Report on the SIGDial 2021 Special Session on Summarization of Dialogues and Multi-Party Meetings (SummDial)

Tirthankar Ghosal
ÚFAL, MFF
Charles University, Czech Republic
ghosal@ufal.mff.cuni.cz

Muskaan Singh
ÚFAL, MFF
Charles University, Czech Republic
singh@ufal.mff.cuni.cz

Anja Nedoluzhko
ÚFAL, MFF
Charles University, Czech Republic
nedoluzhko@ufal.mff.cuni.cz

Ondřej Bojar
ÚFAL, MFF
Charles University, Czech Republic
bojar@ufal.mff.cuni.cz

Abstract

The SummDial special session on summarization of dialogues and multi-party meetings was held virtually within the SIGDial 2021 conference on July 29, 2021. SummDial @ SIGDial 2021 aimed to bring together the speech, dialogue, and summarization communities to foster cross-pollination of ideas and fuel the discussions/collaborations to attempt this crucial and timely problem. When the pandemic has restricted most of our in-person interactions, the current scenario has forced people to go virtual, resulting in an information overload from frequent dialogues and meetings in the virtual environment. Summarization could help reduce the cognitive burden on the participants; however, multi-party speech summarization comes with its own set of challenges. The SummDial special session aimed to leverage the community intelligence to find effective solutions while also brainstorming the future of AI interventions in meetings and dialogues. We report the findings of the special session in this article. We organized the SummDial special session under the aegis of the EU-funded H2020 European Live Translator (ELITR) project.[1]

Date: 29 July, 2021.

Website: https://elitr.github.io/automatic-minuting/summdial.html.

1 Introduction

Arguably the most conventional and effective form of communication between humans is a conversation in a natural language. With continued efforts to infuse intelligence in machines and fuel

[1] https://elitr.eu

the larger goal of human-machine interaction, automatically comprehending speech and natural language constitutes a fundamental Speech and Natural Language Processing (SNLP) task.

One helpful indicator if an agent (human or machine) has correctly understood the content is to see how well the agent summarizes it considering several evaluation criteria of summarization (e.g., coverage, conciseness, readability, coherence, grammatical correctness, relevance, significance, etc.). Summarization is a challenging SNLP problem. The task and its evaluation are subjective to the agent, and automatic evaluation measures of summarization are still not reliable [Bhandari et al., 2020; Deutsch and Roth, 2021]. Summarizing speech is more complex than summarizing a textual narrative due to various reasons, including noises, incorrectness of the ASRs, discontinuous or incoherent utterances, etc. [Zechner, 2002b]. The task becomes even more challenging when the discourse is a multi-party dialogue or a meeting with multiple participants.

With a sizeable world's working population going virtual, summarizing multi-party dialogues or meetings would be a handy SNLP application. As a significant workforce is working and collaborating remotely because of the pandemic resulting in frequent meetings and ensuing cognitive overload on the participants, imagine how convenient it would be for the participants to just hover over past calendar invites and get concise summaries of the meeting proceedings (*minutes of the meeting*)? How about automatically minuting a multimodal multi-party meeting and generating a multimodal summary? How about consensus on the evaluation measures for the dialogue or meeting summaries? Are minutes and multi-party dialogue summaries the same?

Automatic Minuting is a challenging and not well-defined task. There are no agreed-upon guidelines on how to take minutes, and people adopt different styles to summarize the meeting contents [Nedoluzhko and Bojar, 2019]. The form of the minutes also depends on the meeting's category, the intended audience, and the goal or objective of the meeting. Our special session, SummDial at SIGDial 2021, intended to instigate discussions on these critical challenges. Our goal for this session was to stimulate intense discussions around this topic and set the tone for further interest, research, and collaboration in both Speech and Natural Language Processing communities. A special session on Speech Summarization was held in 2006[2]. Hence, we thought it might be good to gauge the current community interest and have a focused session on this topic. There have been several prominent research on summarizing meetings and dialogues in the SNLP community over the years[3] which signifies the interest and progress made on this topic. We witnessed enthusiastic community participation and interest in our four-hour-long session. We also conducted a 30-minute breakout session on *Multi-party Dialogues and Meeting Summarization, Automatic Minuting* before the special session. We detail the event in the subsequent sections of this report.

2 Call for Papers

For our special session at SIGDial 2021,[4] we invited regular and work-in-progress papers that report:

[2] http://homepages.inf.ed.ac.uk/jeanc/SpeechSummarization06.html

[3] a handy repository of compilation and evolution of summarization research papers http://pfliu.com/pl-summarization/summ_paper.html

[4] https://www.sigdial.org/files/workshops/conference22/

- Current research in multi-party dialogue summarization for summarizing meetings, spoken dialogue, using speech, text, or multimodal data (audio, video),
- Challenges in manual and automatic dialogue summarization evaluation,
- New methods and metrics for manual and automatic dialogue summarization evaluation,
- Challenges and methods in summarizing transcripts in different domains, including legal, educational, political, social, etc.
- Datasets and corpora for dialogue summarization,
- Techniques of data collection, pre-processing, adaptation,
- Ethical issues and possible solutions,
- New systems for dialogue or meeting summarization, or new evaluations of existing systems,
- Qualitative or quantitative comparisons of speech-specific summarization systems and summarization systems imported from the text domain,
- Tools for meeting transcript generation and automatic summarization,
- Topic detection and span identification in meeting transcripts for multi-topic summarization,
- Position papers to reflect on the current state of the art in this topic, take stock of where we have been, where we are, where we are going and where we should go.

We received acceptance notification of our special session from SIGDial 2021 chairs on February 25, and our first call for papers went live on March 2. Researchers had to choose to submit long, short, late-breaking, work-in-progress, or position papers. Regular submissions (long and short) followed the SIGDial 2021 submission process and timeline (April 10 deadline) as they appeared in the SIGDial 2021 proceedings. Late-breaking, Work-In-Progress, and Position Papers had a later submission deadline on June 15. All submission deadlines followed 23:59 GMT-11. Our paper category descriptions went as follows:

Long papers. must describe substantial, original, completed, and unpublished work. Wherever appropriate, concrete evaluation and analysis should be included. These papers would go through the same peer-review process by the SIGDial program committee as papers submitted to the main SIGdial track. These papers will appear in the main SIGdial proceedings and are presented with the main track. Long papers must be no longer than eight pages, including title, text, figures, and tables. An unlimited number of pages is allowed for references. Two additional pages are allowed for appendices containing sample discourses/dialogues and algorithms, and an extra page is allowed in the final version to address reviewers' comments.

Short papers. must describe original and unpublished work. These papers would go through the same peer-review process by the SIGDial program committee as papers submitted to the main SIGdial track. These papers will appear in the main SIGdial proceedings and are presented with the main track. Please note that a short paper is not a shortened long paper. Instead, short papers should have a point that can be made in a few pages, such as a small, focused contribution, a negative result, or an interesting application nugget. It should be no longer than four pages, including title, text, figures, and tables. An unlimited number of pages is allowed for references. One additional page is allowed for sample discourses/dialogues and algorithms, and an extra page is allowed in the final version to address reviewers' comments. An unlimited number of pages are allowed for references.

Late-breaking and Work-in-progress papers. will showcase ongoing work and focused relevant contributions. Submissions need not present original work. Late-breaking and work-in-

progress papers should be no longer than four pages, including title, text, figures and tables, and references. These will be reviewed by the SummDial program committee and posted on the special session website. These papers will be presented as lightning talks or posters during the session. Authors will retain the copyright to their work so that they may submit it to other venues as their work matures.

Position papers. will give voice to authors who wish to take a position on a topic listed above or the field of spoken, dialogue, meeting summarization. Submissions need not present original work and should be two to six pages in length, including title, text, figures and tables, and references. These will be reviewed by the SummDial program committee and posted on the special session website. These papers will be presented as lightning talks or posters during the session. Authors will retain the copyright to their work so that they may submit it to other venues.

3 Format of Special Session

SummDial at SIGDial 2021 had one keynote talk of 45 minutes, one panel discussion of 60 minutes, three long and three short papers, each for 20 minutes. All the sessions were conducted virtually over Zoom. The recording of the session is available here[5]. We carried out the Q&A over Zoom chat and also over the dedicated slack channel provided to us by the SIGDial 2021 organizers. At one point in time, there were about 50 participants in the session.

4 Keynote Speaker

We were delighted to have **Klaus Zechner**[6] from Educational Testing Service, United States as our keynote speaker. His pioneering works on summarization of meeting speech and dialogues helped shape the investigations in this topic further [Zechner and Waibel, 2000; Zechner, 2001a, 2002a]. Klaus Zechner received his Ph.D. from Carnegie Mellon University in 2001 for research on automated speech summarization. This work was published at SIGIR 2001 and in Computational Linguistics (2002). Klaus Zechner is now a Senior Research Scientist in the Natural Language Processing Lab in the Research and Development Division of Educational Testing Service (ETS) in Princeton, New Jersey, USA. Since joining ETS in 2002, he has been pioneering research and development of technologies for automated scoring of non-native speech, leading large R&D projects dedicated to the continuous improvement of automated speech scoring technology. He holds more than 20 patents on technology related to SpeechRater, an automated speech scoring system he and his team have been developing at ETS. SpeechRater is currently used operationally as sole score for the TOEFL Practice Online (TPO) Speaking assessment and, in a hybrid scoring approach, also for TOEFL iBT Speaking. Klaus Zechner authored more than 80 peer-reviewed publications in journals, book chapters, conference and workshop proceedings, and research reports. He also edited a book on automated speaking assessment that was published by Routledge in 2019; it provides an overview of the current state-of-the-art in automated speech scoring of spontaneous non-native speech.

[5] https://tinyurl.com/summdial-recording
[6] https://scholar.google.com/citations?user=eVYrz4EAAAAJ&hl=en

Kindly note that the speaker himself authors the following abstract.

Title of the Talk: ***Who Discussed What With Whom: Is Meeting Summarization A Solved Problem?***

Abstract: While creating audio and video records of multi-party meetings has become easier than ever in recent years, obtaining access to the key contents or a summary of a meeting is non-trivial. In this talk, I will first provide an overview of the main differences between multi-party meetings and news articles – the prototypical domain for most research on summarization so far. In the second part of the talk, a few example approaches to meeting summarization will be presented and discussed, spanning from early research to late-breaking system papers. Finally, I will conclude with thoughts about the current state-of-the-art of the field of meeting summarization and open issues that still need to be addressed by the research community.

Discussion: The discussion that ensued following the keynote talk in the question-answering session included:

- Multimodal summarization of meetings: to track participant emotions to make a better summary, derive inferences, or comprehend disagreements.
- Taking care of temporal aspects in meetings which are not quite obvious in news article summarization
- Handling small talks, irony, or sarcasm in meeting conversations so that they do not appear in the summary
- A "drill-down summarization" of meetings would be a good idea to address the conciseness vs. coverage conundrum in minutes. Readers would have the flexibility to tailor the minutes according to their information needs or level of detailedness.
- *Relevance, Readability, Coverage* are important factors for human evaluation of meeting minutes.

5 Panel Discussion

We had a panel discussion on the topic **Dialogue and Meeting Summarization: Taking Stock and Looking Ahead, Towards Automatic Minuting** with four panelists who are very prominent in the summarization and dialogue community. Our co-organizer, Ondřej Bojar, moderated the panel. Our panelists were Ani Nenkova, Diyi Yang, Chenguang Zhu. Our keynote speaker, Klaus Zechner, also joined the discussion.

- **Ani Nenkova**[7] is a Principal Scientist at Adobe Research, leading the Document Intelligence Lab at Adobe-Maryland. Her main areas of research are computational linguistics and artificial intelligence, with emphasis on developing computational methods for the analysis of text quality and style, discourse, affect recognition, and summarization. She obtained her Ph.D. degree in computer science from Columbia University. Ani is a co-editor-in-chief of the Transactions of the Association for Computational Linguistics (TACL). She was a member

[7] https://www.cis.upenn.edu/~nenkova/

of the editorial board of Computational Linguistics (2009–2011) and an associate editor for the IEEE/ACM Transactions on Audio, Speech, and Language Processing (2015–2018). She regularly serves as an area chair/senior program committee member for ACL, NAACL, and AAAI.

- **Diyi Yang**[8] is an assistant professor in the School of Interactive Computing at Georgia Tech. Her research focuses on Computational Social Science, and Natural Language Processing. Diyi received her Ph.D. from Language Technologies Institute at Carnegie Mellon University. Her work has been published at leading NLP/HCI conferences, and also resulted in multiple paper award (nominations) from EMNLP 2015, ICWSM 2016, SIGCHI 2019, CSCW 2020, SIGCHI 2021. She is named as one of Forbes 30 Under 30 in Science in 2021, and a recipient of IEEE AI 10 to Watch in 2020.
- **Chenguang Zhu**[9] is a Principal Research Manager in Microsoft Cognitive Services Research Group. His research in NLP covers text summarization, knowledge graphs, and task-oriented dialogue. Dr. Zhu has led teams to achieve first place in multiple NLP competitions, including CommonsenseQA, CommonGen, FEVER, CoQA, ARC, and SQuAD v1.0. He holds a Ph.D. degree in Computer Science from Stanford University.

The main objective of this panel was to take stock of the progress in meeting and dialogue summarization from domain experts, discuss the challenges, and chalk out future directions. We decided to keep the panel around the following topics:

- How did our panelists decide to choose multi-party dialogue summarization as their area of research?
- Characteristic of summarization in specific genres: text, speech, dialogues, meeting
- Multi-party meeting summarization evaluation
- Datasets and data acquisition
- Methods and system architectures
- Would starting a shared task cycle help address the various challenges in this domain?

It is exactly twenty years since Klaus Zechner's seminal thesis "Automatic Summarization of Spoken Dialogues in Unrestricted Domains" [Zechner, 2001b] came out. The SNLP community has made much progress in between in several areas, especially with the advent of the Deep Learning era[10]. Increased computational power and resources have enabled us to harness the inherent capabilities of deep neural networks, which were otherwise not possible in earlier days. In case of some problems like machine translation, sometimes the state-of-the-art is able to match the human gold standard [Popel et al., 2020]. The industry has started investing resources in SNLP[11]. Quite often, we hear about some gigantically large language models with billions of parameters[12] surpassing the human benchmarks on some downstream NLP tasks on some leaderboards [13].

[8] https://www.cc.gatech.edu/~dyang888/

[9] https://www.microsoft.com/en-us/research/people/chezhu/

[10] https://ruder.io/nlp-imagenet/

[11] https://gradientflow.com/2021nlpsurvey/

[12] https://venturebeat.com/2021/10/11/microsoft-and-nvidia-team-up-to-train-one-of-the-worlds-largest-language-models/

[13] https://venturebeat.com/2021/01/06/ai-models-from-microsoft-and-google-already-surpass-human-performance-on-the-superglue-language-benchmark/

However, for the problem of meeting summarization, we probably did not make that gigantic leap since Zechner's thesis.

The panelists started by discussing their first steps into summarization and, more specifically, meeting and dialogue summarization. It is a significant problem to address in the current scenario when most of our interactions have gone virtual due to the pandemic. While the entire conversation is available for public viewing on the SummDial website,[14] we try to summarize the crucial points that came up during the panel.

- There is *no ideal meeting summary*. The definition of an ideal meeting summary should come from the behavioral perspective of different readers. Industry who run meeting tools may step in here and do a user study (obviously with appropriate permissions and privacy, ethical considerations). It is important to consider the subjectivity associated with the task - for whom has the summary been created?
- We should not have just one reference summary but multiple summaries written by different meeting participants to train our systems. Non-participant minutes suffer in information quality due to their lack of context.
- Meeting transcripts are long text documents. Hence capturing the entire semantics of what was discussed in the meeting is challenging. It may be helpful to represent meetings as topical segments or discourse relations or in some graphical form to counter the information management in the long discourse.
- A line of investigation could be to generate user-centric "personalized" minutes based on question-answering the meeting transcripts.
- Although ROUGE [Lin and Hovy, 2003] is well-past its life expectancy, we still do not have a strong alternative. Reference summaries are subjective as well. A line of thought is that if we can discard reference summaries [Louis and Nenkova, 2013] and instead use the transcript for evaluation. Maybe one can align the target summaries with the transcript itself and see what the *coverage quotient* of the minutes is. However, reference summaries are essential to training supervised systems. More research should be directed towards *ROUGE-less, reference-less* summarization to have a better answer to this proposition.
- Human evaluation in this task is critical yet very difficult, especially for a non-participant. Even for active participants, the minutes could differ hugely in content. Our Task C in the AutoMin shared task [Ghosal et al., 2021] is motivated precisely towards this point: *decide whether two minutes belong to the same meeting.*
- Available meeting summarization datasets like AMI and ICSI or even the AutoMin shared task dataset are small-scale; it is almost impossible to use them to train a deep network. Dataset development or data acquisition in this domain is challenging primarily because of ethical and privacy reasons. Otherwise, the pandemic has posed a unique opportunity before us where thousands of meetings are being recorded and minuted every day. Data banks[15] are pretty popular in the healthcare domain, and maybe we could try setting up such data banks following all ethical and privacy regulations. We would need the community support to donate their meetings and minutes to such a repository to continue associated research.
- How about using the large-scale language models like GPT-3 [Brown et al., 2020] to generate synthetic meeting transcripts? Care should be taken so that these models do not leak the

[14] https://elitr.github.io/automatic-minuting/summdial.html
[15] https://en.wikipedia.org/wiki/Data_bank

user-sensitive information (which was used to train it) during generation.

- A vital aspect to consider in the summaries or minutes is to address the authority or background of the speakers. E.g., a project leader's speech would probably be more critical than a vendor's in a project meeting.
- Maybe we should focus on important sub-tasks associated with this problem like *topic-segmentation, topical highlights, multiple summary training, discourse relations, significance identification, etc.* Then accumulate the findings towards the larger problem.
- Unsupervised methods, graph-based methods, multimodal summarization, infusing discourse relations, or relevant linguistic information in transformer models could be other directions to explore for this problem.
- Start the shared task cycle for this problem. Our AutoMin shared task could be the first instance of this. The recent astonishing performance of machine translation models for text and speech could be primarily attributed to the various shared tasks in WMT[16], IWSLT[17] over the years.

6 Presented Papers

As mentioned earlier, we had six accepted papers in SummDial. Out of the six accepted papers, four were accepted in the SIGDial 2021 main conference and appeared in the SIGDial 2021 proceedings. The other two were specific to SummDial and non-archival.

- **Coreference-Aware Dialogue Summarization** by Liu et al. [2021]. In this work, the authors investigate different approaches to explicitly incorporate coreference information in neural abstractive dialogue summarization models to tackle challenges like unstructured information exchange in dialogues, informal interactions between speakers, and dynamic role changes of speakers as the dialogue evolves. Their experiments implied that it is useful to utilize coreference information in dialogue summarization. This paper was also the **best paper award winner in SIGDial 2021**.
- **Weakly Supervised Extractive Summarization with Attention** by Zhuang et al. [2021]. In this work, the authors develop a general framework that generates extractive summarization as a byproduct of supervised learning tasks for indirect signals via the help of an attention mechanism. They demonstrate that their models can reliably select informative sentences and words for automatic summarization.
- **Incremental Temporal Summarization in Multi-party Meetings** by Manuvinakurike et al. [2021]. The authors develop a dataset for incremental temporal summarization in a multi-party dialogue. They leverage the question generation paradigm to automatically generate questions from the dialogue to draw the attention of the user towards the contents they need to summarize; a kind of personalized summary generation of the meeting proceedings which is rightly motivated by the fact that not all participants would have similar information needs in the minutes.
- **Mitigating Topic Bias when Detecting Decisions in Dialogue** by Karan et al. [2021].

[16] https://aclanthology.org/venues/wmt/
[17] https://iwslt.org

Here, the authors explore the task of detecting decision-related utterances in multi-party dialogue. They experimented with traditional machine learning and transformer-based deep learning approaches. They found that models rely more on topic-specific words that decisions are about rather than on words that more generally indicate decision making.

- **Creating a Dataset of Abstractive Summaries of Turn-labeled Spoken Human-Computer Conversations** In this work, the authors presented a novel dataset of abstractive summaries of turn-labeled spoken human-computer conversations in Dutch. They also include a baseline transformer-based summarization model; the dataset can also be used for investigating automatic dialogue turn splitting and turn labeling.
- **Dynamic Sliding Window for Meeting Summarization** by Liu and Chen [2021]. In this work, the authors propose a dynamic sliding window strategy to counter the challenge of summarizing long meeting transcripts. Their "divide and conquer" strategy based on BART [Lewis et al., 2020] achieved outputs of higher factual consistency than the base model.

7 SIGDial 2021 Break-out session

In addition to the special session, we also conducted a breakout session on **Multi-party Dialogues and Meeting Summarization, Automatic Minuting**[18] at SIGDial 2021. The motivation of conducting this special session was to have a community brainstorming session on:

Can we imagine a future where automatically the minutes are sent to the participants immediately after the meeting and just via hovering over the past meeting invites one can see the minutes of the meeting?

We also intended to host the 30-minute breakout session to have a quick community take on the following topic-relevant issues and set the stage for our special session.

1. Why is multi-party meeting or dialogue summarization challenging?
2. What do you think about resource creation in this genre? What are the challenges/obstacles? We see there are only a few resources (AMI, ICSI, etc.), did we miss anything important, e.g., because it is too local, non-English?
3. Evaluation: How important is human evaluation here? For automatic evaluation, is it time to do away with ROUGE? What are the alternatives?
4. What do you think about using off-the-shelf text summarization models here? What are the considerations that one may need to take care of?
5. What would be the characteristic of ideal minutes of the meeting?
6. What, according to you, should be the research directions/sub-problems for the NLP and Speech community on this problem?

Around 20 people attended the breakout session, which was just before the opening ceremony of the conference. The major points that came out during the discussions were:

- Participants pointed out that the characteristics of good minutes include: if all the topics

[18] https://tinyurl.com/sigdial2021-minuting-breakout

discussed in the meeting are touched upon (coverage), participation ratio (who was doing most of the talking or driving the conversation), if the important action items are properly extracted (the ToDo list). Also, the evaluation criteria for minutes will depend on the type of meeting. Different meetings have different agendas, expectations. Care should be taken so that one speaker does not "hijack" the meeting and the minutes do not contain only their points but also have minute items from other participants. Another important issue is to encompass the human controllability factor or generate "personalized" summaries. Not every participant or a non-participant would have the same information-need from a meeting. A marriage between "personalization" and "summarization" would be an interesting direction to pursue to counter the "subjectivity" associated with this task. Another way could be treating the summary generation as a question answering task where the user would be able to query the meeting transcript and get their personalized summary, something which has been tried with the QMSumm dataset [Zhong et al., 2021] as a query-based multi-domain meeting summarization task. Evaluation of the minutes in such cases would be much easier and more objective, like if the user's information need is satisfied, which can be found by looking into the answers in response to the user queries.

- Participants also talked about the trade-off between "conciseness" and "coverage" in meeting summaries or minutes. For a subjective task as this, it may be worthwhile to generate "slider summaries" where the user can tailor the minute with the level of details they would want to consume. A more practical variant could be the "hypertext summaries" where the reader gets an abstract view of the meeting in the minutes but can zoom in to more details by clicking on the hypertexts.
- One participant pointed out that it would be helpful to have a "taxonomy of meetings". Since there are various meetings with different goals and content, the taxonomy has to be meeting category-specific.
- One participant opined that "somethings are better not automated" and for this particular use-case may be a "human-in-the-loop" summarization would help owing to a variety of reasons, ethical issues and privacy being the primary ones.
- Treating the meeting minutes generation as a "slot-filling task" according to some preset agenda items can be another possible way to ensure "coverage".
- All participants agreed that evaluation for this application is challenging as it is complicated to compare with a reference summary which is itself very subjective. Evaluation via crowdsourcing is not very reliable as validating the understanding of the crowd workers is not possible. Crowdsourced annotations are fine for tasks that have shorter inputs. But for a task as this where the annotator has to comprehend the entire discourse of a meeting (sometimes not only via the transcript but also via the audio/video recordings), we would need very specialized people to do so. Here lies the conundrum about who creates the reference summaries; a meeting participant would always have a better understanding and context of the meeting proceedings than a non-participant. It could be a good investigation objective to study the minutes created by participants and non-participants and see which are more informative to meeting participants and absentees.
- One participant argued how about treating minutes' evaluation as an entailment problem? Could we automatically answer if the minute statements are consistent with the transcript facts?

- ROUGE has been there as a de-facto automatic summarization metric for quite some time. But ROUGE has its own limitations. Some participants pointed out that there are some new summarization metrics in the town like BERTScore [Zhang et al., 2020], which are encouraging. A summarization evaluation toolkit would be a useful instrument to study and validate the various metrics against different categories of meetings and minutes by different creators. We refer to the SummEval[19] [Fabbri et al., 2021] and SacreROUGE[20] [Deutsch and Roth, 2020] packages here which caters to the aforesaid requirement.
- There is a dearth of large-scale, real-life meeting datasets. However, there are some recent multiparty dialogue summarization datasets like DialogSum [Chen et al., 2021a], SamSum [Gliwa et al., 2019], MediaSum [Zhu et al., 2021] which can be taken as a proxy. One can also train their deep models on such datasets or related tasks like podcast summarization (The Spotify Podcast Dataset [Clifton et al., 2020]) and see how the learning transfers to the meeting summarization task. One participant suggested that one can make use of the publicly available debates as a data source. However, the domain and style would be different from multi-party project meetings.
- Existing datasets like AMI [McCowan et al., 2005] or ICSI [Janin et al., 2003] contains meetings which are conducted in a staged environment. However, staged meetings cannot resemble the spontaneous conversations in actual meetings. Again people are not comfortable sharing their free flow conversations in actual meetings, which might contain personal or sensitive information. As part of our ELITR project, we too made a call for donating meeting conversations (audio, transcripts)[21], but received very less response. One way out could be to properly de-identify the data, get explicit consent from the participants, omit the conversations that may include personally identifiable or sensitive information. We followed these steps while we prepared our dataset for the AutoMin[22] shared task at Interspeech 2021. We invited the participants to explore our dataset, which consists of meetings in English and Czech with multiple summaries/minutes written by several different annotators for almost every meeting.
- We require more community events like the **AutoMin** shared task [Ghosal et al., 2021] on automatic minuting to make progress in this very relevant, timely, and important NLP application. The DialogSum challenge [Chen et al., 2021b] at INLG 2022[23] is one such event which we look forward to.
- Our experience says that generating minutes of the meeting is a tedious task and more so for a non-participant in the meeting. The data creation is demanding in terms of costs, expertise, availability, and also in terms of retaining the interest and attention of the annotator.
- One of our participants helpfully pointed the audience to the CALO[24] meeting assistance project [Tür et al., 2010] that attempted to integrate numerous AI technologies into a cognitive assistant.
- The panel agreed that rigorous studies on how text summarization approaches can be suit-

[19] https://github.com/Yale-LILY/SummEval
[20] https://github.com/danieldeutsch/sacrerouge
[21] https://elitr.eu/recipe-for-miracles-to-happen/
[22] https://elitr.github.io/automatic-minuting/
[23] https://cylnlp.github.io/dialogsum-challenge/
[24] https://en.wikipedia.org/wiki/CALO

ably applied to multi-party dialogues and meeting summarization are to be conducted [Singh et al., 2021] and the pitfalls to be identified.
- Finally, the participants concur that in the current time, when it is possible to record the virtual meetings so easily, there is an ample opportunity to fuel the concerned research. The small/large corporations, academia can come forward to donate data from their project meetings to create a large-scale community dataset to spearhead research in this domain.

Due to paucity of time, we had to cut short our session and carry the discussion forward in the SummDial special session.

8 Conclusions and Future Directions

With the intense discussions during the breakout session, panels, and community-wide participation in the event, we believe SummDial got the desired headstart. According to NLPExplorer[25], the SummDial URL[26] was one of the top-visited 10 URLs in #NLProc Twitter in 2021. We envisage that the community would take the learnings and findings forward, and we would be able to discuss/brainstorm some more challenges and updates in the next iteration of SummDial in 2022. The next directions, challenges in multi-party dialogues, and meeting summarization are already spelled out loud in our panel and breakout sessions. To re-iterate, we need to prioritize and re-prioritize *large-scale dataset creation on automatic minuting, study the trade-off between conciseness and coverage in generating minutes, generating personalized summaries, organize more shared tasks like* ***AutoMin and DialogSum****, develop better evaluation schema, and study effects of transfer learning, multitasking from associated tasks.* As researchers in this domain, we had a great learning and enriching experience in SummDial, and we hope our participants had too. We witnessed encouraging participation in our AutoMin shared task from many attendees of SummDial. We are motivated and look forward to continuing this community-building exercise and organizing events for this very relevant and significant task for the SNLP community.

9 About the Organizers

SummDial @ SIGDial 2021 was organized by:

- **Tirthankar Ghosal**[27] is a researcher at the Institute of Formal and Applied Linguistics, Charles University Prague, Czech Republic. His main research interests are NLP/ML for Scientific Discourse Processing and Peer Reviews, Text/Dialogue Summarization, Argumentation Mining.
- **Muskaan Singh**[28] is a researcher with the Institute of Formal and Applied Linguistics, Charles University, Czech Republic. Her main research interests are Machine Translation and Automatic Summarization of Speech/Dialogues.

[25] http://lingo.iitgn.ac.in:5001
[26] http://lingo.iitgn.ac.in:5001/twitter
[27] https://member.acm.org/~tghosal
[28] https://ufal.mff.cuni.cz/muskaan-singh-0

- **Anja Nedoluzhko**[29] is a researcher at the Institute of Formal and Applied Linguistics, Charles University, Prague. Her main research interests concern phenomena exceeding the sentence boundary (coreference, bridging, discourse analysis).
- **Ondřej Bojar**[30] is an associate professor at the Institute of Formal and Applied Linguistics, Charles University, Prague. His main research interest is machine translation, but he was also involved in treebanking and lexicographic projects. He has led several large-scale NLP projects and is also the primary investigator of the EU-funded H2020 ELITR project [Bojar et al., 2020] whose Automatic Minuting module can be seen as the origin of this special session.

Acknowledgements

We thank the organizers, volunteers of SIGDial 2021, especially the chairs, for providing us with the requisite support and infrastructure to host SummDial online. We thank our speakers, panelists, authors for their valuable talks and inputs. We extend our gratitude to the program committee for helping us craft the program. Our program committee members were:

- Shantipriya Parida, Idiap Research Institute, Switzerland
- Sovan Kumar Sahoo, Indian Institute of Technology Patna, India
- Sandeep Kumar, Indian Institute of Technology Patna, India
- Tirthankar Ghosal, Institute of Formal and Applied Linguistics, Charles University, Czech Republic
- Muskaan Singh, Institute of Formal and Applied Linguistics, Charles University, Czech Republic
- Anja Nedoluzhko, Institute of Formal and Applied Linguistics, Charles University, Czech Republic
- Ondřej Bojar, Institute of Formal and Applied Linguistics, Charles University, Czech Republic

Lastly, we thank the participants of SummDial for enthusiastically taking part in the special session. We hope to continue the discussions around this important topic with new updates in the next iteration of this special session in 2022. We organized SummDial as part of our involvement in the European Live Translator (ELITR) project, which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460.

References

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online, November

[29] https://ufal.mff.cuni.cz/anna-nedoluzhko
[30] https://ufal.mff.cuni.cz/ondrej-bojar

2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.751. URL https://aclanthology.org/2020.emnlp-main.751.

Ondřej Bojar, Dominik Macháček, Sangeet Sagar, Otakar Smrž, Jonáš Kratochvíl, Ebrahim Ansari, Dario Franceschini, Chiara Canton, Ivan Simonini, Thai-Son Nguyen, Felix Schneider, Sebastian Stücker, Alex Waibel, Barry Haddow, Rico Sennrich, and Philip Williams. ELITR: European live translator. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 463–464, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.53.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.449. URL https://aclanthology.org/2021.findings-acl.449.

Yulong Chen, Yang Liu, and Yue Zhang. Dialogsum challenge: Summarizing real-life scenario dialogues. In Anya Belz, Angela Fan, Ehud Reiter, and Yaji Sripada, editors, *Proceedings of the 14th International Conference on Natural Language Generation, INLG 2021, Aberdeen, Scotland, UK, 20-24 September, 2021*, pages 308–313. Association for Computational Linguistics, 2021b. URL https://aclanthology.org/2021.inlg-1.33.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.519. URL https://aclanthology.org/2020.coling-main.519.

Daniel Deutsch and Dan Roth. SacreROUGE: An open-source library for using and developing summarization evaluation metrics. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 120–125, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlposs-1.17. URL https://aclanthology.org/2020.nlposs-1.17.

Daniel Deutsch and Dan Roth. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.conll-1.24.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 04 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00373. URL https://doi.org/10.1162/tacl_a_00373.

Tirthankar Ghosal, Muskaan Singh, Anja Nedoluzhko, and Ondřej Bojar. Overview of the First Shared Task on Automatic Minuting (AutoMin) at Interspeech 2021. In *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021. URL http://dx.doi.org/10.21437/AutoMin.2021-1.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003*, pages 364–367. IEEE, 2003. doi: 10.1109/ICASSP.2003.1198793. URL https://doi.org/10.1109/ICASSP.2003.1198793.

Mladen Karan, Prashant Khare, Patrick Healey, and Matthew Purver. Mitigating topic bias when detecting decisions in dialogue. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 542–547, Singapore and Online, July 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.sigdial-1.56.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Chin-Yew Lin and Eduard H. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Marti A. Hearst and Mari Ostendorf, editors, *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*. The Association for Computational Linguistics, 2003. URL https://aclanthology.org/N03-1020/.

Zhengyuan Liu and Nancy F. Chen. Dynamic sliding window for meeting summarization. *CoRR*, abs/2108.13629, 2021. URL https://arxiv.org/abs/2108.13629.

Zhengyuan Liu, Ke Shi, and Nancy Chen. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519, Singapore and Online, July 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.sigdial-1.53.

Annie Louis and Ani Nenkova. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300, June 2013. doi: 10.1162/COLI_a_00123. URL https://aclanthology.org/J13-2002.

Ramesh Manuvinakurike, Saurav Sahay, Wenda Chen, and Lama Nachman. Incremental temporal summarization in multi-party meetings. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 530–541, Singapore and Online, July 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.sigdial-1.55.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. The ami meeting corpus. In *Proceedings of the 5th international conference on methods and techniques in behavioral research*, volume 88, page 100. Citeseer, 2005. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.95.6326.

Anna Nedoluzhko and Ondrej Bojar. Towards automatic minuting of the meetings. In *ITAT*, 2019. URL http://ceur-ws.org/Vol-2473/paper3.pdf.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1): 1–15, 2020. URL https://www.nature.com/articles/s41467-020-18073-9.

Muskaan Singh, Tirthankar Ghosal, and Ondřej Bojar. An empirical analysis of text summarization approaches for automatic minuting. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, November 2021. Association for Computational Linguistics.

Gökhan Tür, Andreas Stolcke, L. Lynn Voss, Stanley Peters, Dilek Hakkani-Tür, John Dowding, Benoît Favre, Raquel Fernández, Matthew Frampton, Michael W. 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. The CALO meeting assistant system. *IEEE Trans. Speech Audio Process.*, 18(6): 1601–1611, 2010. doi: 10.1109/TASL.2009.2038810. URL https://doi.org/10.1109/TASL.2009.2038810.

Klaus Zechner. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In W. Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel, editors, *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*,

pages 199–207. ACM, 2001a. doi: 10.1145/383952.383989. URL https://doi.org/10.1145/383952.383989.

Klaus Zechner. Automatic summarization of spoken dialogues in unrestricted domains. 2001b. URL https://isl.anthropomatik.kit.edu/downloads/Zechner_Klaus_thesis.pdf.

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

Klaus Zechner. Summarization of spoken language-challenges, methods, and prospects. *Speech technology expert eZine*, 6, 2002b. URL http://www.cs.cmu.edu/~./zechner/ezine.ps.

Klaus Zechner and Alex Waibel. DIASUMM: flexible summarization of spontaneous dialogues in unrestricted domains. In *COLING 2000, 18th International Conference on Computational Linguistics, Proceedings of the Conference, 2 Volumes, July 31 - August 4, 2000, Universität des Saarlandes, Saarbrücken, Germany*, pages 968–974. Morgan Kaufmann, 2000. URL https://aclanthology.org/C00-2140/.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.472. URL https://aclanthology.org/2021.naacl-main.472.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.474. URL https://aclanthology.org/2021.naacl-main.474.

Yingying Zhuang, Yichao Lu, and Simi Wang. Weakly supervised extractive summarization with attention. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 520–529, Singapore and Online, July 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.sigdial-1.54.