SmartMeeting: Automatic Meeting Transcription and Summarization for In-Person Conversations

Yuanfeng Song[1,2], Di Jiang[2], Xuefang Zhao[2], Xiaoling Huang[2]
Qian Xu[2], Raymond Chi-Wing Wong[1], Qiang Yang[1,2]
[1]The Hong Kong University of Science and Technology [2]AI Group, WeBank Co., Ltd, China
{songyf,raywong,qyang}@cse.ust.hk

ABSTRACT

Meetings are a necessary part of the operations of any institution, whether they are held online or in-person. However, meeting transcription and summarization are always painful requirements since they involve tedious human effort. This drives the need for automatic meeting transcription and summarization (AMTS) systems. A successful AMTS system relies on systematic integration of multiple natural language processing (NLP) techniques, such as automatic speech recognition, speaker identification, and meeting summarization, which are traditionally developed separately and validated offline with standard datasets. In this demonstration, we provide a novel productive meeting tool named *SmartMeeting*, which enables users to automatically record, transcribe, summarize, and manage the information in an in-person meeting. SmartMeeting transcribes every word on the fly, enriches the transcript with speaker identification and voice separation, and extracts essential decisions and crucial insights automatically. In our demonstration, the audience can experience the great potential of the state-of-the-art NLP techniques in this real-life application.

ACM Reference Format:
Yuanfeng Song, Di Jiang, Xuefang Zhao, Xiaoling Huang, Qian Xu, Raymond Chi-Wing Wong, Qiang Yang. 2021. SmartMeeting: Automatic Meeting Transcription and Summarization for In-Person Conversations. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3474085.3478556

1 INTRODUCTION

Meetings are a common way for people to communicate, share ideas and reach a common general understanding about tasks and their progress. With the great effort involved for people to manually transcribe and extract crucial insights from meetings, automatic meeting transcription and summarization (AMTS) techniques have drawn great attention in the research community [2, 3, 9, 17]. However, limited work has focused on transferring the advanced AMTS techniques into real-life systems and applications.

In this demonstration, we work towards a practical AMTS system. We present a tool named *SmartMeeting*, which is designed to provide users with a powerful, easy-to-use productive meeting solution. SmartMeeting is a complete meeting system that empowers users to automatically record, transcribe, summarize, and manage the information in an in-person meeting. Through this demonstration, the users will have the ultimate opportunity to experience the mechanisms of how ASR, speaker identification, voice separation, and meeting summarization techniques can be seamlessly integrated with a vivid and interactive approach. We expect that it will show some insights on transferring advanced NLP techniques into real-life applications.

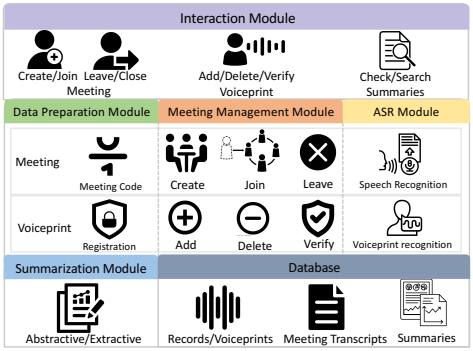


Figure 1: System Overview of SmartMeeting

2 SYSTEM DESCRIPTION

An overview of the SmartMeeting system is shown in Fig. 1. We will discuss its three core components: Transcription by ASR, Transcript Enrichment, and Meeting Summarization.

Transcription by ASR The first step of SmartMeeting is to transcribe the speech recordings made by each participant's devices into their corresponding text. Even though the end-to-end ASR models have drawn great attention in the research community, the hybrid ones with several distinct components - an AM, a PM, and an LM, still dominate the industry due to their modularization and robustness [15]. Hence, in SmartMeeting, we employ the widely-used hybrid ASR pipeline, and the final recognition result $\mathbf{w}^*$ for a given acoustic input $\mathbf{a}$ is defined as follows:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} (\log P_{LM}(\mathbf{w}) + \lambda \log P_{AM}(\mathbf{a}|\mathbf{w})),$$

where P_{LM} is the score given by the LM, P_{AM} is the score given by the AM, and λ is a trade-off parameter. We build the AM using the Kaldi "Chain" model [7], and also trained a trigram LM. To further boost the performance, we also use an self-developed *Learning-to-Rescore* mechanism [10, 11], enhanced with BERT [1], to rescore the N-best list and select the best candidate.

Transcript Enrichment Meeting audio recordings are usually recorded from participants scattered around the same meeting room, resulting in similarly recognized transcripts in different qualities since participants may have different hardware devices, and distances from the microphones. Therefore, the system has to take into account each speaker's recordings while generating transcripts

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10.
https://doi.org/10.1145/3474085.3478556

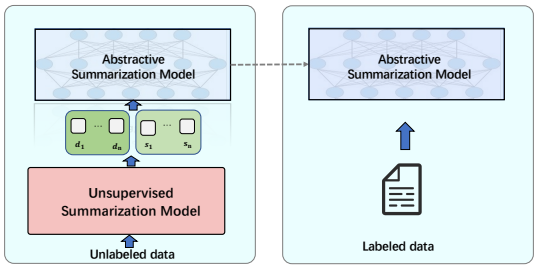


Figure 2: Weakly Supervised Pre-training and Fine-tuning for `WSNeuSummary`

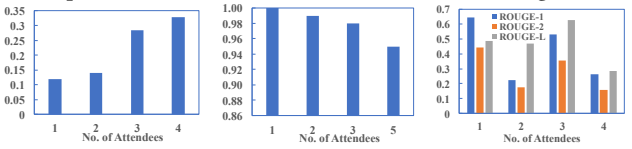
and summaries. The transcript enrichment in SmartMeeting mainly includes four steps: 1) voiceprint-based speakers diarization for each transcript, 2) speaker labeling for each separated utterance, 3) quality evaluation for each utterance, and 4) context selection and merging, and finally the redundant transcripts from different devices/sources are processed into a high-quality transcript. We use a CNN with a self multi-head attention, similar to [8], for voice separation and speaker identification to do the segmentation and clustering based on the speakers' voiceprints.

Meeting Summarization Due to recent advancements in neural networks, DNN-based summarization models have also dominated this area and achieved promising performance [12, 14, 16]. However, these DNN-based summarization models usually assume the availability of large-scale training data, which is hardly possible in our scenario since labeling real-life data is costly and requires great human labor. In SmartMeeting, we propose a novel weakly supervised pre-training mechanism named `WSNeuSummary` to alleviate the above-mentioned problem. As shown in Fig. 2, `WSNeuSummary` works in a two-step strategy: pre-training using weal supervision and fine-tuning with limited labeled instances. Specifically, `WSNeuSummary` 1) adopts an unsupervised summarization model [6] to generate a large volume of weak labels, 2) pre-trains an abstraction summarization model using the weakly-labeled data, and 3) fine-tunes the pre-trained abstraction summarization model with limited human-annotated data. The abstraction summarization model in SmartMeeting is a straightforward transformer-based network [13], which also incorporates BERT [1] to encode the transcripts into hidden representations.

3 PERFORMANCE ANALYSIS

Character Error Rate We examine the effectiveness of SmartMeeting's transcription in terms of CER, and the results are listed in Fig. 3a. CER [4] is the ultimate metric to evaluate the performance of an ASR system for character-based languages such as Chinese, and the lower the value the better the system performance. From the results, we can see that the ASR performance is significantly affected by the number of attendees. With more attendees involved in a meeting, the quality of each recording significantly decreases since people are scattered around the meeting room.

Speaker Attribution Accuracy We also examine the accuracy of the speaker attribution, with the results shown in Fig. 3b. The



(a) Comparison of CER (b) SA Accuracy (c) Comparison of ROUGE

Figure 3: Performance Analysis

(a) Meeting Joining with a Meeting ID (b) Realtime Transcription and Summarization (c) Summarization Browsing and Searching

Figure 4: The User Interfaces of the SmartMeeting System

speaker attribution relies on the user identification technique with the voiceprint of the users after conducting the voice separation. SmartMeeting generally achieves a relatively high accuracy (around 90%) for speaker attribution.

ROUGE Fig. 3c shows the ROUGE-1, ROUGE-2, and ROUGE-L scores of generated summaries with a different number of meeting attendees. ROUGE scores are a series of widely accepted metrics to evaluate the quality of machine-generated summaries [5]. Generally speaking, the higher the ROUGE value, the better the quality of the summaries. The results show a gradual decline in performance with more meeting attendees involved, which is consistent with the general common understanding that more members result in a much more complicated meeting content.

4 DEMONSTRATION OVERVIEW

Fig. 4 gives a series of screenshots of the SmartMeeting application.

User Registration If this is the first time a user uses the system, SmartMeeting will require the user to grant access to the microphone and also do the registration with his/her voiceprint and a user name. The voiceprint is used by SmartMeeting to identify each meeting participant and conduct voice separation.

Meeting Management Management functionalities include meeting creation, starting, joining, ending, and deletion. A coordinator creates or schedules each meeting with a meeting ID, and each attendee can join this meeting using the same ID (Fig. 4a). The coordinator has the right to start, end the meeting, and manage attendees, while each attendee can join and leave the meeting.

Real-time Transcription After the meeting starts, the recording and transcription happen in the background. As shown in Fig. 4b, each user can view the real-time transcripts, enriched by speaker identification and voice separation. Each utterance, separated from the transcripts, is also labeled with the user name of the speaker. The users can easily verify or correct the text if needed. The transcripts are stored only on the user's device for privacy concerns.

Meeting Summarization Each user can view and edit the summarization after a meeting. The summarization identifies tasks and classifies insights. The users can easily edit or correct the text and synchronize the summarization with all the attendees.

Summarization Searching SmartMeeting also supports the users to browse through and search across all the historical meeting highlights by keywords (Fig. 4c).

Acknowledgement We thank anonymous reviewers for their helpful comments. The research of Raymond Chi-Wing Wong is supported by GZETDZ18EG06.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*.

[2] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. 369–376.

[3] Som Gupta and SK Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications* 121 (2019), 49–65.

[4] Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication* 38, 1-2 (2002), 19–28.

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

[6] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. 404–411.

[7] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for ASR based on lattice-free MMI.. In *Interspeech*. 2751–2755.

[8] Pooyan Safari and Javier Hernando. 2019. Self multi-head attention for speaker recognition. *arXiv preprint arXiv:1906.09890* (2019).

[9] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368* (2017).

[10] Yuanfeng Song, Di Jiang, Xiaoling Huang, Yawen Li, Qian Xu, Raymond Chi-Wing Wong, and Qiang Yang. 2020. GoldenRetriever: A Speech Recognition System Powered by Modern Information Retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4500–4502.

[11] Yuanfeng Song, Di Jiang, Xuefang Zhao, Qian Xu, Raymond Chi-Wing Wong, Lixin Fan, and Qiang Yang. 2019. L2RS: A Learning-to-Rescore Mechanism for Automatic Speech Recognition. *arXiv preprint arXiv:1910.11496* (2019).

[12] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1171–1181.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010.

[14] Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuan-Jing Huang. 2020. Heterogeneous Graph Neural Networks for Extractive Document Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6209–6219.

[15] Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, et al. 2020. Transformer-based acoustic modeling for hybrid speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6874–6878.

[16] Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural Latent Extractive Document Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 779–784.

[17] Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. End-to-End Abstractive Summarization for Meetings. *arXiv preprint arXiv:2004.02016* (2020).