The Second Automatic Minuting (*AutoMin*) Challenge: Generating and Evaluating Minutes from Multi-Party Meetings

Tirthankar Ghosal[*], Marie Hledíková[*], Muskaan Singh[$], Anna Nedoluzhko[*], Ondřej Bojar[*]
[*]Charles University, Faculty of Mathematics and Physics, ÚFAL, Czech Republic
[$]Speech and Audio Processing Group, IDIAP Research Institute, Martigny, Switzerland
(ghosal,hledikova,nedoluzhko,bojar)@ufal.mff.cuni.cz, msingh@idiap.ch

Abstract

We would host the *AutoMin* generation challenge at INLG 2023 as a follow-up of the first AutoMin shared task at Interspeech 2021. Our shared task primarily concerns the automated generation of meeting minutes from multi-party meeting transcripts. In our first venture, we observed the difficulty of the task and highlighted a number of open problems for the community to discuss, attempt, and solve. Hence, we invite the Natural Language Generation (NLG) community to take part in the second iteration of AutoMin. Like the first, the second AutoMin will feature both English and Czech meetings and the core task of summarizing the manually-revised transcripts into bulleted minutes. A *new* challenge we are introducing this year is to devise efficient metrics for evaluating the *quality* of minutes. We will also host an optional track to generate minutes for European parliamentary sessions.

We carefully curated the datasets for the above tasks. Our ELITR Minuting Corpus has been recently accepted to LREC 2022 and publicly released.[1] We are already preparing a new test set for evaluating the new shared tasks. We hope to carry forward the learning from the first AutoMin and instigate more community attention and interest in this timely yet challenging problem. INLG, the premier forum for the NLG community, would be an appropriate venue to discuss the challenges and future of *Automatic Minuting*. The main objective of the AutoMin GenChal at INLG 2023 would be to come up with efficient methods to automatically generate meeting minutes and design evaluation metrics to measure the quality of the minutes.

1 Introduction

Ever since most of our interactions went virtual, the need for automatic support to run online meetings became essential. Due to frequent meetings and the resulting context switching, people are experiencing an information overload (Fauville et al., 2021) of epic proportions. Hence a tool to automatically summarize a meeting proceeding would be a valuable addition to the virtual workplace. Automatic minuting (Nedoluzhko and Bojar, 2019) is close to summarization; however, there are subtle differences. While summarization is motivated towards generating a concise and coherent summary of the text, minuting is more inclined towards adequately capturing the contents of the meeting (*where coverage is probably more significant than coherence and conciseness*). Summarizing spoken multi-party dialogues (Bhattacharjee et al., 2022) comes with its own challenges: incorrect/noisy automated speech recognition (ASR) outputs, long discourse, topical shifts, the dialogue turns, redundancies and small talk, etc. Hence we deem automatic minuting to be more difficult than text summarization (Figure 2 in Appendix A shows an envisaged demonstration of the task).

With the *AutoMin* challenge, we want to explore the various problems associated with the task and their potential solutions from the perspective of a multi-year joint community initiative. Apart from the main task of summarizing meeting transcripts into concise, bulleted minute items, another crucial task is to develop efficient evaluation measures to judge the quality of the automatically generated minutes. It is a known fact that the current popular methods of automatic summarization evaluation (e.g., ROUGE (Lin, 2004)) do not guarantee critical quality parameters like *fluency, adequacy, grammatical correctness, etc.* (Ghosal et al., 2021a,b), which is why we have to primarily rely on human evaluation metrics in our shared task. The proposed instance of the AutoMin challenge will venture into developing automatic/semi-automatic evaluation metrics to measure the *quality* of generated minutes. Summarizing the participants' ideas for this challenge and the anticipated follow-up dis-

[1] http://hdl.handle.net/11234/1-4692

Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges, pages 1 - 11
July 17-22, 2022 ©2022 Association for Computational Linguistics

cussions, we will try to define an ideal meeting summary. Since the task suffers from resource scarcity, we would launch an initiative where interested parties could donate their meetings to prepare a public multimodal, multilingual dataset of real meetings.

2 First AutoMin @ Interspeech 2021

The AutoMin[2] shared task at Interspeech 2021 (Ghosal et al., 2021a) was a first of its kind with this problem. It generated considerable interest in the speech and NLP community. Twenty-seven teams from diverse geographical regions registered, and finally, ten teams (both from academia and industry) actively participated in the challenge. Almost 70 people attended the shared task virtual event. The first AutoMin consisted of one main task (Task A) and two supporting tasks (Task B and Task C), relying on a dataset of transcripts and minutes from primarily technical meetings in English and Czech (Nedoluzhko et al., 2022).

Considering the current non-availability of large-scale domain datasets on multiparty meeting summarization, the best recipe for automatic minuting that evolved out from the first AutoMin is roughly the following: training a deep neural model on available dialogue summarization datasets (SAMSum (Gliwa et al., 2019); DialSum (Chen et al., 2021); etc.) and further fine-tuning it on the minuting or meeting summarization datasets (AMI (Mccowan et al., 2005); ICSI (Janin et al., 2003); AutoMin (Ghosal et al., 2021a)), accompanied by some intelligent pre- and post-processing steps.

3 Task Overview

We would continue with the tasks in the previous AutoMin challenge in the current iteration. However, the new additions would be: (1) *automatically generating the meeting minutes of parliamentary sessions* as part of Task A, and (2) *designing appropriate evaluation schema/metrics to evaluate the generated minutes* as a new Task D.

3.1 Task A

*The **main task** consists of automatically generating minutes from multiparty meeting conversations* provided in the form of transcripts. The objective is to generate minutes as bulleted lists, summarizing the main contents of the meeting, as opposed to usual paragraph-like text summaries. This task would run for the meetings in the ELITR Minuting Corpora (Nedoluzhko et al., 2022) and the *new data we curated from the European parliamentary sessions*.[3] Note that the nature of meetings as well as the corresponding minutes are very different in the two datasets (technical project meetings vs. parliamentary sessions).

3.2 Task B

Given a pair of a meeting transcript and a manually-created minute, the task is to identify whether the minute belongs to the transcript.

During our data preparation from meetings on similar topics, we found that this task could be challenging due to the similarity of the discussed content and anchor points like named entities, e.g., in recurring meetings of the same project on the one hand, and the differences in the style of minuting, on the other hand. Another reason is that some minutes do not capture the central points in the meeting because the external scribes did not understand the context correctly and created minutes that miss significant issues discussed in the meeting or are simply too short.

3.3 Task C

Task C is a variation of Task B. *Given a pair of minutes, the task is to identify whether the two minutes belong to the same meeting or to two different ones.* This task is important as we want to uncover how minutes created by two different persons for the same meeting may differ in content and coverage.

3.4 Task D (New Task)

Given a meeting transcript, a candidate minute, and a set of one or more reference minutes, assign a score indicating the quality of the candidate minute.

The participating evaluation methods can focus on diverse aspects of minutes quality, such as the coverage of content discussed, the adequacy of the description, the readability, etc. We will evaluate the submitted scores with respect to correlation with human judgements in terms of *adequacy*, *fluency* and *grammatical correctness* from AutoMin 2021 human evaluations, and possibly in terms of additional criteria.

In other words, there is not a single evaluation criterion for submissions to Task D. Task D should

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

[3] https://emeeting.europarl.europa.eu/emeeting/committee/en/archives

	English		Czech	
Meeting Minuted	#meetings	#hours	#meetings	#hours
Once	30	22	8	2
Twice	65	65	20	20
More than twice	25	22	31	31
Total meetings	120	109	59	53

Table 1: Basic transcript and minutes statistics for ELITR Minuting Corpus.

Language	English	Czech
# of Meetings	120	59
avg. words per transcript	7,066	8,534
avg. words per summary	373	236
avg. turns per transcript	727	1,205
avg. number of speakers	5.9	7.6

Table 2: Text statistics of ELITR Minuting Corpus.

be treated as a joint exploration rather than an optimization exercise.

4 Dataset Description

We provide the AutoMin 2023 participants with ELITR Minuting Corpus (Nedoluzhko et al., 2022); however, we would allow them to use any external datasets if they explicitly describe them in their system reports.

4.1 ELITR Minuting Corpus for Task A

In our ELITR Minuting Corpus (Nedoluzhko et al., 2022), a meeting usually contains one manually corrected transcript, one original minute (created by a meeting participant; in some cases, these minutes are a detailed agenda which got further updated during or after the meeting), and one or more generated minutes (by annotators).

Table 1 presents our dataset's statistics regarding the number of meetings and hours. We separately count meetings for which we have only one, two, and more than two (up to 11) minutes. For English meetings, either (i) our annotators created both minutes or (ii) one minute was written by one of the participants before or after the meeting and another by our annotator. In contrast, most meetings in the Czech portion of the dataset are minuted at least twice, and more than half of the Czech portion of ELITR Minuting Corpus is minuted 3-5 times.

To address GDPR issues (privacy of participants), we de-identify any information concerning Person, Organisation, Project and Location (in specific cases) names were replaced with the lexical substitute strings [PERSON*number*], [ORGANIZATION*number*], [PROJECT*number*] and [LOCATION*number*] respectively. Additionally, we replaced the names of annotators mentioned in minutes with [ANNOTATOR*number*].

Table 2 reports summary statistics of the texts in ELITR Minuting Corpus and Figure 4 in Appendix B shows a sample minute from the corpus.

4.2 EuroParlMin for Task A

We curate this dataset from the publicly available European parliamentary sessions by using the transcripts in the EuroParl dataset (Koehn, 2005) and crawling the corresponding minutes from the EU parliament website.[4]

We automatically create a set of transcript–minute pairs (∼2000). This dataset is new, and we would make this available to the shared task participants.

4.3 Test Data for Task D

There is no training data for Task D (except training data available for Tasks A–C anyway).

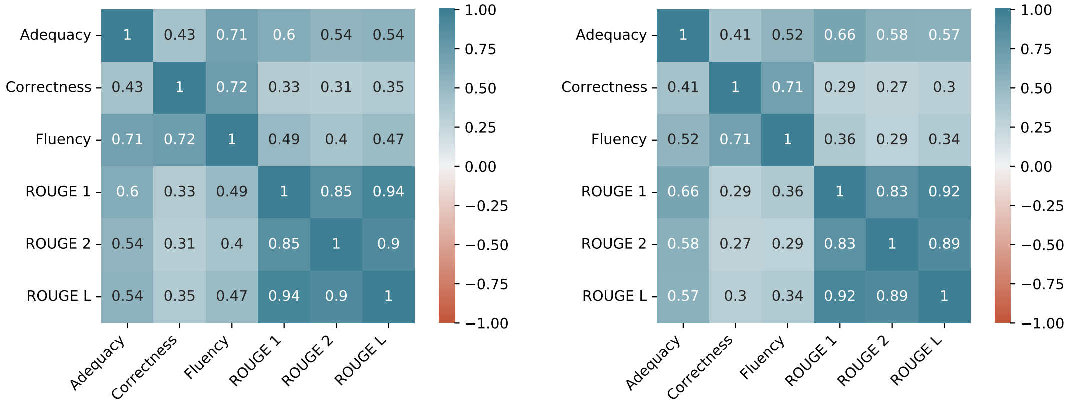
The test data for Task D consists of participants' submissions to AutoMin 2021. Our human evaluators rated each submitted minute by the ten different participating teams (some had multiple submission runs) in three criteria: *adequacy*, *fluency*, and *grammatical correctness* on the test set. Additionally, we plan to design some methods of minute scoring based on the (manual) alignments between the transcript and the minute Polák et al. (2022). These alignments are included in ELITR Minuting Corpus for many of the meetings and their manually created minutes, which can be used as training data. We will also prepare these alignments for AutoMin 2021 submissions, i.e., automatic minutes.

We will use these annotations (*adequacy*, *fluency*, *grammatical correctness* and different scores based on the alignments) as different possible ground truth values for participants in Task D.

It is up to the participants of Task D to propose which type of criterion their metric will focus on. We will evaluate each submission against all available ground truths.

We prepared data for Task B and Task C from ELITR Minuting Corpus (leaving the meetings we selected to run AutoMin 2021).

[4] https://www.europarl.europa.eu/committees/en/meetings/minutes



(a) Average across multiple references.

(b) Maximum across multiple references.

Figure 1: Correlations of metrics (human and automatic) used in AutoMin 2021 across all participants. Each cell represents the Pearson correlation between the two types of measurements of a given meeting. With multiple reference minutes, the automatic scores are aggregated with (a) average and (b) maximum. Two independent judges assigned manual scores, and to arrive at a single score per minute, we again aggregated them with average or maximum.

5 Evaluation Campaign

5.1 Human Evaluation

We will perform human evaluation on the submissions in Task A (both English and Czech) with the usual metrics: *adequacy, fluency, relevance, and grammatical correctness* (Kryscinski et al., 2020) on a Likert scale of 1-5.

1. **Adequacy** assesses if the minute adequately captures the major topics discussed in the meeting, also considering coverage (all such topics reflected).
2. **Fluency** reflects if the minute consists of fluent, coherent texts and is readable to the evaluator.
3. **Grammatical Correctness** checks the level to which the minute is grammatically consistent.
4. **Relevance** signifies whether the important content from the source transcript appear in the candidate minutes.

Along with that, we will launch a pilot evaluation of the submitted minutes via our ALIGNMEET tool (Polák et al., 2022). An alignment maps each turn of the transcript to either one line of the minute's file in which it is summarized, a "problem" label, both or neither. The alignments are done in such a way that whole discussions are aligned to the minutes lines (e.g., speaker A agreeing to a statement by speaker B is aligned to the same minutes line as speaker B's original statement; see Figure 3 for an example of an alignment in Appendix A).

There will be no manual evaluation for Tasks B, C, and D.

5.2 Automatic Evaluation

For our automatic evaluation of Task A, we will still rely on the widely popular text summarization metric ROUGE (Lin, 2004) in its three variants: ROUGE-1, ROUGE-2, ROUGE-L. Additionally, we will use BERTScore (Zhang et al., 2019) and/or BARTScore (Yuan et al., 2021) which are currently being used to evaluate generation tasks.

For Tasks B and C, which are actually classification tasks, we will use accuracy and class-wise F1 scores.

Task D will not be evaluated by a single criterion. As mentioned above, all submissions to Task D will be evaluated in terms of Pearson correlation against all manual and all other automatic evaluation scores.

Figure 1 plots the heatmaps of Pearson correlations between various types of evaluations of minutes. For automatic scores (ROUGE variants), we utilize multiple reference minutes, where available, and average or maximize over them. For manual scores (Adequacy, Correctness, and Fluency), we average or maximize the score assigned by two annotators to get a single score for a given minute.

We see that all ROUGE score types are significantly correlated with each other but not much related to the manual scores. The highest correlation is between ROUGE 1 and Adequacy in the (b) plot in Figure 1, reaching 0.66, which is approximately the same level of correlation as between Correctness and Fluency. Any variants of the automatic score do not reflect Correctness and Fluency. Figure 5 in Appendix C shows one of the good minutes generated by a participating team in the First AutoMin shared task.

6 Baseline Evaluations

We provide our participants with the baseline codes for automatic minuting (Task A) here.[5] The details of the experiments are described in Singh et al. (2021). It includes initial exploration using *off-the-shelf* text summarization models for future investigations. For generating abstractive meeting minutes we use BART (Lewis et al., 2019), BERTSUM (Liu and Lapata, 2019), BERT2BERT (Rothe et al., 2020), LED (Beltagy et al., 2020), Pegasus (Zhang et al., 2020), Roberta2Roberta (Liu et al., 2019), and T5 (Raffel et al., 2019) models. For extractive meeting summaries we use *TF-IDF*-based summarizer (Christian et al., 2016), an unsupervised extractive summarizer, TextRank (Mihalcea and Tarau, 2004), LexRank (Erkan and Radev, 2004), Luhn Algorithm (Luhn, 1958), and LSA (Gong and Liu, 2001) based summarizer. These off-the-shelf text summarization models are not the best candidates for generating minutes which calls for further research on this challenging task for meeting-specific summarization or minuting models.

7 AutoMin 2023 Procedure and Timeline

Table 3 summarizes the tentative timeline for AutoMin 2023. We would create and host a server to handle the shared task system submissions. We would use START or EasyChair for paper submissions and reviewing. We would also set up a program committee to review the system submissions and help the authors improve their reports.

8 Diversity and Inclusion

As our commitment to diversity and inclusion, like the previous iteration, we would like to make our event *open-to-all* (and possibly hybrid) in consultation with INLG 2023 chairs. We would especially reach out to organizations like Widening NLP[6] (where our first organizer is also a chair) to help us reach the underrepresented groups and communities and encourage them to participate. We would also look for funding from industries/labs interested in the application of this research to sponsor resources (especially compute) and/or travel/registration of our participants in need of those logistics.

July 2022	Announcement at INLG 2022
August 2022	Call for Participation
September 2022	Training Data Release
December 2023	Test Data Release
February 2023	System Submission
March 2023	Evaluation Notification
April 2023	System Report Submission
May 2023	System Report Review Notification
May 2023	Camera-ready Submission
June 2023	Proceedings appear in ACL Anthology
July 2023	Second AutoMin at INLG 2023

Table 3: Tentative Timeline for second AutoMin at INLG 2023 (may change depending on INLG 2023 schedule)

9 Conclusion

AutoMin is a very timely yet complex task for the speech and natural language processing community. Given the array of problems this task had to offer, we are very excited to continue this iteration of the generation challenge at INLG 2023. We look forward to uncovering the several linguistic phenomena and insights that should go into action while a machine writes a minute and see how much we have progressed towards an acceptable automated minuting output. In that essence, Task A and Task D are of more interest to the NLG and summarization community than Task B and Task C.

10 Ethical Considerations

For our ELITR Minuting Corpus, all meeting participants consented to make the data publicly available. Please refer to Nedoluzhko et al. (2022); Ghosal et al. (2021a) for a detailed description of our de-identification and participant-consent process. We would follow the same conditions to prepare the hidden test set. The EuroParl (Koehn, 2005) data, as well as the minutes for those parliamentary sessions, is publicly available (on the EuroParl website). Hence, there should not be any privacy or ethical issues.

Acknowledgement

The shared task would receive funding from the grant 19-26934X (NEUREM3) of the Czech Science Foundation.

[5] https://github.com/ELITR/Minuting_Baseline_Experiments

[6] http://www.winlp.org

References

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

Saprativa Bhattacharjee, Kartik Shinde, Tirthankar Ghosal, and Asif Ekbal. 2022. A multi-task learning approach for summarization of dialogues. In *Proceedings of the 15th International Conference on Natural Language Generation*, page TBA, Maine, US. Association for Computational Linguistics.

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

Hans Christian, Mikhael Pramodana Agus, and Derwin Suhartono. 2016. Single document automatic text summarization using term frequency-inverse document frequency (tf-idf). *ComTech: Computer, Mathematics and Engineering Applications*, 7(4):285–294.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

G. Fauville, M. Luo, A.C.M. Queiroz, J.N. Bailenson, and J. Hancock. 2021. Zoom exhaustion & fatigue scale. *Computers in Human Behavior Reports*, 4:100119.

Tirthankar Ghosal, Ondřej Bojar, Muskaan Singh, and Anja Nedoluzhko. 2021a. 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*, pages 1–25.

Tirthankar Ghosal, Muskaan Singh, Anna Nedoluzhko, and Ondřej Bojar. 2021b. Report on the sigdial 2021 special session on summarization of dialogues and multi-party meetings (summdial). *ACM SIGIR Forum*, December 2021:1–17.

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

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The icsi meeting corpus. pages 364–367.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MTSUMMIT*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

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

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 2005. The ami meeting corpus. In *In: Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research. L.P.J.J. Noldus, F. Grieco, L.W.S. Loijens and P.H. Zimmerman (Eds.), Wageningen: Noldus Information Technology*.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Anna Nedoluzhko and Ondrej Bojar. 2019. Towards automatic minuting of the meetings. In *ITAT*.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. Elitr minuting corpus: A novel dataset for automatic minuting from multi-party meetings in english and czech. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3174–3182, Marseille, France. European Language Resources Association.

Peter Polák, Muskaan Singh, Anna Nedoluzhko, and Ondřej Bojar. 2022. Alignmeet: A comprehensive tool for meeting annotation, alignment, and evaluation. In *Proceedings of The 13th Language Resources and Evaluation Conference*, page To Appear.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Kartik Shinde, Nidhir Bhavsar, Aakash Bhatnagar, and Tirthankar Ghosal. 2021. Team ABC @ AutoMin 2021: Generating Readable Minutes with a BART-based Automatic Minuting Approach. In *Proc. First Shared Task on Automatic Minuting at Interspeech 2021*, pages 26–33.

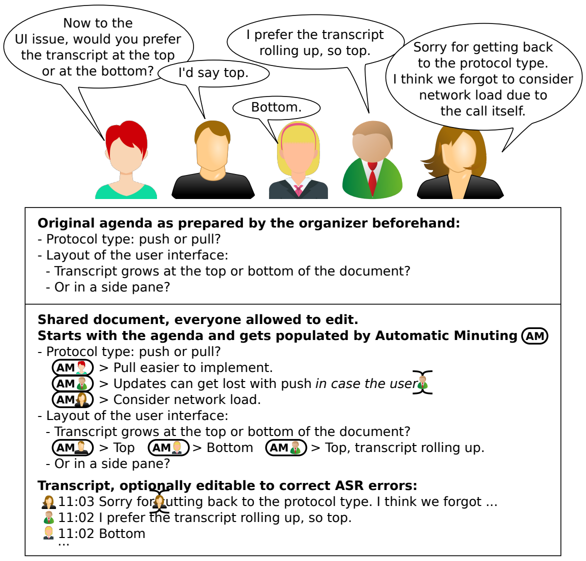
Muskaan Singh, Tirthankar Ghosal, and Ondrej Bojar. 2021. An empirical performance analysis of state-of-the-art summarization models for automatic minuting. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 50–60, Shanghai, China. Association for Computational Lingustics.

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

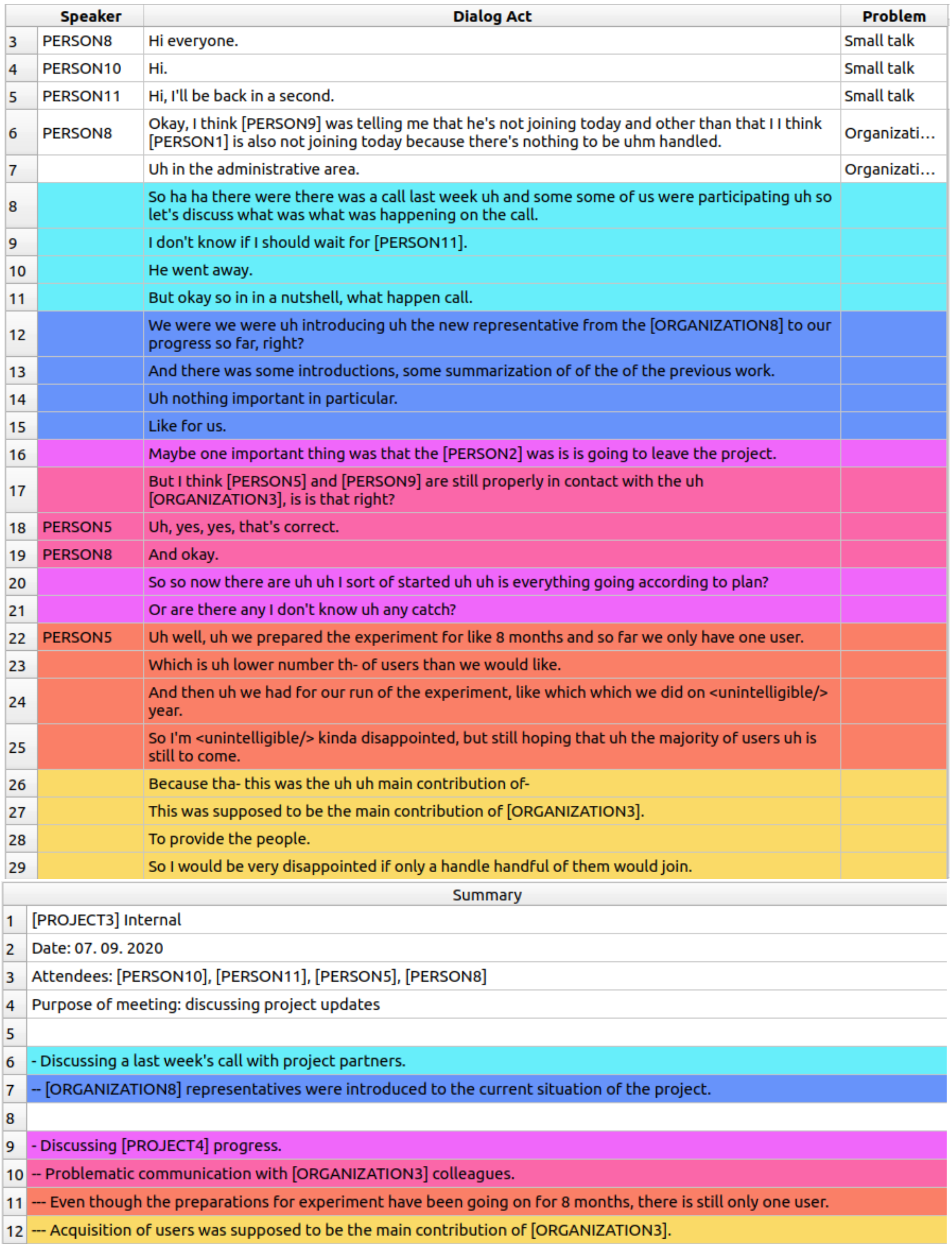

Now to the UI issue, would you prefer the transcript at the top or at the bottom?
I'd say top.
Bottom.
I prefer the transcript rolling up, so top.
Sorry for getting back to the protocol type. I think we forgot to consider network load due to the call itself.

Original agenda as prepared by the organizer beforehand:
- Protocol type: push or pull?
- Layout of the user interface:
  - Transcript grows at the top or bottom of the document?
  - Or in a side pane?

Shared document, everyone allowed to edit.
Starts with the agenda and gets populated by Automatic Minuting AM
- Protocol type: push or pull?
  AM > Pull easier to implement.
  AM > Updates can get lost with push *in case the user*
  AM > Consider network load.
- Layout of the user interface:
  - Transcript grows at the top or bottom of the document?
  AM > Top AM > Bottom AM > Top, transcript rolling up.
  - Or in a side pane?

Transcript, optionally editable to correct ASR errors:
11:03 Sorry for utting back to the protocol type. I think we forgot ...
11:02 I prefer the transcript rolling up, so top.
11:02 Bottom
...

Figure 2: Envisioning Automatic Minuting



	Speaker	Dialog Act	Problem
3	PERSON8	Hi everyone.	Small talk
4	PERSON10	Hi.	Small talk
5	PERSON11	Hi, I'll be back in a second.	Small talk
6	PERSON8	Okay, I think [PERSON9] was telling me that he's not joining today and other than that I I think [PERSON1] is also not joining today because there's nothing to be uhm handled.	Organizati...
7		Uh in the administrative area.	Organizati...
8		So ha ha there were there was a call last week uh and some some of us were participating uh so let's discuss what was what was happening on the call.	
9		I don't know if I should wait for [PERSON11].	
10		He went away.	
11		But okay so in in a nutshell, what happen call.	
12		We were we were uh introducing uh the new representative from the [ORGANIZATION8] to our progress so far, right?	
13		And there was some introductions, some summarization of of the of the previous work.	
14		Uh nothing important in particular.	
15		Like for us.	
16		Maybe one important thing was that the [PERSON2] was is is going to leave the project.	
17		But I think [PERSON5] and [PERSON9] are still properly in contact with the uh [ORGANIZATION3], is is that right?	
18	PERSON5	Uh, yes, yes, that's correct.	
19	PERSON8	And okay.	
20		So so now there are uh uh I sort of started uh uh is everything going according to plan?	
21		Or are there any I don't know uh any catch?	
22	PERSON5	Uh well, uh we prepared the experiment for like 8 months and so far we only have one user.	
23		Which is uh lower number th- of users than we would like.	
24		And then uh we had for our run of the experiment, like which which we did on <unintelligible/> year.	
25		So I'm <unintelligible/> kinda disappointed, but still hoping that uh the majority of users uh is still to come.	
26		Because tha- this was the uh uh main contribution of-	
27		This was supposed to be the main contribution of [ORGANIZATION3].	
28		To provide the people.	
29		So I would be very disappointed if only a handle handful of them would join.	

	Summary
1	[PROJECT3] Internal
2	Date: 07. 09. 2020
3	Attendees: [PERSON10], [PERSON11], [PERSON5], [PERSON8]
4	Purpose of meeting: discussing project updates
5	
6	- Discussing a last week's call with project partners.
7	-- [ORGANIZATION8] representatives were introduced to the current situation of the project.
8	
9	- Discussing [PROJECT4] progress.
10	-- Problematic communication with [ORGANIZATION3] colleagues.
11	--- Even though the preparations for experiment have been going on for 8 months, there is still only one user.
12	--- Acquisition of users was supposed to be the main contribution of [ORGANIZATION3].

Figure 3: Example of an alignment viewed in ALIGNMEET. Dialogue Acts with white background are not aligned to minutes, other colors indicate alignment to minutes line of the same color. Problems are shown in the right column of the transcript view.

B Sample Reference Minutes in ELITR Minuting Corpus

```
Date: 2019/04/01
Attendees: [PERSON10], [PERSON2], [PERSON3], [PERSON7], [PERSON11], [PERSON8], [
    PERSON1]
Purpose of meeting: Technical prepare for [ORGANIZATION6] congress

Agenda:
- Start recording.
- Date for [PROJECT1] call.
- Collecting photos and videos from Trade Fair.
- Confirmation of proposed scheme of wiring for [ORGANIZATION6] Congress.
- Digital interface to audio mix pult.
- Microphones.
- Get a contact for someone from [ORGANIZATION4], who will handle the presentation
    platform.
- Will [ORGANIZATION4] also try get their ASR.
- When will the python version of [ORGANIZATION4] platform sample connector.

Summary of meeting:

[PERSON3], [PERSON7]:
- After reminder missing vote for [PROJECT1] call date was chosen the April 16th.

[PERSON3], [PERSON7]:
- Ask for photos from the trade fair. Will be sent to e-mail immediately.

[PERSON3], [PERSON7], [PERSON11]:
- It is needed to specify the settings for workshop in June and [ORGANIZATION6]
    congress.
The hardware will provide outside company.
It is supposed to translating and transcribing the main session.
There will be rented tablets and is supposed that everyone will have their cell
    phones.
It is needed to connect the microphones to the mean audio mixer and then to have
    digital output to the booth for listening and ASR.
Any of the separate notebooks after the ASR can provide input to the multilingual
    translation system.
Proposal that every input language has uhm have to have its own ehm session with the
     mediator, this will be implemented by [PERSON2].
It is needed original sound from the microphones as possible from booth main
    microphone of the plenary session, ideally the digital signal captured at
    microphone.
Languages: English, German, Czech, French, Italian, Spanish, Russian.
There is experience only with Dante, but it is very expensive and doesn't simplify
    setting.
It is needed one PC for each language, one PC per input channel.
It is recomended to keep audio data and network traffic separated.
Will be demand one direct microphone output from the main microphone.
And one direct microphone output from each of the booths and for these booth
    microphones we demand that only the predefined languages is spoken at that
    channel.
Proposal to say get booth analog output as a call back and digital interface scholar
     choice.
[ORGANIZATION4] will let know what digital audio should be specify in the
    documentation until Tuesday.

[PERSON3], [PERSON11], [PERSON7]:
- It is needed to demand also Microphones.
Ask for definition all the individual microphones that the speakers will use.
After discussion they agreed that there will be preferred wired microphone for main
    stage.
Until Tuesday [PERSON7] will provide specification for main stage wired microphones
    and interpreters booths large microphones and also for wireless.

[PERSON3], [PERSON7], [PERSON11]:
- Presentation platform will have to be different for the workshop in June and for
    the [ORGANIZATION6] congress, because the setting is different.
Explain idea.
[PERSON2] will be coding this thing.
```

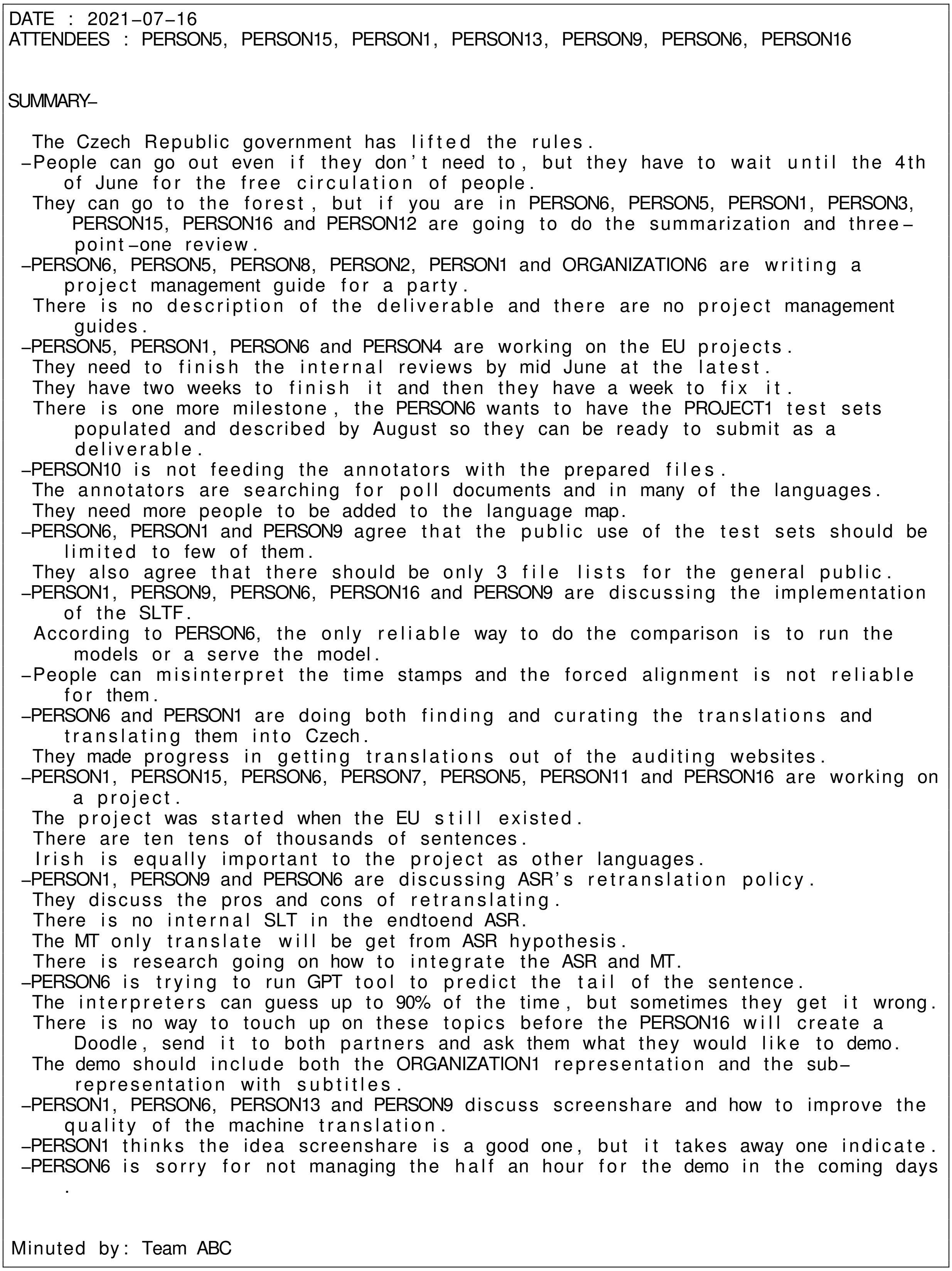
```
[PERSON3], [PERSON7]:
- [ORGANIZATION4] won't try their own ASR.

[PERSON3], [PERSON7]:
- Ask when the python connector to the [ORGANIZATION4] platform would be ready.
People using python at the [ORGANIZATION8] will help with this point.
It will be published at public website.

Minutes submitted by: [ANNOTATOR1]
```

Figure 4: A sample minute taken by our external annotators

C Sample Candidate Minutes by one participating team in AutoMin 2021



```
DATE : 2021-07-16
ATTENDEES : PERSON5, PERSON15, PERSON1, PERSON13, PERSON9, PERSON6, PERSON16


SUMMARY-

  The Czech Republic government has lifted the rules.
 -People can go out even if they don't need to, but they have to wait until the 4th
    of June for the free circulation of people.
  They can go to the forest, but if you are in PERSON6, PERSON5, PERSON1, PERSON3,
     PERSON15, PERSON16 and PERSON12 are going to do the summarization and three-
     point-one review.
 -PERSON6, PERSON5, PERSON8, PERSON2, PERSON1 and ORGANIZATION6 are writing a
    project management guide for a party.
  There is no description of the deliverable and there are no project management
     guides.
 -PERSON5, PERSON1, PERSON6 and PERSON4 are working on the EU projects.
  They need to finish the internal reviews by mid June at the latest.
  They have two weeks to finish it and then they have a week to fix it.
  There is one more milestone, the PERSON6 wants to have the PROJECT1 test sets
     populated and described by August so they can be ready to submit as a
     deliverable.
 -PERSON10 is not feeding the annotators with the prepared files.
  The annotators are searching for poll documents and in many of the languages.
  They need more people to be added to the language map.
 -PERSON6, PERSON1 and PERSON9 agree that the public use of the test sets should be
    limited to few of them.
  They also agree that there should be only 3 file lists for the general public.
 -PERSON1, PERSON9, PERSON6, PERSON16 and PERSON9 are discussing the implementation
    of the SLTF.
  According to PERSON6, the only reliable way to do the comparison is to run the
     models or a serve the model.
 -People can misinterpret the time stamps and the forced alignment is not reliable
    for them.
 -PERSON6 and PERSON1 are doing both finding and curating the translations and
    translating them into Czech.
  They made progress in getting translations out of the auditing websites.
 -PERSON1, PERSON15, PERSON6, PERSON7, PERSON5, PERSON11 and PERSON16 are working on
     a project.
  The project was started when the EU still existed.
  There are ten tens of thousands of sentences.
  Irish is equally important to the project as other languages.
 -PERSON1, PERSON9 and PERSON6 are discussing ASR's retranslation policy.
  They discuss the pros and cons of retranslating.
  There is no internal SLT in the endtoend ASR.
  The MT only translate will be get from ASR hypothesis.
  There is research going on how to integrate the ASR and MT.
 -PERSON6 is trying to run GPT tool to predict the tail of the sentence.
  The interpreters can guess up to 90% of the time, but sometimes they get it wrong.
  There is no way to touch up on these topics before the PERSON16 will create a
     Doodle, send it to both partners and ask them what they would like to demo.
  The demo should include both the ORGANIZATION1 representation and the sub-
     representation with subtitles.
 -PERSON1, PERSON6, PERSON13 and PERSON9 discuss screenshare and how to improve the
    quality of the machine translation.
 -PERSON1 thinks the idea screenshare is a good one, but it takes away one indicate.
 -PERSON6 is sorry for not managing the half an hour for the demo in the coming days
    .


Minuted by: Team ABC
```

Figure 5: A sample minute from Team ABC (Shinde et al., 2021) in AutoMin 2021 (Ghosal et al., 2021a)