Katja Filippova. 2010. Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 322–330, Beijing, China. Coling 2010 Organizing Committee.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan. Association for Computational Linguistics. https://doi.org/10.3115/1075096.1075167

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: A survey. *Artificial Intelligence Review*, 47(1):1–66. https://doi.org/10.1007/s10462-016-9475-9

Prakhar Ganesh and Saket Dingliwal. 2019. Restructuring conversations using discourse relations for zero-shot abstractive dialogue summarization. *arXiv preprint arXiv:1902.01615*.

Nikhil Garg, Benoit Favre, Korbinian Reidhammer, and Dilek Hakkani Tür. 2009. ClusterRank: A graph based method for meeting summarization. Technical report, Idiap. https://doi.org/10.21437/Interspeech.2009-456

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

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *arXiv preprint arXiv:2209.12356*.

Jonathan L. Gross and Jay Yellen. 2003. *Hand book of Graph Theory*. CRC Press. https://doi.org/10.1201/9780203490204

Yanzhu Guo, Chloé Clavel, Moussa Kamal Eddine, and Michalis Vazirgiannis. 2022. Questioning the validity of summarization datasets and improving their factual consistency. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5716–5727, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems*, 28.

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer. https://doi.org/10.1007/978-3-319-24261-3_7

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

Filip Ilievski, Pedro Szekely, and Bin Zhang. 2021. Cskg: The commonsense knowledge graph. In *European Semantic Web Conference*, pages 680–696. Springer. https://doi.org/10.1007/978-3-030-77385-4_41

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)*, volume 1, pages I–I. IEEE.

Karen Sparck Jones. 1999. Automatic summarizing: Factors immarizing: Factors and directions. *Advances in Automatic Text Summarization*, page 1.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

Moussa Kamal Eddine, Guokan Shang, Antoine Tixier, and Michalis Vazirgiannis. 2022. FrugalScore: Learning cheaper, lighter and

faster evaluation metrics for automatic text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1305–1318, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.93

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. BARThez: A skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.740

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Jia Jin Koay, Alexander Roustai, Xiaojin Dai, Dillon Burns, Alec Kerrigan, and Fei Liu. 2020. How domain terminology affects meeting summarization performance. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5689–5695, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.499

Jia Jin Koay, Alexander Roustai, Xiaojin Dai, and Fei Liu. 2021. A sliding-window approach to automatic creation of meeting minutes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 68–75, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-srw.10

Danielle Kost. 2020. You're right! You are working longer and attending more meetings. *Harvard Business School Working Knowledge*.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.384

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1051

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.750

Alex Lascarides and Nicholas Asher. 2008. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing Meaning*, pages 87–124. Springer. https://doi.org/10.1007/978-1-4020-5958-2_5

Seolhwa Lee, Kisu Yang, Chanjun Park, João Sedoc, and Heuiseok Lim. 2021. Who speaks like a style of vitamin: Towards syntax-aware dialogue summarization using multi-task learning. *IEEE Access*, 9:168889–168898. https://doi.org/10.1109/ACCESS.2021.3124556

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880,

Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1210

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

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F. Chen. 2019. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821. IEEE.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165. https://doi.org/10.1147/rd.22.0159

Xiaoqiang Luo, Radu Florian, and Todd Ward. 2009. Improving coreference resolution by using conversational metadata. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 201–204, Boulder, Colorado. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A framework for the analysis of texts. Technical report, University of Southern California Marina Del Rey Information Sciences Inst.

Pedro Henrique Martins, Zita Marinho, and Andre Martins. 2022. ∞-former: Infinite memory transformer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5468–5485, Dublin, Ireland. Association for Computational Linguistics.

Iain Mccowan, Jean Carletta, Wessel Kraaij, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Wilfried Post, Dennis Reidsma, and P. Wellner. 2005. The AMI meeting corpus. *International Conference on Methods and Techniques in Behavioral Research*.

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. https://doi.org/10.3115/v1/P14-1115

Yashar Mehdad, Giuseppe Carenini, Frank Tompa, and Raymond T. Ng. 2013. Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 136–146, Sofia, Bulgaria. Association for Computational Linguistics.

Joseph E. Mroz, Joseph A. Allen, Dana C. Verhoeven, and Marissa L. Shuffler. 2018. Do we really need another meeting? The science of workplace meetings. *Current Directions in Psychological Science*, 27(6):484–491. https://doi.org/10.1177/0963721418776307

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010. Generating and validating abstracts of meeting conversations: A user study. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2012. Using the omega index for evaluating abstractive community detection. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 10–18, Montréal, Canada. Association for Computational Linguistics.

Anna Nedoluzhko and Ondrej Bojar. 2019. Towards automatic minuting of the meetings. In *ITAT*, pages 112–119.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR Minuting Corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC-2022)*, Marseille, France. European Language Resources Association (ELRA). In print.

Fumio Nihei and Yukiko I. Nakano. 2019. Exploring methods for predicting important utterances contributing to meeting summarization. *Multimodal Technologies and Interaction*, 3(3):50. https://doi.org/10.3390/mti3030050

Fumio Nihei, Yukiko I. Nakano, and Yutaka Takase. 2016. Meeting extracts for discussion summarization based on multimodal nonverbal information. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 185–192. https://doi.org/10.1145/2993148.2993160

Fumio Nihei, Yukiko I. Nakano, and Yutaka Takase. 2018. Fusing verbal and nonverbal information for extractive meeting summarization. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3279981.3279987

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. 2014. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 45–53, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

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*, pages 1–9, (Hybrid) Dublin, Ireland, and Virtual. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wit-1.1

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1502

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.

Matthew Purver, Patrick Ehlen, and John Niekrasz. 2006. Detecting action items in multi-party meetings: Annotation and initial experiments. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 200–211. Springer. https://doi.org/10.1007/11965152_18

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

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. *Journal of Machine Learning Research*, 21(1).

G. J. Rath, A. Resnick, and T. R. Savage. 1961. The formation of abstracts by the selection of sentences. part i. sentence selection by men and machines. *American Documentation*, 12(2):139–141. https://doi.org/10.1002/asi.5090120210

Korbinian Riedhammer, Dan Gillick, Benoit Favre, and Dilek Hakkani-Tür. 2008. Packing the meeting summarization knapsack. In *Ninth Annual Conference of the*

International Speech Communication Association. https://doi.org/10.21437/Interspeech.2008-604

Steven G. Rogelberg, Cliff Scott, and John Kello. 2007. The science and fiction of meetings. *MIT Sloan Management Review*, 48(2):18–21.

Tobias Rohde, Xiaoxia Wu, and Yinhan Liu. 2021. Hierarchical learning for generation with long source sequences. *arXiv preprint arXiv:2104.07545*.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39. https://doi.org/10.1145/3485766

Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. 1997. Automatic text structuring and summarization. *Information Processing & Management*, 33(2):193–207. https://doi.org/10.1016/S0306-4573(96)00062-3

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *CoRR*, abs/2110.08207.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj,

Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeňek Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Karen Fort, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pámies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer. https://doi.org/10.1007/978-3-319-93417-4_38

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Guokan Shang. 2021. *Spoken Language Understanding for Abstractive Meeting Summarization*. Ph.D. thesis, Institut Polytechnique de Paris.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1062

Guokan Shang, Antoine Tixier, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2020a. Energy-based self-attentive learning of abstractive communities for spoken language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 313–327, Suzhou, China. Association for Computational Linguistics.

Guokan Shang, Antoine Tixier, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2020b. Speaker-change aware CRF for dialogue act classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 450–464, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.40

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7007–7014. https://doi.org/10.1609/aaai.v33i01.33017007

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.415

Antoine Tixier, Polykarpos Meladianos, and Michalis Vazirgiannis. 2017. Combining graph degeneracy and submodularity for unsupervised extractive summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 48–58, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4507

Gokhan Tur, Andreas Stolcke, Lynn Voss, John Dowding, Benoit Favre, Raquel Fernández, Matthew Frampton, Michael Frandsen, Clint Frederickson, Martin Graciarena, Dilek Hakkani-Tur, Donald Kintzing, Kyle Leveque, Shane Mason, John Niekrasz, Stanley Peters, Matthew Purver, Korbinian Riedhammer, Elizabeth Shriberg, and Fan Yang. 2009. The CALO meeting speech recognition and understanding system, pages 69–72. https://doi.org/10.1109/SLT.2008.4777842

Gokhan Tur, Andreas Stolcke, Lynn Voss, Stanley Peters, Dilek Hakkani-Tur, John Dowding, Benoit Favre, Raquel Fernandez, Matthew Frampton, Mike Frandsen, Clint Frederickson, Martin Graciarena, Donald Kintzing, Kyle Leveque, Shane Mason, John Niekrasz, Matthew Purver, Korbinian Riedhammer, Elizabeth Shriberg, Jing Tien, Dimitra Vergyri, and Fan Yang. 2010. The

CALO meeting assistant system. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1601–1611. https://doi.org/10.1109/TASL.2009.2038810

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.eval4nlp-1.2

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in Neural Information Processing Systems*, 28.

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

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.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *ICLR 2022*.

Feng Xiachong, Feng Xiaocheng, and Qin Bing. 2021. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 964–975, Huhhot, China. Chinese Information Processing Society of China. https://doi.org/10.1007/978-3-030-84186-7_9

Shasha Xie and Yang Liu. 2010. Using confusion networks for speech summarization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 46–54, Los Angeles, California. Association for Computational Linguistics.

Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2021)*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1174

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485. https://doi.org/10.1162/089120102762671945

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating text generation with BERT. In *8th International Conference on*

Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.30

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.112

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019a. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1053

Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019b. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference*, pages 3455–3461. https://doi.org/10.1145/3308558.3313619

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. DialogLM: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773. https://doi.org/10.1609/aaai.v36i10.21432

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. https://doi.org/10.18653/v1/2021.naacl-main.472

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629. https://doi.org/10.24963/ijcai.2018/643

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.19

Yicheng Zou, Bolin Zhu, Xingwu Hu, Tao Gui, and Qi Zhang. 2021. Low-resource dialogue summarization with domain-agnostic multi-source pretraining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 80–91, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.7