# Summarization of Dialogues and Conversations At Scale

**Diyi Yang**
Computer Science Department
Stanford University
diyiy@stanford.edu

**Chenguang Zhu**
Microsoft Azure Cognitive
Services Research
chezhu@microsoft.com

## 1 Introduction

Conversations are the natural communication format for people. This fact has motivated the large body of question answering and chatbot research as a seamless way for people to interact with machines. The conversations between people however, captured as video, audio or private or public written conversations, largely remain untapped as a source of compelling starting point for developing language technology. Summarizing such conversations can be enormously beneficial: automatic minutes for meetings or meeting highlights sent to relevant people can optimize communication in various groups while minimizing demands on people's time; similarly analysis of conversations in online support groups can provide valuable information to doctors about the patient concerns.

Summarizing written and spoken conversation poses unique research challenges—text reformulation, discourse and meaning analysis beyond the sentence, collecting data, and proper evaluation metrics. All these have been revisited by researchers since the emergence of neural approaches as the dominant approach for solving language processing problems. In this tutorial, we will survey the cutting-edge methods for summarization of conversations, covering key sub-areas whose combination is needed for a successful solution.

## 2 Tutorial Outline

This will be a **three-hour** tutorial devoted to the **cutting-edge** topic of conversation summarization. Our tutorial will include three sessions. Each session will be 40 minutes, followed by 10 minutes for Q&A and 10 minutes for a break. Each part includes an overview of the corresponding topic and widely used methods and a deep dive into representative research. The detailed tutorial schedule can be found in Table 1.

| **Slot** | **Theme** |
|---|---|
| *Session 1: Introduction to Conversation Summarization* | |
| 14:15 – 14:20 | Tutorial presenters introduction |
| 14:20 – 14:35 | Introduce the task, its history and impact |
| 14:35 – 15:15 | Compare document and conversation summarization (CS), and datasets |
| 15:15 – 15:30 | Coffee Break |
| *Session 2: Pretraining and Methods* | |
| 15:30 – 15:50 | Pretraining in Conversation summarization |
| 15:50 – 16:30 | Classic models in summarizing conversations, dialogue, and meetings |
| 16:30 – 16:45 | Coffee Break |
| *Session 3: Evaluation and Challenges* | |
| 16:45 – 17:00 | Structures and knowledge to improve conversation summarization |
| 17:00 – 17:30 | Evaluation metrics and issues |
| 17:30 – 17:45 | Open questions and challenges |
| 17:45 – 18:00 | Conclusion |

Table 1: Tutorial schedule.

### 2.1 Introduction

Compared to summarizing news reports or encyclopedia articles, summarizing conversations—an essential part of human-human/machine interaction where most important pieces of information are scattered across various utterances of different speakers—remains relatively under-investigated. Yet capabilities for automatic dialog summarization hold the promise to facilitate information access, especially in corporate (or large group) settings. For instance, participants in corporate meetings usually want to get high-level synopsis of the meeting content and action items to review after meeting. People who miss the meeting also want to quickly get the main topics discussed in the meeting. Another scenario is customer service, where customer calls with the agents can be summarized, categorized and analyzed. This can help agents to find frequent problems to answer, product issues to follow, etc. in order to improve service quality.

*Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics Tutorial Abstracts*, pages 13–18
May 2-4, 2023 ©2023 Association for Computational Linguistics

### 2.1.1 Datasets

Text summarization was dominated by unsupervised methods for decades (Nenkova and McKeown, 2011, 2012), due to the lack of suitable size datasets for the task. The field has been transformed by the introduction of large-scale datasets such as CNN/DailyMail (Hermann et al., 2015) and XSUM (Narayan et al., 2018). The past decade also saw the emergence of many dialogue and conversation datasets such as MultiWoz (Budzianowski et al., 2018) and Ubuntu (Lowe et al., 2015). However, the progress in datasets for conversation summarization is comparatively limited. Existing public datasets in this field are either small in scale or limited to a specific domain. We argue that there are two main reasons. First, unlike news articles, conversations usually happen in relatively private environment, which raises privacy concerns for release of public datasets. Secondly, a conversation is typically quite long, and the conversation participants often have different standpoints and language styles with frequent topic changes. These all propose great challenges for labelers to produce accurate summaries as ground-truth labels.

Despite these difficulties, new research datasets for conversation summarization have been developed recently. For instance, SAMSum (Gliwa et al., 2019a) hires linguists to write summaries for 16,369 messenger-like open-domain daily conversations. MediaSum (Zhu et al., 2021) leverages public transcripts with overviews and topic descriptions from CNN and NPR to produce 463.6K dialogues with summaries. DialogSum (Chen et al., 2021) hires annotators to write summaries for 13,460 dialogues from several public dialogue corpora. The CovoSumm dataset (Fabbri et al., 2021) provides 250 development and 250 test summaries for dialogs from broad domains covering news article comments, discussion forms and debates, community question answering, and email threads. QMSum (Zhong et al., 2021b) consists of 1,808 query-summary pairs over 232 meetings. We will provide a systematic review of these newly released resources.

## 2.2 Methods

### 2.2.1 Pretraining

Pre-trained language representations are at the core of most NLP technologies (Devlin et al., 2018; Radford et al., 2018; Liu et al., 2019b). They provide representations that capture language meaning from large amount of data, easily tunable for specific downstream tasks. For instance, the super language models with hundreds of billions of parameters, e.g., GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022) and OPT (Zhang et al., 2022), have demonstrated strong capabilities in many different tasks, including dialogue summarization.

These foundation models are typically trained on web corpus, comprising mostly of articles and monologues, partly because of their general availability and partly because conversational online exchanges often contain more problematic stereotypes. However, matching pre-training data to downstream tasks—both in the type of textual data and the self-supervised objective for pretraining—is important for performance in downstream tasks. Motivated by this, several pre-trained dialogue models have been proposed for conversation summarization (Feng et al., 2021). These methods pretrain the model with self-supervised tasks, such as recovering altered conversation turns, predicting speakers, reordering shuffled turns, predicting masked utterances in the conversation. These tasks only require unlabeled conversation corpus, i.e., no labeled summary is needed, which greatly expands the availability of data that can be used to improve the quality of pre-trained models. It has been shown that after such pre-training on conversation corpus, the performance on downstream conversational summarization tasks can be greatly improved.

For instance, DialogLM (Zhong et al., 2021a) continues to pre-train the UniLMv2 (Bao et al., 2020) model with several window-based denoising tasks such as recovering the conversation text after randomly reshuffling the turns. The pre-training leads to considerable improvement on both conversation understanding and summarization tasks. HMNet (Zhu et al., 2020a) takes a different approach by creating pseudo meetings from news summarization datasets. The model is trained to produce the summary which is the concatenation of the 4 news articles' summaries. Experiments show that the pre-training can increase the ROUGE-1 metric on AMI (McCowan et al., 2005) and ICSI (Janin et al., 2003) by 4~5 points.

We will overview pretraining methods specific to dialog summarization and will sythesize the impact of data and objectives on downstream tasks across the published literature.

#### 2.2.2 Abstractive Summarization Models

Unlike news, conversations can rarely be summarized meaningfully with extractive approaches. Abstractive summarization is the default expectation for dialogues. Other than directly applying document summarization models to conversational settings (Gliwa et al., 2019b), models tailored for conversation are designed to achieve the state-of-the-art performances such as modeling conversations in a hierarchical way (Zhao et al., 2019; Zhu et al., 2020b). The rich structured information in conversations are also explored and leveraged such as dialogue acts (Goo and Chen, 2018), key point/entity sequences (Liu et al., 2019a; Narayan et al., 2021), topic segments (Liu et al., 2019c; Li et al., 2019), stage developments (Chen and Yang, 2020), discourse relations (Chen and Yang, 2021b; Feng et al., 2020a). Recent work has also explicitly models coreference information to deal with the complex coreference phenomenon in dialogues (Liu et al., 2021). External information like commonsense knowledge has also been incorporated to help understand the global conversation context as well (Feng et al., 2020b).

We will cover how classic statistical models are used to summarize dialogues and multi-party meetings, as well as the recent techniques in using large pretrained language models and diverse neural architectures that take into account conversation characteristics. We will also go through and discuss how to evaluate dialogue summarization models, ranging from classical ROUGE to the recent automatic metrics like BERTScore, as well as multiple widely used qualitative measures.

### 2.3 Open Questions

We will conclude the tutorial with a discussion of more exploratory work around the actual usability of summarization approaches. For example, naturally occurring conversations are long, so segmentation and representations for long text become necessary. Processing time and processing cost for many of the methods is high, both because of the complexity of analyzing discourse and topics and because of the length of the input. We will conclude the section by covering estimates of costs for conversation summarization.

We will cover a discussion on robustness of methods, i.e. their ability to generalize across datasets rather than needing finetuning for each dataset. Finetuning different versions of the model for each dataset is not practical, as maintaining and deploying different fine-tuned versions is less realistic to be done in practice.

Finally, we will discuss scenarios for user evaluation of conversation summarization technology. Such extrinsic evaluations would be needed to move the technology from the realm of research to technological reality. We will include a short segement in which tutorial participants will brainstorm study designs, to validate specific claims about the utility of summarization models.

We will also discuss potential biases and ethical issues related to conversation summarization. This last part of the tutorial is meant to introduce open research questions, so that newcomers to the field of conversation summarization can be equipped to make their own contributions in some of the areas of open questions.

## 3 Tutorial Presenters

**Diyi Yang** is an assistant professor in the Computer Science Department at Stanford University. Her research focuses on dialogue summarization, learning with limited and noisy text data, user-centric language generation, and computational social science. Diyi has organized four workshops at NLP conferences: Widening NLP Workshops at NAACL 2018 and ACL 2019, Casual Inference workshop at EMNLP 2021, and NLG Evaluation workshop at EMNLP 2021. She also gave a tutorial at ACL 2022 on Learning with Limited Data.

**Chenguang Zhu** is a Principal Research Manager in Microsoft Azure Cognitive Services Research Group, where he leads the Knowledge & Language Team. His research in NLP covers text summarization, task-oriented dialogue and knowledge graph,. Dr. Zhu has led teams to achieve first places in multiple NLP competitions. He holds a Ph.D. degree in Computer Science from Stanford University. Dr. Zhu has given talks at Stanford University, Carnegie Mellon University and UC Berkeley. He has given tutorials on Knowledge-Augmented Methods for Natural Language Processing at ACL 2022 and WSDM 2023. He is also the main organizer of The Workshop on Knowledge Augmented Methods for NLP at AAAI 2023.

## 4 Diversity Considerations

The conversation summarization techniques we introduce is language agnostic. Thus, they can be applied to data in various languages and localities with some extent of adaption. Code-switch and multilingual models for conversation summarization can scale this work beyond English.

Our presenter team will share our tutorial with a worldwide audience by promoting it on social media. Our presenters span over junior (Diyi Yang) and senior researchers (Chenguang Zhu) with a female. Diyi is from academia and Chenguang is from industry. Thus, we have diversified instructors which will also help encourage diverse audience. Diyi has experience co-organizing Widening NLP Workshops at both NAACL and ACL, and actively works on inviting undergraduate students to research and promoting diversity such as by speaking at AI4ALL and local high-schools at Atlanta. We will work with ACL/NAACL D&I teams, and consult resources such as the BIG directory to diversify our audience participation.

## 5 Reading List and Prerequisite

### 5.1 Prerequisite

The prerequisite includes familiarity with basic machine learning and deep learning models, especially those typically used in modern NLP for summarization, including sequence to sequence learning, transformers, etc. Furthermore, this tutorial assumes background in basic probability, linear algebra, and calculus. We will also provide introduction materials and additional readings.

**Reading List**

1. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization (Gliwa et al., 2019c)

2. A Hierarchical network for abstractive meeting summarization with cross-domain pretraining (Zhu et al., 2020a)

3. Dialoglm: Pre-trainedmodel for long dialogue understanding and summa-rization (Zhong et al., 2021a)

4. Multi-View Sequence-to-Sequence Models with Conversational Structure for Abstractive Dialogue Summarization (Chen and Yang, 2020)

5. Simple Conversational Data Augmentation for Semi-supervised Abstractive Dialogue Summarization (Chen and Yang, 2021a)

6. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization (Zhong et al., 2021b)

7. ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining (Fabbri et al., 2021)

### 5.2 Breadth

While we will give pointers to dozens of relevant papers over the course of the tutorial, we plan to cover around 7-8 research papers in close detail. Only 1-2 of the “deep dive” papers will come from the presenter team.

## 6 Tutorial Details

### 6.1 Audience Size

We expect the audience size to be around 100 for a physical conference, and around 150 for a virtual conference. Our tutorial will likely bring a similar audience as the SummDial: A SIGDial 2021 Special Session on Summarization of Dialogues and Multi-Party Meetings https://elitr.github.io/automatic-minuting/summdial.html.

### 6.2 Preferable venues

ACL-IJCNLP, EMNLP and EACL would be preferable, as they would fit better with the organizers’ schedules and our tutorial’s emphasis on machine learning. We would like to have access to Gather.Town for interactive Q&A for the online portion of the tutorial.

### 6.3 Open Access

We will put the slides, code, and other teaching materials online for public access, as well as consent to adding the video recording of our tutorial in the ACL Anthology.

## 7 Ethics Statement

Certain conversation data might come from private dialogues between people. Thus, privacy considerations must be take to ensure all data that is released conforms to regulations and are under consent. As conversations and large-pretrained language models may have bias in various forms, summarization models may contain the same form of bias and should be reviewed and modified if necessary.

## References

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pages 642–652. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

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.

Jiaao Chen and Diyi Yang. 2021a. Simple conversational data augmentation for semi-supervised abstractive conversation summarization.

Jiaao Chen and Diyi Yang. 2021b. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, and Yue Zhang. 2021. Dialogsum challenge: Summarizing real-life scenario dialogues. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Xiachong Feng, Xiaocheng Feng, Bing Qin, Xinwei Geng, and Ting Liu. 2020a. Dialogue discourse-aware graph convolutional networks for abstractive meeting summarization.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2020b. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. *arXiv preprint arXiv:2010.10044*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019a. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019b. 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. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019c. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *EMNLP-IJCNLP 2019*, page 70.

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

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*, pages 1693–1701.

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. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*, 1:I–I.

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.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019a. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD19, page 1957–1965, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. 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. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 88:100.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simoes, and Ryan McDonald. 2021. Planning with entity chains for abstractive summarization.

Ani Nenkova and Kathleen R. McKeown. 2011. Automatic summarization. *Found. Trends Inf. Retr.*, 5(2-3):103–233.

Ani Nenkova and Kathleen R. McKeown. 2012. A survey of text summarization techniques. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 43–76. Springer.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference*, WWW '19, page 3455–3461, New York, NY, USA. Association for Computing Machinery.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021a. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *arXiv preprint arXiv:2109.02492*.

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

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020a. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. arXiv:2004.02016.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020b. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.