This document is part of the Research and Innovation Action "European Live Translator (ELITR)". This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460.

Deliverable D5.2

Final Report on Summarization

Tirthankar Ghosal (CUNI), Marie Hledíková (CUNI), Ondřej Bojar (CUNI), Muskaan Singh (CUNI)

Dissemination Level: Public

Final (Version 1.0), 31st March, 2022

Grant agreement no.	825460
Project acronym	ELITR
Project full title	European Live Translator
Type of action	Research and Innovation Action
Coordinator	doc. RNDr. Ondřej Bojar, PhD. (CUNI)
Start date, duration	1st January, 2019, 39 months
Dissemination level	Public
Contractual date of delivery	31st March, 2022
Actual date of delivery	31st March, 2022
Deliverable number	D5.2
Deliverable title	Final Report on Summarization
Type	Report
Status and version	Final (Version 1.0)
Number of pages	101
Contributing partners	CUNI
WP leader	PV
Author(s)	Tirthankar Ghosal (CUNI), Marie Hledíková (CUNI), Ondřej Bojar (CUNI), Muskaan Singh (CUNI)
EC project officer	Luis Eduardo Martinez Lafuente
The partners in ELITR are:	▪ Univerzita Karlova (CUNI), Czech Republic ▪ University of Edinburgh (UEDIN), United Kingdom ▪ Karlsruher Institut für Technologie (KIT), Germany ▪ PerVoice SPA (PV), Italy ▪ alfatraining Bildungszentrum GmbH (AV), Germany
Partially-participating party	▪ Nejvyšší kontrolní úřad (SAO), Czech Republic

For copies of reports, updates on project activities and other ELITR-related information, contact:

doc. RNDr. Ondřej Bojar, PhD., ÚFAL MFF UK
Malostranské náměstí 25
118 00 Praha, Czech Republic
bojar@ufal.mff.cuni.cz
Phone: +420 951 554 276
Fax: +420 257 223 293

Copies of reports and other material can also be accessed via the project's homepage:

`http://www.elitr.eu/`

Contents

1 Executive Summary

This deliverable reports on our progress in WP5: Automatic Minuting module. We describe the associated research and activities carried out during the second (i.e., concluding) reporting period of the ELITR project for realizing the WP5 objectives.

By "automatic minuting", we mean the task of automatically obtaining meeting minutes. The characteristics of the task are detailed and illustrated in the rest of this deliverable and the attached papers in various stages of the reviewing and publication process.

In our initial proposal, the Automatic Minuting work package was structured into the following sub-tasks:

- T5.1: Meeting Segmentation
- T5.2: Segment-Level Summarization
- T5.3: Document-Level Summarization
- T5.4: Sequence to Structure

Automatic Minuting is a new problem in Speech and Natural Language Processing (NLP). Considering the challenges involved in this task and the limited capacity of the ELITR team, we took a step back and re-oriented our research to achieve the best possible outcome for both ELITR's promised goals and the research community in general.

Ultimately, our activities in Automatic Minuting span several ELITR work packages (WP1 Data, WP5 Automatic Minuting, WP6 Integration, and WP7 Dissemination). Our work on minuting integration into a system prototype was recently described in "D6.5: Demonstrator of Automatic Minuting". The current deliverable thus focuses on the remaining activities, providing the encompassing view, namely:

- Resource Creation for Automatic Minuting (Section 2),
- Automatic Minuting Experiments (Section 3),
- Community Events (SummDial and AutoMin, see Section 4).

We included the original sub-tasks T5.1, T5.2, T5.3, and T5.4 within our Automatic Minuting experiments component.

In the following sections, we will describe the creation of our dataset, the many baseline experiments we performed on Automatic Minuting, our final proposed method encompassing the T5.1, T5.2, T5.3, and T5.4 sub-tasks, and the NLP community events we organized as part of investigating the challenges and initiating community participation for this critical yet timely problem. We conclude in Section 5.

2 Dataset Creation

One of the main reasons that *Automatic Minuting* still remains a challenging problem is because of the lack of appropriate datasets. Privacy and ethical issues are the main concerns for developing real-life meeting corpora and probably the reason why we do not see significant efforts in curating such datasets.

Throughout the duration of the ELITR project, we kept working on our own collection of meeting transcripts and their minutes. Intermediate progress on this has been reported in WP1 deliverables.

Two releases of the data outside of the ELITR consortium happened in 2021 and 2022:

AutoMin 2021 Dataset was a release restricted to the participants of the AutoMin shared task. Full details on the release, including the processing steps through which we prepared the data are presented in the AutoMin Overview paper, see Appendix F.

Attachments to the AutoMin Overview paper contain sample minutes from our dataset.

ELITR Minuting Corpus is the final public release of the dataset. A paper describing this public version is currently under review for the LREC conference (anticipated notification on April 5, 2022). The draft of this paper is attached to this deliverable as Appendix A, but several important updates have happened since the submission, including the introduction of Marie Hledíková to our team.

In the rest of this section, we describe the further steps we carried out with the dataset after AutoMin and then the final status of the dataset, the ELITR Minuting Corpus.

2.1 Data Improvements beyond AutoMin Release

The main differences between AutoMin 2021 Dataset and the final ELITR Minuting Corpus release are the addition of manual alignments between meeting transcripts and minutes, and an additional round of censoring. Both are described below.

2.1.1 Alignments between Transcripts and Minutes

The ELITR Minuting Corpus has been created with machine or deep learning in mind. When creating automatic systems for minuting, reliable evaluation of candidate outputs is essential. In the AutoMin shared task, we used manual judgements of adequacy, fluency and grammatical correctness on three Likert scales, see AutoMin Overview Paper (Appendix F).

We see such a summarizing evaluation as a simplification which we had to make in order to evaluate participants' submissions in the limited timeframe. Ideally, we would prefer a more principled semi-automatic (and later perhaps fully automatic) evaluation strategy. Specifically, we planned to ask annotators to manually align items in the minutes to parts of the transcript, ask them to indicate the adequacy, fluency and grammaticality not only at the document level but also at the level of the individual alignments ("hunks"), and also automatically calculate "coverage", i.e. the portion of DAs that are aligned to an item in the minute.

We implemented an annotation tool, ALIGNMEET,[1] for this purpose and ran a small study with it. The full description of ALIGNMEET and the study results were submitted to LREC 2022 and we are still waiting for the review. The full version of the submitted paper is available in Appendix B.

As a basis for future research on minuting and the suggested type of minuting evaluation, we equipped a subset of ELITR Minuting Corpus minutes with manual alignments created using ALIGNMEET. See Figure 1 for an example of an alignment.

An alignment maps each Dialogue Act (DA) of the transcript to either one line of the minutes file in which it is summarized, a "problem" label, both or neither. The alignments are done in

[1] https://github.com/ELITR/alignmeet

	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.	

(a) Transcript excerpt

	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].

(b) Relevant section of minutes

Figure 1: Example of an alignment viewed in ALIGNMEET. DAs 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.

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).

Most of the provided minutes cover the transcripts completely and mention all important points, however, almost all transcripts contain sections that do not belong in the minutes for various reasons, or are mentioned in the minutes, but are somehow problematic or interesting. For these, we have defined the following problem types, which our annotators were assigning to sections in transcripts as needed:

1. **Organizational:** Organizational talk not directly related to the subject of the meeting (e.g. discussing technical issues with the video call).
2. **Speech incomprehensible:** It is not clear what the speaker is saying.
3. **Other issue:** There is another reason the DA should not be summarized in the minutes.
4. **Small talk:** Small talk or conversation unrelated to the subject of the meeting (e.g. discussing the weather).
5. **Censored:** A section of the transcript removed during the ethical check.

It is possible for a single DA to be aligned both to a minutes line and a problem. Some minutes also do not cover the full range of notable topics, therefore some DAs remain unaligned completely.

2.1.2 Final Ethical Check and Censoring

For the AutoMin 2021 Dataset, the data was de-identified of personally identifiable information (PII), as described in Appendix F. For the purposes of the public release, however, we additionally performed an ethical check and censoring. This was necessary because some information was considered sensitive and could not be released publicly, even in its de-identified form. This was only done on transcripts because the sensitive information appeared exclusively in small-talk parts of the transcript and was not included in the minutes.

Our goal was to censor minimally, so as to keep the data as authentic as possible. Most importantly, we did not want to remove all small talk, as it is an important part of the data. All removed sections were replaced by a `<censored/>` tag to clearly mark the edit.

The ethical check was performed completely manually. We instructed our annotators to remove any parts of the transcript which were too personal, potentially harmful or otherwise sensitive. What counts as too sensitive in this context is to a large degree subjective and it is impossible to define it clearly. As such, there are no hard-and-fast rules we used, and a lot was left up to the annotators' discretion.

To summarize the process briefly, the sections that we decided to remove mostly fall into one of the following categories:

- Remarks that were deemed too personal, e.g. discussions of participants' family arrangements, children or personal plans or medical information. The medical conditions are actually explicitly defined as private data, but our texts were already de-identified, so keeping them in would not violate GDPR. We nevertheless preferred to remove them.
- Remarks that were inappropriate or could harm someone's good name, e.g. overly negative opinions on individual people, organizations or projects, as well as disclosure of bad practices or bad habits by individuals. Again, the previous de-identification prevented from attaching such remarks to concrete people, but the identity of the teams could be in principle guessed from the discussed content.

Altogether we removed 485 lines which add up to 6472 words. While the ethical check was being performed, we also instructed the annotators to correct any leftover typos, badly de-identified PII and other potential problems at the same time.

2.2 ELITR Minuting Corpus Final Status

This section describes the final status of ELITR Minuting Corpus in its publicly released form.

2.2.1 Final Data Layout and Annotation

The final release contains the directories `elitr-minuting-data-cs` and `elitr-minuting-data-en`. These contain Czech and English meetings, respectively. They are further divided into train, dev, test and test2 sets.[2] We added test2 to the collection while the AutoMin shared task was running, as an additional set of fully independent test instances.

Each meeting has a directory containing the following file types:

- **transcript_MANX_annotYY.txt**: the transcript (X – number of consecutive manual revisions if 2 or more, YY – ID of annotator who did them, one file)
- **minutes_ORIG.txt**: the original agenda, typically somewhat expanded into minutes, written by meeting organizer (zero or one file)
- **minutes_GENER_annotYY.txt**: the minutes files written by our annotators (YY – ID of the annotator who wrote it, one or more files)
- **alignment+<transcript_filename>+<minutes_filename>**: the alignment between the transcript and minutes (zero or more files, at most one per each minutes file)

Each line of the **transcript file** contains one Dialogue Act (DA) and has one of these formats:

- (SPEAKER) DA
- DA

The second option means that the DA was spoken by the same speaker as the immediately preceding DA.

Speaker IDs are in the format (PERSON*number*). Other de-identified instances of PII are replaced by identifier strings in the format [ENTITY*number*] or [ENTITY]. Entity is one of the following strings:

- PERSON
- ORGANIZATION
- PROJECT
- LOCATION
- ANNOTATOR
- URL
- NUMBER
- PASSWORD
- PHONE
- PATH
- EMAIL

[2]The division of the data into these sections was kept from AutoMin, see Appendix F; AutoMin just uses the names Test-I and Test-II instead of `test` and `test2`.

Lang	Set	Number of Meetings	Number of Minutes Total	Max per Meeting	Avg±Std.Dev per Meeting	with Alignment
cs	dev	10	32	5	3.2±0.8	20
cs	test	10	30	5	3.0±0.9	23
cs	test2	6	6	1	1.0±0.0	6
cs	train	33	79	3	2.4±0.6	6
en	dev	10	28	8	2.8±2.1	18
en	test	18	55	11	3.1±2.1	49
en	test2	8	10	2	1.2±0.5	8
en	train	84	163	8	1.9±0.9	36

Table 1: Overall statistics of the public release of ELITR Minuting Corpus.

- OTHER

Transcripts further contain the following special tags:

- *<another_language>...</another_language>* or *<another_language/>*: speech in a different language than the rest of the transcript
- *<typing/>*: sounds of typing
- *<parallel_talk>...</parallel_talk>* or *<parallel_talk/>*: speakers talking over each other
- *<cough/>*: coughing
- *<other_yawn/>*: yawning
- *<censored/>*: a section of the transcript has been censored
- *<laugh/>*: laughter
- *<unintelligible/>*: speech is not comprehensible
- *<other_sigh/>*: sighing
- *<talking_to_self/>*: speaker talking to themselves
- *<other_noise/>*: another further unspecified noise

The **alignment files** are in the form of space separated data in three columns: transcript DA line number, minutes line number to which it is aligned or "None" if unaligned and the ID of the problem with this DA or "None". DAs with no alignment or a problem are not present and indices start at 1.

2.2.2 Resulting Corpus Statistics

Table 1 provides the overall statistics of ELITR Minuting Corpus.

We see that the English portion contains 84 meetings in the training part, with up to 8 independently created minutes for one meeting. The average number of minutes per meeting is close to 2. In total, the training set was equipped with 163 minutes. For the test set, we selected meetings which have even more manual minutes: up to 11 and 3 on average.

The last column in Table 1 indicates how many minutes we have with the minute-to-transcript manual alignment. Again, we promoted the annotation of the English test set with 49 aligned minutes in total.

Table 2 reports on the size of the transcripts in the dataset. Overall, there are about 500k Czech words and almost 850k English words in the transcripts (across the train/dev/test divisions).

		Total		Per Meeting:			
Lang	Set	# Meetings	# Words	# Words	# Lines	# Speakers	# People
cs	dev	10	90.1k	9.0k±2.3k	1273±352	7.3±5.3	26.7±11.6
cs	test	10	80.7k	8.1k±3.3k	1097±481	7.3±5.3	23.4±10.2
cs	test2	6	52.9k	8.8k±2.2k	1297±642	7.8±5.7	31.0±19.1
cs	train	33	279.8k	8.5k±3.5k	1201±491	8.3±5.0	24.6±11.8
en	dev	10	64.3k	6.4k±2.4k	763±406	5.1±3.1	12.1±6.0
en	test	18	118.1k	6.6k±2.5k	675±333	6.1±2.5	11.5±5.1
en	test2	8	56.3k	7.0k±2.8k	756±285	5.4±2.6	14.1±6.6
en	train	84	609.3k	7.3k±4.3k	732±425	6.1±2.5	10.8±5.2

Table 2: Transcript size statistics of ELITR Minuting Corpus. We report averages ± standard deviations.

		Total	Per Average Minute:		
Lang	Set	# Minutes	# Words	# Lines	# People
cs	dev	32	264±120	33±9	7.9±5.5
cs	test	30	231±78	34±7	7.7±6.0
cs	test2	6	399±224	55±26	7.8±6.0
cs	train	79	222±125	34±12	7.7±5.0
en	dev	28	228±150	30±12	5.1±2.3
en	test	55	278±84	36±9	5.6±1.8
en	test2	10	468±287	60±34	7.5±4.5
en	train	163	422±458	46±35	5.8±3.0

Table 3: Minuting size statistics of ELITR Minuting Corpus.

The Czech transcripts are, on average, a little longer than the English ones: between 8k and 9k words in Czech compared to 6k–7k words in English. The same applies to the average number of lines (1.1k–1.2k Czech lines vs. about 700 English lines).

In terms of the number of speakers (but ignoring their balance; so perhaps the majority of utterances comes from a single speaker in a meeting), Czech meetings are again a little bigger (7–8 speakers) compared to the English ones (5–6 speakers). The number of mentioned people is considerably higher for Czech meetings: 20–30 people are mentioned in the transcript on average. In English, about half as many people were mentioned, 10–14 on average.

Table 3 summarizes the statistics of the minutes in ELITR Minuting Corpus. For instance the 18 English test set meetings have in total 55 minutes. We report averages of averages in the subsequent columns: an average minute (across the multiple minutes created for a given meeting) has 278.50 words and 35.89 lines. About 5.56 persons are mentioned in a minute on average.

In general, the number of persons mentioned in the minutes is closer to the number of speakers in the given meeting than to the number of persons mentioned in the meeting but there are exceptions of all kinds, of course.

2.3 ELITR Minuting Corpus Availability

ELITR Minuting Corpus is publicly available at the Lindat repository:

`http://hdl.handle.net/11234/1-4692`

3 Experiments and Methods

In this section, we describe the various summarization experiments that we performed for the automatic minuting task.

Our initial experiments are described in Section 3.1. As mentioned earlier and summarized in Section 4.2, we organized the AutoMin shared task. Among the ELITR consortium members, UEDIN and CUNI (Team ABC) participated in the shared task and their submissions are summarized here in Sections 3.2 and 3.3.

3.1 Initial Automatic Minuting Experiments

We conducted several experiments on our AutoMin dataset and other publicly available meeting summarization datasets (AMI, ICSI) with the *state-of-the-art* text summarization models for automatically generating meeting minutes. These experiments were described in Singh et al. (2021), published at PACLIC and included here in Appendix C.

Specifically in our initial explorations to set the baselines for future investigations, we use *off-the-shelf* text summarization models without worrying much about domain-specific training. For generating abstractive meeting summaries/minutes, we employ the following pre-trained off-the-shelf text summarization models: 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). For extractive meeting summaries we use: a *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.

For automatic evaluation, we use the usual text summarization evaluation metrics (ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), BLEU (Papineni et al., 2002)). We also carry out a human evaluation of the generated summaries. Human evaluators rated our minutes/meeting summaries against their *Adequacy, Fluency, and Grammatical Correctness* on a Likert scale of 1 to 5 (1 signifying the worst, 5 the best).

We conclude that off-the-shelf text summarization models are not the best candidates for generating minutes which calls for further research on meeting-specific summarization or minuting models. We found that off-the-shelf transformer-based summarization models perform comparatively better than other categories of summarization algorithms; however, they are still far from generating a good multi-party meeting summary/minutes. With this takeaway, we carried out our investigations on transformer-based models and ultimately came up with a pipelined automatic minuting deep neural architecture which we discuss in Section 3.3.

3.2 UEDIN's Approach in AutoMin Shared Task

For the AutoMin shared task, UEDIN developed a minuting system that combines extractive summarization with logistic regression-based filtering and certain rule-based pre- and post-processing steps. The extractive summarizer was a modified version of *lecture summarizer*,[3], which uses BERT-based sentence embeddings to cluster the sentences in the document, then extracts sentences that are as near as possible to the centroid of each cluster. We set the summarizer to over-generate, then used a filter trained on a portion of hand-labelled data to select the best sentences for the minutes. The rule-based pre- and post-processing helps to remove many of the artifacts of spoken language, creating more fluent output, and formatting it in the style of minutes.

An examination of the output shows that the minutes give a sense of the meeting's discussions, but displays several problems arising from the extractive summarizer – for example some of the extracted sentences are not understandable without context, and there is no mechanism for the model to use reported speech or to express discussions as (e.g.) "PERSON1 and

[3] `https://github.com/dmmiller612/lecture-summarizer`

PERSON4 discussed the topic". We attempted to address these problems using an abstractive summarizer, but found that it was difficult to overcome the length restrictions in such summarizers.

Appendix G provides the full submission description by Williams and Haddow (2021).

3.3 CUNI's Minuting Method in AutoMin and Final Proposed Approach

Team ABC, supervised by CUNI, participated in the AutoMin shared task. We proposed a BART-based minuting model, which is trained on the SAMSum dataset Gliwa et al. (2019) and fine-tuned on our AutoMin dataset.

The full paper by Shinde et al. (2021) has been already included as Appendix A in Deliverable D6.5 Demonstrator of Automatic Minuting, so we do not reproduce it here again.

However, we further leveraged our shared task submission and developed a BART-based minuting pipeline that serves as the basis of our Automatic Minuting work package.

Our final approach encompassed the WP5 sub-tasks T5.1 and T5.2 and, to some extent, T5.3 and T5.4, respectively to generate the meeting minutes (we demonstrated the automatically generated minute examples in Appendix B in Deliverable D6.5). Specifically, our final proposed pipeline model for automatic minuting consists of a pre-processing module (redundancy and small talk elimination, linear segmentation, topical segmentation of the meeting transcript), the BART-based minuting/summarization module, and a post-processing module (integration, redundancy elimination, sentence compression, information filtering).

The associated paper is under review but can be found in Appendix D for reference.

4 Community Events

As part of our effort to include the Speech and Natural Language Processing community in discussing the challenges of this problem and participating in proposing innovative solutions, we organized two community events in 2021.

- SummDial:[4] A SIGDial 2021 Special Session on Summarization of Dialogues and Multi-Party Meetings
- AutoMin:[5] The First Shared Task on Automatic Minuting @ Interspeech 2021

4.1 SummDial Special Session

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.

The 4.5 hours long virtual session had one keynote talk, one-panel discussion, three long papers, and three short paper presentations. Our keynote speaker was Klaus Zechner[6] from Educational Testing Service, United States. His pioneering works on summarization of meeting speech and dialogues helped shape the investigations in this topic further (Zechner and Waibel, 2000; Zechner, 2001, 2002). 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 panelists were Ani Nenkova from the University of Pennsylvania and Adobe Research, US, Diyi Yang from Georgia Institute of Technology, US, Chenguang Zhu from Microsoft Cognitive Services, US. Klaus Zechner, who was our keynote speaker, also joined the discussion.

[4] https://elitr.github.io/automatic-minuting/summdial.html
[5] https://elitr.github.io/automatic-minuting/index.html
[6] https://scholar.google.com/citations?user=eVYrz4EAAAAJ&hl=en

Paper presentations in SummDial included various topics on meeting and dialogue summarization. Liu et al. (2021) presented their work on coreference-aware dialogue summarization. Zhuang et al. (2021) discussed their work on weakly-supervised extractive summarization of dialogues with attention. Manuvinakurike et al. (2021) presented a dataset for incremental temporal summarization in a multi-party dialogue. In Karan et al. (2021), authors explored the task of detecting decision-related utterances in multi-party dialogues while mitigating topical bias. In another work, the corresponding authors (presentation-only; the paper is not published yet) presented a novel dataset of abstractive summaries of turn-labeled spoken human-computer conversations in Dutch. Finally, Liu and Chen (2021) proposed a dynamic sliding window strategy to counter the challenge of summarizing long meeting transcripts.

The critical points that came up during the special session were: 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*** *(Ghosal et al., 2021) and* ***DialogSum*** *(Chen et al., 2021b), develop better evaluation schema, and study effects of transfer learning, multitasking from associated tasks.*

Appendix E reproduces the full report as published in ACM SIGIR Forum (Ghosal et al., 2022). We are looking forward to hosting the next iteration of SummDial. Unfortunately, our foreseen venue, SIGDial 2022, has received too many event proposals and did not select ours; we will thus submit the proposal to another related venue.

4.2 AutoMin 2021 Shared Task

The AutoMin shared task at Interspeech 2021 (Ghosal et al., 2021) 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 took active participation in the challenge. Our AutoMin shared task 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 as discussed in Section 2.

The ***main task*** *consisted of automatically generating minutes from multiparty meeting conversations* provided in the form of transcripts. The objective was to generate minutes as bulleted lists, summarizing the main contents of the meeting, as opposed to usual paragraph-like text summaries. Task B definition was *given a pair of a meeting transcript and a manually-created minute; the task was to identify whether the minute belongs to the transcript.* 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 two different ones.*

Some unique features of AutoMin 2021 were:

- the first shared task on generating minutes from real multiparty meetings,
- a meeting dataset on a language (Czech) other than English,
- multiple reference minutes created by different annotators to allow observing the variance of outputs when humans are carrying out the task,
- source-based manual evaluation, to avoid evaluation bias induced by a particular reference minute.[7]

Considering the current non-availability of large-scale domain datasets on multiparty meeting summarization (even AutoMin dataset is small-scale), the the best recipe for automatic minuting that evolved out from the AutoMin shared task looks like this: training a deep neural

[7]We use the common English word “minutes” to refer to a meeting summary in general. In cases where we need to highlight the existence of multiple such summaries for a given meeting, we also use the non-standard singular “a minute” to refer to one of them.

model on available dialogue summarization datasets (SAMSum (Gliwa et al., 2019), DialSum (Chen et al., 2021a), etc.) and further fine-tuning it on the minuting or meeting summarization datasets (AMI (Mccowan et al., 2005), ICSI (Janin et al., 2003), AutoMin), accompanied by some intelligent pre and post-processing steps.

Appendix F provides the full overview paper, currently in the publication process.

5 Conclusion

We are glad that the Automatic Minuting work package of the ELITR project attempted all the activities we envisaged. Automatic Minuting is a novel task for the speech and NLP community, where we delved into a few uncharted territories. Starting from building a complex dataset, getting the ethical and privacy issue clearance, carrying out baseline experiments, coming up with an alignment tool for minuting evaluation, then organizing a community event to take stock of the state-of-the-art and discuss challenges with the community, finally organizing a new shared task and building a novel minuting method upon the best-performing system, it feels like we have come a full circle. It was not at all easy considering the complexity and novelty associated with the problem. However, we are now uniquely placed and equipped with relevant knowledge to carry out further streamlined research on this problem and build the community around it.

Considering the fact that the world is adapting to the new normal of remote workplaces, Automatic Minuting would be a super valuable tool for professionals. There are several challenges that we still need to address before we reach a point where we could imagine a scenario where meeting participants can hover over past calendar invites and they get the minutes of the meeting. The Automatic Minuting work package in ELITR hopefully made the first step and set the momentum towards that goal.

The following papers have resulted from WP5. They are all available in the appendices.

1. *AutoMin: A Novel Dataset for Automatic Minuting from Multi-Party Meetings in English and Czech* (under review)
2. *ALIGNMEET: A Comprehensive Tool for Meeting Annotation, Alignment, and Evaluation* (under review)
3. *An Empirical Analysis of Text Summarization Approaches for Automatic Minuting* (published at PACLIC 2021)
4. *A Pipeline Method for Generating Minutes from Multi-Party Meeting Proceedings* (under review)
5. *Report on the SIGDial 2021 Special Session on Summarization of Dialogues and Multi-Party Meetings (SummDial)* (Ghosal et al., 2022) (published in the December 2021 issue of the SIGIR Forum)
6. *Overview of the First Shared Task on Automatic Minuting (AutoMin) at Interspeech 2021* (Ghosal et al., 2021) (in the publication process)
7. *Team UEDIN @ AutoMin 2021: Creating Minutes by Learning to Filter an Extracted Summary* (Williams and Haddow, 2021) (in the publication process)
8. *Team ABC @ AutoMin 2021: Generating Readable Minutes with a BART-based Automatic Minuting Approach* (Shinde et al., 2021) (reproduced in the Appendix of the deliverable D6.5; in the publication process)

References

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

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 *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313, Aberdeen, Scotland, UK, August 2021b. Association for Computational Linguistics. URL https://aclanthology.org/2021.inlg-1.33.

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

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

Tirthankar Ghosal, Ondřej Bojar, Muskaan Singh, and Anja Nedoluzhko. 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, 2021. doi: 10.21437/AutoMin.2021-1. URL http://dx.doi.org/10.21437/AutoMin.2021-1.

Tirthankar Ghosal, Muskaan Singh, Anja Nedoluzhko, and Ondřej Bojar. Report on the sigdial 2021 special session on summarization of dialogues and multi-party meetings (summdial). *SIGIR Forum*, 55(2), mar 2022. ISSN 0163-5840. doi: 10.1145/3527546.3527561. URL https://doi.org/10.1145/3527546.3527561.

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.

Yihong Gong and Xin Liu. 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, 2001.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The icsi meeting corpus. pages 364–367, 2003.

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, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

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

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

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

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.

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

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.

I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 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*, 2005.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

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

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

Kartik Shinde, Nidhir Bhavsar, Aakash Bhatnagar, and Tirthankar Ghosal. Team abc @ automin 2021: Generating readable minutes with a bart-based automatic minuting approach. In *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, pages 1–10, 2021. doi: 10.21437/AutoMin.2021-2. URL http://dx.doi.org/10.21437/AutoMin.2021-2.

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, 11 2021. Association for Computational Lingustics. URL https://aclanthology.org/2021.paclic-1.83.

Philip Williams and Barry Haddow. Team uedin @ automin 2021. In *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, pages 1–4, 2021. doi: 10.21437/AutoMin.2021-10. URL http://dx.doi.org/10.21437/AutoMin.2021-10.

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, 2001. doi: 10.1145/383952.383989. URL https://doi.org/10.1145/383952.383989.

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

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/.

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

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

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`.

A AutoMin Dataset Paper (under review)

AutoMin: A Novel Dataset for Automatic Minuting from Multi-Party Meetings in English and Czech

Anna Nedoluzhko, Muskaan Singh, Tirthankar Ghosal, Ondřej Bojar
UFAL, MFF, Charles University, Prague, Czech Republic
(nedoluzhko,singh,ghosal,bojar)@ufal.mff.cuni.cz

Abstract

Taking minutes is an essential component of every meeting, although the goals, style, and procedure of this activity ("minuting" for short) can vary. Minuting is a rather unstructured writing activity and is affected by who is taking the minutes and for whom the intended minutes are. With the rise of online meetings, automatic minuting would be an important benefit for the meeting participants as well as for those who might have missed the meeting. However, automatically generating meeting minutes is a challenging problem due to a variety of factors including the quality of automatic speech recorders (ASRs), availability of public meeting data, subjective knowledge of the minuter, etc. In this work, we present the first of its kind dataset on *Automatic Minuting*. We develop a dataset of English and Czech technical project meetings which consists of transcripts generated from ASRs, manually corrected, and minuted by several annotators. Our dataset, *AutoMin*, consists of 113 (English) and 53 (Czech) meetings, covering more than 160 hours of meeting content. Upon acceptance, we will publicly release (`aaa.bbb.ccc`) the dataset as a set of meeting transcripts and minutes, excluding the recordings for privacy reasons. A unique feature of our dataset is that most meetings are equipped with more than one minute, each created independently. Our corpus thus allows studying differences in what people find important while taking the minutes. We also provide baseline experiments for the community to explore this novel problem further. To the best of our knowledge **AutoMin** is probably the first resource on minuting in English and also in a language other than English (Czech).

(A) Meeting Transcript segment:
(PERSON7) Uh, here is the organization of the [PROJECT9] presentations. So do do you have any preference or d- do you have any idea how do we do it?
(PERSON45) I thought sort of you'd ask with doing it-
(PERSON7) Yeah.
(PERSON45) And the, coordinating. So what what's your propose? I mean, what we have proposed in the a in the offline track seems quite a reasonable. [...]
(PERSON7) Uh, uh, so let's start with the um, um, with the uh, uh, the the postponed review. So [PERSON42], uh, please, let let us know what this doodle is. This is that we need to figure out, the date.
(PERSON36) We should give uh, our project officer the new ah, a new date. And I see more people finally voted it, so- [...]
(PERSON7) Whether we want get little time extension, uh, little time extension uh, of the project. So I don't know if [PERSON36] is aware any date until we should make our uh, mind.
(PERSONI) Um, if we um, ask for an extension, I will be <unintelligible> automatically.
(PERSON7) Okay.

(B) Meeting minutes segment:
- remote presentations organization of the [PROJECT9]
  - Discussion about the results: agreement on the pre-recorded presentation for the [PROJECT1] system paper
  - One slot to present overall results
- The postponed review:
  - doodle with voting for a new date,
  - possible to decide already now
- A time extension of the project
  - 2 or 3 months probably
  - Voting to mid the next week: to fill the table how many months and the reason for that

Figure 1: An example of a meeting transcript and meeting minutes segments from **AutoMin**. As the data has been anonymized, "PERSON*number*" and "PROJECT*number*" denote persons' and projects' placeholders respectively.

1. Introduction

A significant portion of the working population has their mainstream interaction and meetings virtual these days. Amongst many other things, the COVID-19 pandemic has led people to discover innovative ways to continue their work and adapt to the "new normal". Hence virtual meetings are now an integral part of life for the working population. As one has to attend more and more meetings, it requires a considerable effort to note down and retrieve the desired information from the meeting as and when required. Frequent meetings and ensuing context switching hence gives rise to undesired information overload on the participants. For this, usually there is a designated participant or a scribe who jots down the *minutes of the meeting* (see Figure 1) which can consist of important issues, actions points, decisions, or proposed activities discussed during the course of the meeting. Manually writing minutes takes time and distracts attention from the discussion. Hence we believe that an automatic minuting solution will be an useful application of natural language processing for the professional community. However, the task is complicated. Automatic Minuting (AM) systems would need reliable ASR technologies combined with efficient multi-party dialogue processing. Automatic minuting as a task seems close to meeting summarization. However, the goals of these two tasks are somewhat different. Whereas *meeting summarization* intends to sum up the central concepts of the meeting (can disregard some non-central points) while preserving fluency and coherence in the output summary, *meeting minuting* is motivated more towards topical coverage and churning out the action points (Nedoluzhko and Bojar, 2019; Zhu et al., 2020). Thus, the resulting minutes can be a structured bulleted list of critical meeting information where fluency or coherence may be less critical. There is a dearth of such automatic minuting datasets in the community and our current work attempts to fill that gap. Our dataset is also unique in the fact that it includes meetings in Czech and not just English as all similar datasets we are aware of.

The two existing benchmark meeting datasets in English, AMI (Mccowan et al., 2005) and ICSI (Janin et al., 2003) are aimed at meeting summarization. They contain meeting tran-

scripts, extractive summaries (selected relevant transcript lines), and abstractive summaries in the form of coherent paragraphs.

AutoMin is comparable in size to AMI and ICSI, but we differ in three significant aspects: (i) we focus on minuting, so our summaries are organized as bulleted lists, typical for actual meeting minutes; (ii) our dataset includes meetings in two languages, English and Czech, and (iii) we provide multiple minutes for the same meeting, consisting of minutes taken by actual meeting participants and also by specially-trained annotators.

2. Related Work

Given the lack of proper minuting datasets, we survey few existing datasets on meeting and dialogue summarization, which seem closely related. The past decade featured many dialogue summarization datasets (Mccowan et al., 2005; Janin et al., 2003; Zhu et al., 2021; Gliwa et al., 2019; Liu et al., 2019a; Rameshkumar and Bailey, 2020; Krishna et al., 2020; Budzianowski et al., 2020; Clifton et al., 2020). However, resources for meeting summarization are relatively few, probably due to higher annotation costs and privacy issues (Zhu et al., 2021).

Among the meeting datasets, the AMI and ICSI are the most commonly used ones for meeting summarization experiments. The AMI Meeting corpus (Mccowan et al., 2005) contains 100 hours of meeting discussions, two thirds of which are, however, meetings acted artificially according to a scenario. The open-source corpus contains audio/video recordings, manually corrected transcripts, and a wide range of annotations such as dialogue acts, topic segmentation, named entities, extractive and abstractive summaries. The ICSI corpus (Janin et al., 2003) contains 70 hours of regular computer science working teams meetings in English. The speech files range in length from 17 to 103 minutes and involve from 3 to 10 participants. Interestingly, the corpus contains a significant portion of non-native English speakers, varying in fluency from nearly-native to challenging-to-transcribe. Other meeting collections are substantially smaller (e.g., NIST Meeting Room (Michel et al., 2006) or ISL (Burger et al., 2002)), unprocessed (e.g. various official meetings or recorded debates), or do not represent well the "project meetings" domain (e.g. proceedings of parliaments or city coucils).

Some recently released conversational datasets are 463.6K transcripts with short abstractive summaries of Public Radio (NPR) and CNN television interviews from multiple domains. DiDi (Liu et al., 2019a) is a large (328.9K) dialogue dataset of customer service inquiries, but it is not published under an open license. The SAMSum (Gliwa et al., 2019) is a manually annotated dialogue dataset for abstractive summarization with messenger-like artificially created conversations. The dataset is distributed uniformly with two, three, or more than three participants on the topic of booking and general inquiry. The CRD3 conversational dataset (Rameshkumar and Bailey, 2020) is an example of conversations in the gaming domain with multiple lengthy abstractive summaries varying in levels of detail. It is considerably longer in dialogue length than similar conversational dialogue datasets. The MultiWOZ (Budzianowski et al., 2020) dataset consists of natural multi-domain touristic dialogues and their summaries created by random workers on Amazon Mechanical Turk. There are also some other dialogue datasets, such as Spotify podcast (Clifton et al., 2020) with 105,360 podcast episodes some of which may contain dialogues , the collection of doctor–patients conversations (Krishna et al., 2020) and some others.

1 compares our dataset with relevant others, distinguishing meeting collections (top) and other dialogue corpora (bottom of the table). Among the meeting collections, only *AutoMin* has minutes in the form of structured bullet points. The AMI and ICSI corpora have coherent textual abstractive summaries, mostly one-paragraph abstracts and a list of some action points (decisions, problems, progress, etc.).

3. Dataset Description

This section describes our dataset, which consists of de-identified project meetings transcripts in English and Czech and their corresponding minutes. The English part includes project meetings from the computer science domain, with prevailing non-native speakers of English. The discussions in the Czech part are from computer science and public administration domains; all Czech meeting participants are native speakers of Czech. The duration of the meetings varies from 10 minutes to more than 2 hours, but most meetings are about one hour long. Meetings shorter than half an hour are rather exceptions, whereas meeting longer than two hours are topic-oriented mini-workshops, also rather occasional.

In AutoMin, 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). Original minutes are missing for some meeting sessions, but each meeting must contain at least one generated minute. To conform to GDPR and consents of the participants of the meetings, we release only the transcripts and minutes in a de-identified form, not the audio.

3.1. Data Collection

The minuting corpus consists of primarily online meetings, where each participant has their device and is usually wearing a headset with a microphone. Depending on the remote conferencing platform, the meetings are recorded directly by the platform (sometimes as separate channels per speaker, sometimes as one joint channel); rarely, an external sound recording software had to be used to record the audio. There are also few in-person meetings (before the Covid-19 pandemic), all recorded with a single microphone in the middle of the conference room. The recordings have been automatically transcribed using our own in-house ASR systems for English and Czech. The ASR outputs contain no diarization (segmentation to individual speakers). Since most meeting participants of the English meetings are not native speakers of English and due to the highly varying recording conditions and domain-specific terminology, the ASR outputs are often of low quality. Along with the recordings, we also collected original minutes prepared by one of the meeting

Dataset	A	B	C	D	E	F	G	H	I	J
Our data (English)	MM	project meetings	✓	✓	✓	113	9,537	578	242	5.7
Our data (Czech)	MM	project meetings	✓	✓	✓	53	11,784	292	579	3.6
ICSI	MS	project meetings	✓	✗	✓	61	9,795	638	456	6.2
AMI	MS	project meetings	✗	✗	✓	137	6,970	179	335	4
MEDIASum	DS	radio+TV interview	✓	✗	✓	463,596	1,554	14	30	6.5
SAMSUM	DS	booking+inquiry	✗	✗	✓	16,369	84	20	10	2.2
CRD3	DS	games	✓	✓	✓	159	31,803	2,062	2,507	9.6
DiDi	DS	customer service	✓	✗	✗	328,880	/	/	/	2
MultiWoz	DS	tourist enquiry	✓	✗	✓	10,438	180	92	14	2

Table 1: Comparison of dialogue and meeting summarization datasets. Notation: A – category (DS – dialogue summarization, MM – meeting minuting, MS – meeting summarization), B – domain, C – real dialogues (not acted ones), D – multiple summaries for a single transcript, E – open source, F – number of meetings, G – avg. words per transcript, H – avg. words per summary, I – avg. turns per transcript, J - avg. number of speakers.

participants. These minutes are stored together with the specially created minutes (described in 3.3.).

3.2. Data Pre-Processing

The obtained ASR transcripts are given to specially hired annotators for manual correction. Annotators were asked to proceed with the following steps:

- Break the transcript into smaller segments corresponding to natural linguistic points in the speech such as sentence or phrase boundaries, speech vs. silence/pauses, or utterances of one speaker. As a general rule, no segment should be longer than a minute, but most of them are much shorter;
- Diarize the transcripts, i.e., the speakers' codes are given at the beginning of each speaker's utterance in round brackets;
- Correct the transcript according to the agreed guidelines (in short: one sentence per line, focus on recognizing the sequence of words, preserve errors in grammar, add punctuation and letter casing).

Some of the transcripts have been corrected in several steps, in consultation with the meeting participants to ensure higher quality with fewer typos and misunderstandings (as the hired annotators were usually not the meeting participants).

3.3. Creating Minutes

The next step is generating meeting minutes. To get as realistic minutes as possible, we intentionally do not give precise guidelines on creating them. Annotators are supported with examples of minutes and are free to use existing web resources on the topic. However, there are some general recommendations on creating minutes, such as being concise, concrete, avoid overusing person names, and focusing on topical coverage, action points, and decisions.
From the formal point of view, meeting minutes in our dataset mostly have some metadata, such as the name, date, and purpose of the meeting, the list of attendees, and the minuting author's name. The minutes were mainly generated by the same annotator who corrected the transcript for the given meeting. Due to our free-form instructions, the human-generated minutes vary in length and type. Shorter minutes contain just a few action items (less than half a page). Longer minutes may be up to two (occasionally even more) pages.

	English		Czech	
Meeting Minuted	#meetings	#hours	#meetings	#hours
Once	24	22	2	2
Twice	64	65	20	20
More than twice	25	22	31	31
Total meetings	113	109	53	53

Table 2: Basic transcript and minutes statistics for AutoMin.

The added value of our dataset is that we create multiple minutes for the same meeting. Summarizing long multi-party and multi-topic dialogues is a complicated task, and the generated minutes are very subjective. Having numerous independently created minutes for the same transcript allows studying the differences in what people find important while taking the minutes. We plan to use these observations when proposing better manual and automatic evaluation metrics and also use these observations for designing optimal strategies for automatic minutes creation.

3.4. De-Identification

Having corrected transcripts and created minutes, we de-identified the whole dataset. We follow the GDPR norms and remove/mask any personally identifiable information (PII) such as names, addresses, or any other relevant information from the transcripts and the minutes. Additionally, we decided to de-identify any information concerning projects and organizations because this could indirectly reveal the person involved. Except for specific cases, we did not de-identify locations, languages, or names of software, workshops, etc. Moreover, having de-identified persons, projects, and organizations, we consider that the names of these entities cannot lead to personal identification.
Person, Organisation and Project names were replaced with the lexical substitute strings: [PERSON*number*], [ORGANIZATION*number*] and [PROJECT*number*] respectively. We fixed the lexical substitute strings throughout our dataset, so whenever the annotators were able to establish the identity of a given person, the same *string* was used.[1] Before releasing the corpus, we shuf-

[1] In practice, this was complicated by unclear speech, spelling, and lack of knowledge of people's voices.

fled these identifiers within each meeting. In other words, the transcript and all its minutes share the same codes, but different meetings use different randomization. The de-identification was completed using our web-based tool (see `https://github.com/Muskaan-Singh/LREC-2022.git` figure for de-identification), which we specially designed for this purpose).

3.5. Annotator Details

A group of external annotators specially hired for these purposes did a manual correction of the meeting transcripts, minutes creation, and de-identification. All annotators are native speakers of Czech with an excellent command of English. In total, about 20 annotators worked on the project. The annotators have been paid by the hour as per university standards.

3.6. Handling Ethical Issues

All meeting participants gave their consent to make the data publicly available. We provided participants with the list of the meetings they participated in to check the de-identified transcripts and minutes by themselves and ensure that no unwanted personal information are disclosed. In case a participant had any objections, we deleted the corresponding sections from the concerned transcripts and minutes.

While collecting the data, we made two crucial observations. First, people vary significantly in what they consider personal enough to be removed from the public release. Whereas some people do not care about what they discuss, others are cautious about discussing personal issues and relations. Some people object to releasing discussions concerning their ongoing projects. Second, without actually browsing the data planned to be released, the participants cannot effectively give informed consent. For that reason we consider it obligatory to give all participants the possibility ton preview and check the final version of the data before the release.

In the case of our dataset, although we had prior consent of all the participants, we performed one additional check of the de-identified transcript and minute. It revealed the need to completely exclude ten meetings (more than 11 hours) and delete some individual segments from the transcripts of approximately 15 meeting sessions.

4. Dataset Analysis

Table 2 shows the basic statistics of our dataset in terms of 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, all meetings (except for two) in the Czech meetings are minuted at least twice, and more than half of the Czech portion of AutoMin is minuted 3-5 times.

In the following, we discuss the quality of minutes (4.5.) in AutoMin and then analyze the English part of our corpus in comparison with the 137 meetings of AMI (Mccowan et al., 2005) and 61 sessions of ICSI (Janin et al., 2003). We also discuss on the level of abstractiveness (4.1.), topic diversity (4.2.), dialogue act diversity (4.3.) and speaker diversity (4.4.).

4.1. Level of Abstractiveness

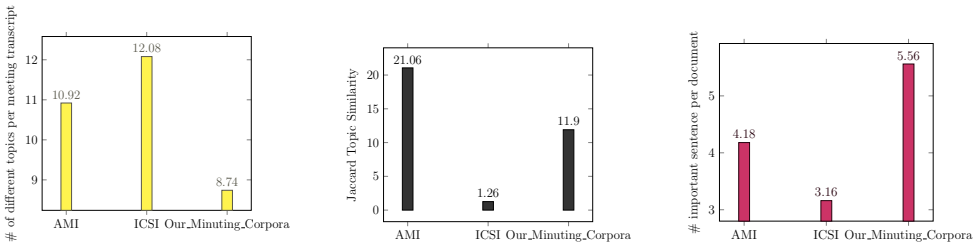
Abstractive summaries involve paraphrasing and are likely to contain words not seen in the transcript. We can thus estimate the *level of abstractiveness* simply by checking what portion of the vocabulary extracted from the minutes is covered by the wording of the transcript. For this analysis, we lemmatize words and exclude stopwords. 3c indicates that close to 30% of word types used in our English minutes do not appear in the transcript, which is twice as many compared to AMI or ICSI.

We also check the distribution of words (excl. stopwords) of the transcript and the minutes. We correlate the number of occurrences of each word in the transcript with the number of occurrences in the minutes. A high Pearson correlation indicates that the minutes are very similar in word distribution to the transcript (persumably being quite verbatim), a low correlation means that the minutes differ. 3d documents that our minutes differ from our transcripts more than what happens in AMI and ICSI.

4.2. Topic Diversity

To demonstrate the multi-topicality of our dataset, we use the Latent Dirichlet Allocation (Blei et al., 2003). Given a set of documents represented as bags of words, LDA automatically identifies "topics" in these documents, representing each topic with a set of keywords relevant to that topic. One of these keywords serves as the topic label. Note that the same word from the documents can serve in multiple topics. We run LDA once for each of the examined datasets, taking both minutes and transcripts in the dataset as the input documents for LDA. We take 100 topics with 20 keywords in each of them and sum the probability for all topic keywords. We further normalize the probability by dividing it by the maximum probability among the 100 topics. If the normalized probability is greater than 0.5, it is treated as relevant topic, other topics are disregarded.

2a reports how many such relevant topics were identified in each document (transcript or minute) on average. To analyze the extent to which the minutes cover the topics discussed in the transcript, we compare the set of topics identified as relevant for a transcript with the set of topics identified as relevant for one of the corresponding minutes using Jaccard similarity (Niwattanakul et al., 2013). 2c plots these similarities averaged over all meetings in the given dataset. Minutes in our dataset appears to cover slightly fewer topics in a meeting than AMI or ICSI. We attribute this to the fact that our annotators may have found some parts of the discussion not worth summarizing. Similarly, based on these topic keywords, we estimate the proportion of relevant sentences in meeting transcripts in Figure 2c. The sentence relevance in transcript is calculated if its occurrence in the minutes/summary is present or not. We score each sentence based on the topic keywords and normalize them by dividing it with the max score. Here we have considered a sentence to be relevant if it has normalized score > 0.7 for topic keywords. Occurrence of relevant sentences indicate how many sentences in our transcript are important and how



(a) Number of topics per documents (b) Coverage (c) Occurrence of Relevant Sentences

Figure 2: Topic diversity of minuting corpora indicated with different topics, their similarity in transcript and minute, and the presence of the summary topic in transcripts

many were just small talks based on topics. The results show the high density of relevant topics in our transcripts.

4.3. Dialogue Act Diversity

We determine the maximum sentence length over the entire transcripts and summaries in Figure 3a. We also determine the position of the maximum length sentence in Figure 3b. It is normalized by the number of sentences in the document so that position is between 0 and 1.

4.4. Speaker Diversity

To observe the biasness in speaker diversity, we calculated the perplexity,entropy in Figure (refer`https://github.com/Muskaan-Singh/LREC-2022.git`) of our minuting dataset. We modeled a different number of speakers and their corresponding count of words. Further we averaged across the dataset. Next we visualize the data distribution by mapping the frequency of parameters across the entire meeting corpora. We plot the number of turns in all meeting corpora and report the presence of multi-party dialogues and summary tokens in Figure (refer `https://github.com/Muskaan-Singh/LREC-2022.git`). We also investigate whether a similar positional bias is present in multi-party dialogues. We record the position of each non-stopword in the transcript that also appears in the summary.To normalize, we partition each transcript into 100 equal-length bins and count the frequency that summary words appear in each bin.

4.5. Data Quality

Estimating the quality of meeting minutes is a very subjective task. People differ in selecting topics which are essential and worthy to be included in the minutes, how much detailed one should be, or how to use different language expressions to describe a meeting action. For some minutes from a series of regular meetings, it could even be challenging to say if they summarize the same session or not. The actual minutes created by meeting participants are sometimes very different from our minutes, both in the formal structure and contents. They may include more information than was discussed in the meeting (for example, because organizers put it there to be addressed, but there was no time for the discussion). On the contrary, they may not include some relevant information. Real project meetings may be open brainstorming sessions where different ideas are discussed, which may or may not have readily identifiable action points or decisions. On the other hand, minutes prepared by our annotators are also subject to human perception. The annotators were involved in manually correcting transcripts, minuting and de-identifying data, but they did not participate in the meetings. Therefore, the minutes maybe different based on the actual annotators, affected by their background, technical knowledge, knowledge of the on-going projects or experience in minuting and annotation, etc.

	R-1	R-2	R-L	R-WE	BLEU
transcript–minute	11.7	7.14	9.09	5.55	23.52
minute–minute	34.28	74.07	24.48	1.33	92.9

Table 3: Automatic Evaluation of Human Annotated Minutes

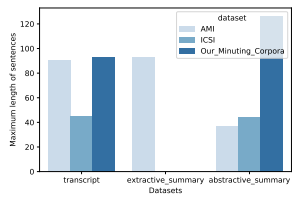
4.6. Manual Evaluation for Human Annotated Minutes

To better understand the quality of minutes in our dataset, we manually evaluated three meetings[2] which had been independently minuted by 8, 8, and 11 people respectively. In five experts, we scored the minutes on the scale of 1 (worst) – 5 (best) according to several generally accepted manual summary estimation criteria: adequacy, topicality, readability, relevance, grammaticality, fluency, coverage, informativeness, and coherence (Kryściński et al., 2019; Zhu et al., 2020; Lee et al., 2020). These criteria are still relatively informal, and their rigorous definition and assessment of inter-annotator agreement are part of our future work.

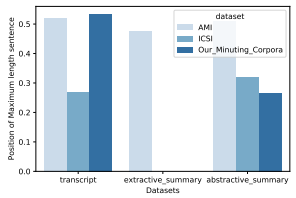
4.7. Automatic Evaluation of Human Annotated Minutes

We analyzed the automatic evaluation(R-1, R-2, R-L, R-WE, BERTScore, BLEU) on the transcript-minute and minute-minute pair. The results empirically shows two minutes of same meeting are lexically very different from each other while the transcript and minute have better lexcial similarity.

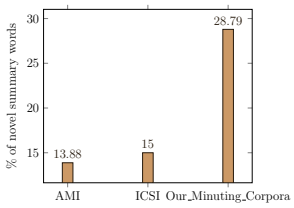
[2]See the supplementary material for all the manually created minutes of the three meetings (labeled A, B, and C).



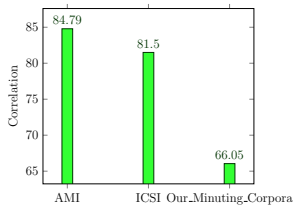
(a) Maximum sentence length over the transcript and minutes



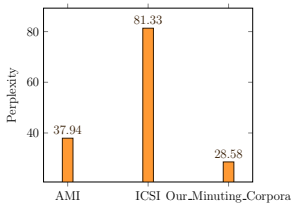
(b) Maximum sentence length position over the transcript and minute



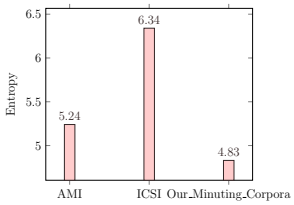
(c) Novel Summary Words



(d) Correlation between Transcript and Minutes Words Distributions



(e) Perplexity



(f) Entropy

Figure 3: (a,b): Maximum sentence length and their position in all the minuting corpora. (c,d): Level of abstractiveness of minuting corpus observed via % of words appearing only in the minutes (c) and lower similarity of word distributions of the transcript and the minutes (d). (e,f): Speakers diversity

	ROUGE_1	ROUGE_2	ROUGE_L	ROUGE_WE	BERTScore	BLEU
BART (Lewis et al., 2019)	24.88	6.36	14.09	6.22	32.08	15.24
BERTSUM (Liu and Lapata, 2019)	20.73	3.67	11.28	4.95	28.94	22.80
BERT2BERT (Rothe et al., 2020)	23.51	5.19	12.03	6.22	19.42	15.54
LED (Beltagy et al., 2020)	9.24	1.28	6.96	0.51	35.80	26.21
Pegasus (Zhang et al., 2020)	22.72	4.55	11.97	4.66	29.12	16.68
Roberta2Roberta (Liu et al., 2019b)	16.67	3.12	9.48	3.13	28.09	**28.90**
T5 (Raffel et al., 2019)	27.01	6.71	14.63	7.59	33.30	16.79
BART-XSum-Samsum[3]	**38.75**	**8.51**	**15.15**	**25.34**	**57.73**	2.69
TF-IDF (Christian et al., 2016)	19.06	3.29	8.45	3.63	25.30	22.43
Unsupervised	23.45	5.04	12.96	2.68	**29.93**	**22.60**
TextRank (Mihalcea and Tarau, 2004)	22.96	5.45	11.94	7.19	17.92	18.32
LexRank (Erkan and Radev, 2004)	22.55	4.14	12.21	5.13	24.94	16.09
Luhn Algorithm (Luhn, 1958)	22.55	4.14	12.21	5.13	24.94	19.05
LSA (Gong and Liu, 2001)	**23.52**	**7.73**	**13.29**	**8.90**	14.61	22.43

Table 4: Quanitative evaluation of summarization methods on AutoMin. The best scores are in bold.

	Adequacy	Fluency	Grammaticality	Coverage
BART (Lewis et al., 2019)	3	3	3.33	3.33
BERTSUM (Liu and Lapata, 2019)	2.66	3.33	3.66	3
BERT2BERT (Rothe et al., 2020)	2.33	2.66	3.66	3
LED (Beltagy et al., 2020)	1.33	1.66	1.66	1.33
Pegasus (Zhang et al., 2020)	3	3	3.66	2.66
Roberta2Roberta (Liu et al., 2019b)	2	2.66	2.66	2.33
T5 (Raffel et al., 2019)	2.66	3	3.66	3
BART-XSum-Samsum[4]	**4**	**4**	**3.5**	**5**
TF-IDF (Christian et al., 2016)	1.66	2	2.66	2
Unsupervised	2.33	2.66	**3.33**	2.33
TextRank (Mihalcea and Tarau, 2004)	2	2.66	2.33	**2.66**
LexRank (Erkan and Radev, 2004)	1.33	2.33	2.66	2.33
Luhn Algorithm (Luhn, 1958)	**2.66**	**2.66**	3	3
LSA (Gong and Liu, 2001)	1.66	2	2	2.66

Table 5: Qualitative evaluation of summarization methods on AutoMin. The best scores are in bold.

5. Evaluation

We evaluate our minuting dataset on three possible use-cases described briefly in `https://github.com/Muskaan-Singh/LREC-2022.git`. Essentially, we consider evaluating our minuting corpora with the existing summarization models. We assess both extractive and abstractive methods of summarization (refer `https://github.com/Muskaan-Singh/LREC-2022.git`). The extractive method, given a transcript, selects a subset of the words or sentences which best represent the discussion of the meeting. While in abstractive, it generates a concise minute that captures the salient notions of the meeting. The generated abstractive minute potentially contains

Table 6: Human evaluation criterion

Criteria	Description
Adequacy	adequately sums up the main contents of the meeting
Fluency	refer to how fluent, coherent, and readable is the output minute text
Grammaticality	grammatical correctness of the minute
Coverage	If the minutes cover the major topics in the meeting transcript

new phrases and sentences that have not appeared in the meeting transcript. Primarily, we experimented with recent models such as 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., 2019b), T5(Raffel et al., 2019), BART_XSum_Samsum[5] and some earlier models such as TextRank(Mihalcea and Tarau, 2004), LexRank(Erkan and Radev, 2004), Luhn(Luhn, 1958), TF-IDF(Christian et al., 2016) and LSA(Gong and Liu, 2001) elaborated in Sec 9.3. We perform quantitative and qualitative analysis on automatically generated minutes.[6] For quantitative analysis, we use the popular automatic summarization metrics like ROUGE (1, 2, L, WE) (Lin, 2004), BERTScore(Zhang et al., 2019) and BLEU (Papineni et al., 2002) which are lexical to evaluate the quality of the summary. The scores are averaged across the datasets. We see that in the abstractive methods, BART-XSum-Samsum performs best in terms of the metrics we took. It is based on transfer learning, where a model is first pre-trained on XSum dataset (Narayan et al., 2018) and further fine-tuned on Samsum corpus(Gliwa et al., 2019). It has been shown to achieve state-of-the-art results on many benchmarks covering summarization; we have presented a sample of the automatically generated output in Sec 1. For qualitative analysis,we ask our annotators to evaluate each automatically generated minute/meeting summary in terms of their *adequacy*, *fluency*, *grammaticality*, and *coverage* using the 5-star Likert rating scale (Likert, 1932) as in 6. We employed three qualified annotators to provide a rating of 1 (worst) to 5 (best) for each criterion to assess the *goodness* of minute given transcript in Table 5. From the table 5 we see BART pre-trained on XSum and fine-tuned on Samsum achieves most readable human evaluation scores.

6. Conclusions and Future Work

In this paper, we present the first version of our AutoMin dataset to generate meeting minutes from meeting transcripts automatically. Our dataset consists of manually corrected transcripts of project meetings in English and Czech and their corresponding minutes jotted by different human scribes. We extensively describe and analyze the annotations (minute creation) both quantitatively, qualitatively and with other meeting datasets as well. Finally, we provide extensive summarization baselines on our dataset. *Automatic Minuting* is a time-critical application of speech and language processing, and we claim that *AutoMin* is a first-of-its-kind dataset to address this use-case. Also, AutoMin is the first meeting dataset to have instances of meetings and minutes in language other than English which we envisage as our attempt to broaden the language diversity for this problem genre. We plan to continue our work and make new versions of the dataset, adding more data (both further collected meetings and newly annotated minutes) and some new annotations, such as topic segmentation and annotating corresponding summaries for them.

[5] https://huggingface.co/lidiya/bart-large-XSum-Samsum

[6] https://anonymous.4open.science/r/minuting-baselines-AB22/README.md

7. Acknowledgment

This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460 (ELITR), and 19-26934X (NEUREM3) of the Czech Science Foundation.

8. Language Resource References

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2020). Multiwoz – a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling.

Burger, S., MacLaren, V., and Yu, H. (2002). The isl meeting corpus: The impact of meeting type on speech style. 01.

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

Clifton, A., Pappu, A., Reddy, S., Yu, Y., Karlgren, J., Carterette, B., and Jones, R. (2020). The spotify podcasts dataset. *arXiv preprint arXiv:2004.04270*.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. (2019). Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Gong, Y. and Liu, X. (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.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The icsi meeting corpus. pages 364–367.

Krishna, K., Khosla, S., Bigham, J. P., and Lipton, Z. C. (2020). Generating soap notes from doctor-patient conversations. *arXiv preprint arXiv:2005.01795*.

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., and Socher, R. (2019). Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Lee, D., Shin, M., Whang, T., Cho, S., Ko, B., Lee, D., Kim, E., and Jo, J. (2020). Reference and document aware semantic evaluation methods for korean language summarization. *arXiv preprint arXiv:2005.03510*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

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

Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Liu, C., Wang, P., Xu, J., Li, Z., and Ye, J. (2019a). Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1957–1965.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Mccowan, I., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (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*.

Michel, M., Ajot, J., and Fiscus, J. (2006). The NIST Meeting Room Corpus 2 Phase 1. In *Machine Learning for Multimodal Interaction, Third International Workshop, MLMI 2006, Bethesda, MD, USA, May 1-4, 2006, Revised Selected Papers*, pages 13–23.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

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

Nedoluzhko, A. and Bojar, O. (2019). Towards automatic minuting of the meetings. In *ITAT*, pages 112–119.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., and Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rameshkumar, R. and Bailey, P. (2020). Storytelling with dialogue: A critical role dungeons and dragons dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134.

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

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Zhu, C., Xu, R., Zeng, M., and Huang, X. (2020). A hierarchical network for abstractive meeting summarization with cross-domain pretraining. *arXiv preprint arXiv:2004.02016*.

Zhu, C., Liu, Y., Mei, J., and Zeng, M. (2021). Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*.

9. Appendix

Kindly refer to the appendix here `https://github.com/Muskaan-Singh/LREC-2022.git`

B ALIGNMEET: A Comprehensive Tool for Meeting Annotation, Alignment, and Evaluation (under review)

ALIGNMEET: A Comprehensive Tool for Meeting Annotation, Alignment, and Evaluation

Peter Polák, Muskaan Singh, Anna Nedoluzhko, Ondřej Bojar
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{surname}@ufal.mff.cuni.cz

Abstract

Meeting summarization is a challenging problem, and even more challenging is to manually create, correct, and evaluate the meeting summary. The severity of the problem grows when the inputs are multi-party dialogues in a meeting setup. To facilitate the research in this area, we present ALIGNMEET, a comprehensive tool for meeting annotation, alignment, and evaluation. The tool aims to provide an efficient and clear interface for fast annotation while mitigating potential error-making. Moreover, we add an evaluation mode that enables a comprehensive quality evaluation of summaries. To the best of our knowledge, there is no such tool available. We release the tool as open source. It is also directly installable from PyPI.

Keywords: meeting summarization, annotation, evaluation

1. Introduction

Meeting summarization is primarily focused on topical coverage rather than on fluency or coherence. It is a challenging and tedious task, even when meeting summaries are created manually. The resulting summaries vary in the goals, style, and they are inevitably very subjective due to the human in the loop. Also, the awareness of the context of the meeting is essential to create adequate and informative summaries.

1.1. Motivation

First, there is a *scarcity of large-scale meeting datasets*: There are a few meeting corpora, such as AMI (McCowan et al., 2005) and ICSI (Janin et al., 2003), which are rather small, on the order of a few dozens of hours each as represented in Table 1. Due to this fact meeting summarization models are usually trained on news (Grusky et al., 2018), stories (Hermann et al., 2015), Wikipedia (Frefel, 2020; Antognini and Faltings, 2020), and other textual corpora, relating poorly to meetings.

Second, when one tries to create such a collection or when a new meeting is to be processed, a *reliable transcript* is needed, which is often impossible for the current automatic speech recognition systems (ASR). It usually requires a large amount of processing to make it suitable for summarization.

Third, meeting transcripts are usually *long text documents* consisting of multi-party dialogues (see Table 1) with multiple topics. Moreover, meeting summaries are also longer compared to text summaries. The manifold structure or length of meeting transcripts and summaries make it difficult to traverse and follow the information for human annotators. Even training is difficult for a neural attention summarization model (Zhu et al., 2020b) with such complexities.

Finally, *evaluation of meeting summarization* requires immediate access to the meeting transcript and sometimes even to the original sound recording to assess the quality of a particular summary point. The length of meeting transcripts and the amount of information quantity contained in a meeting itself put a significant amount of cognitive overload.

1.2. Contribution

We present an efficient, clean, and intuitive comprehensive alignment and evaluation tool which brings the following contributions:

- An annotation platform for data creation and modification with multiple speaker support.
- Alignment between parts of a transcript with corresponding parts of summary.
- A novel evaluation strategy of meeting summaries which we integrate to the tool.

We release the tool as open source.[1] It is also directly installable from PyPI.[2]

2. Related Work

This section studies existing *annotation tools* and *evaluation strategies* for meeting summarization.

2.1. Annotation Tools

Table 2 compares ALIGNMEET with other recent annotation tools for dialogue, conversation and meeting. Most of the tools were designed for data curation. However, only some of them allow modifying existing created datasets (see column *D*). Segmenting the dialogues or turns is possible in some tools (see column *A*) while speaker annotation is possible in almost

[1] https://github.com/ELITR/alignmeet
[2] pip install alignmeet

Category	Dataset	# Meetings	Avg Words (trans)	Avg Words (summ)	Avg Turns (trans)	Avg # of speakers
Meeting	AutoMin (English) (Ghosal et al., 2021)	113	9,537	578	242	5.7
	AutoMin (Czech) (Ghosal et al., 2021)	53	11,784	292	579	3.6
	ICSI (Janin et al., 2003)	61	9,795	638	456	6.2
	AMI (McCowan et al., 2005)	137	6,970	179	335	4.0
Dialogue	MEDIASum (Zhu et al., 2021)	463,596	1,554	14	30	6.5
	SAMSUM (Gliwa et al., 2019)	16,369	84	20	10	2.2
	CRD3 (Rameshkumar and Bailey, 2020)	159	31,803	2,062	2,507	9.6
	DiDi (Liu et al., 2019)	328,880	-	-	-	2.0
	MultiWoz (Budzianowski et al., 2018)	10,438	180	92	14	2.0

Table 1: Dialogue and meeting summarization datasets statistics. The number of words for dialogue, summary, turns, and speakers are averaged across the entire dataset. The meeting dataset statistics have been calculated and dialogue dataset statistics have been derived from Zhu et al. (2021).

all tools (column *B*). ALIGNMEET provides an additional feature of alignment and evaluation of meeting summaries.

We further discuss all these annotation tools in the section mentioned in Table 2.

DialogueView (Heeman et al., 2002) is a tool for annotation of dialogues with utterance boundaries, speech repairs, speech act tags, and discourse segments. It fails to capture inter-annotator reliability. TWIST (Plüss, 2012) is a tool for dialogue annotation consisting of turn segmentation and content feature annotation. The turn segmentation feature allows users to create new turn segments. Further, each segment can be labeled selecting from a pre-defined feature list. This limits the user to pre-defined values. BRAT (Stenetorp et al., 2012) and DOCCANO (Nakayama et al., 2018) are simple web-based annotation tools where you can only edit the dialogue and turns BRAT also provides the user with automated recommendations. INCEpTion (Klie et al., 2018) is a platform for annotation of semantic resources such as entity linking. It provides automated recommendations to the user for annotation. NOMOS (Gruenstein et al., 2005) is an annotation tool designed for corpus development and various other annotation tasks. Its main functionality includes multichannel audio and video playback, compatibility with different corpora, platform independence, information displays for temporal, non-temporal, and related information. This tool is difficult to use by non-technical users and also lacks extensibility. ANVIL (Kipp, 2001) allows multi-modal annotation of dialogues with the granularity in multiple layers of attribute-value pairs. It also provides the feature of statistical processing but lacks the flexibility to add information into the annotation. NITE (Kilgour and Carletta, 2006) is another multi-modal annotation tool aiding in corpora creation. The tool supports the time-alignment of annotation entities such as transcripts or dialogue structure. SPAACy (Weisser, 2003) is a semi-automated tool for annotating dialogue acts. It aids in training corpus creation with grammatical tagging such as topic, mode, polarity. In addition, it produces transcriptions in XML format that require a little post-editing. LIDA (Collins et al., 2019) is one of the most prominent tools for modern task-oriented dialogues with recommendations. However, LIDA does not support more than two speakers in the conversation or additional labeling (e.g., co-reference annotation). MATILDA (Cucurnia et al., 2021) and metaCAT (Liu et al., 2020) address some of the downsides. They add features such as inter-annotator resolution, customizable recommendations, multiple-language support, and user administration. They still lack support for multiple speakers.

All these above-mentioned annotation tools provide annotation for dialogues, but for various textual phenomena. Our tool ALIGNMEET, is specifically designed for meeting data creation or modification, alignment of corresponding meeting transcripts with the corresponding summary, and their evaluation. We also support dialogue and conversational datasets.

2.2. Manual Evaluation

Several researchers, working on summarization have considered qualitative summary evaluation. The qualitative parameters include *accuracy* (Zechner, 2001b; Zechner, 2001a; Goo and Chen, 2018; Nihei et al., 2018; Catherine et al., 2013) usually assesses the lexical similarity between produced text samples and the reference ones utilizing standard metrics such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004). The accuracy is easily computed in some of the applications where reference texts are available. *Grammaticality* measures the capability of a model to produce grammatically correct texts (Liu and Liu, 2009; Mehdad et al., 2013). It is mostly assessed by counting the different types of errors. *Adequacy* (D'Haro et al., 2019; Ma and Sun, 2017; McBurney and McMillan, 2014; Arumae and Liu, 2019; Libovický et al., 2018) rates the amount of meaning expressed in the generated sample given a reference sample. Human participants and categorical scales dominate the assessment process. *Topicality* expresses how well does the generated sample topic matches one of the reference samples (Riedhammer et al., 2008; Arumae and Liu, 2019; Fang et al., 2017). *Naturalness* shows the likelihood of a text being natural or written by a human being rather than automatically generated. Besides accuracy, the rest of the above quality criteria are assessed manually by human experts or survey participants (Zhu and Penn, 2006; Shirafuji et al., 2020). *Relevance* represents how closely are the documents related (Bhatia et al., 2014; Erol et al., 2003; Murray et al., 2010; Zhu et al., 2020a;

Tool	A	B	C	D	E	F	G	I
ALIGNMEET (ours)	✓	✓	✓	✓	✓	✓	✓	Python
MATILDA (Cucurnia et al., 2021)	✓	✓	✓	✓				Python
metaCAT (Liu et al., 2020)	✓	✓	✓					Python
LIDA (Collins et al., 2019)	✓	✓	✓					Python
INCEpTion (Klie et al., 2018)		✓						Java
DOCCANO (Nakayama et al., 2018)		✓						Python
BRAT (Stenetorp et al., 2012)		✓						Python
NITE (Kilgour and Carletta, 2006)		✓	✓	✓			✓	Java
SPAACy (Weisser, 2003)	✓	✓	✓				✓	Perl/Tk
DialogueView (Heeman et al., 2002)	✓	✓						Tcl/Tk
ANVIL (Kipp, 2001)	✓	✓	✓				✓	Java
NOMOS (Gruenstein et al., 2005)	✓	✓	✓	✓			✓	Java
TWIST (Plüss, 2012)	✓							-

Table 2: Annotation Tool Comparison Table. Notation: A – Turn/Dialogue Segmentation, B – Edit Speaker Annotation, C – Data Curation, D – Data Modifications, E – Alignment, F – Evaluation, G – Audio/video playback, H – Programming Language.

Zhang and Fung, 2012; Zhu et al., 2020b; Lee et al., 2020). *Consistency* represents the degree of agreement with the original content (Kryściński et al., 2019; Wang et al., 2020; Lee et al., 2020). *Fluency* represents the quality of expression (Oya, 2014; Wang and Cardie, 2013; Oya et al., 2014; Lee et al., 2020). *Coverage* determines how much of the important content is covered from the source document in the summary (Sonjia and Gina-Anne, 2008; Gillick et al., 2009; Li et al., 2019; Mehdad et al., 2013). *Informativeness* represents the importance of the content captured in the summary (Zhang et al., 2020; Liu and Liu, 2009; Oya et al., 2014; Oya, 2014).

2.3. Automatic Evaluation

The current automatic evaluation of various text summarization tasks (including minuting) is mostly based on ROUGE or similar metrics that utilize n-gram comparisons (from single words to long patterns). Despite being automatic and fast, these metrics are often not able to reflect the quality issues of the text samples (See et al., 2017). Some of the typical problems they miss are grammatical discrepancies, word repetitions, and more. Authors (Novikova et al., 2017; Reiter, 2018) also report that automatic metrics do not correlate well with human evaluations. To overcome these limitations, it is important to simultaneously run human evaluations (following a systematic protocol) of meeting summaries and augment the automatic metric scores with the manual ones.

3. The ALIGNMEET Annotation Tool

ALIGNMEET is a comprehensive annotation and evaluation tool. It supports all stages of the preparation and/or evaluation of a corpus of multi-part meetings, i.e., creation and editing of meeting transcripts, annotating speakers, creating a summary, alignment of meeting segments to a summary, and meeting summary evaluation.

The tool is written in Python using PySide[3] for GUI which makes the tool available on all major platforms (i.e., Windows, Linux, and macOS).

[3] https://www.qt.io/qt-for-python

3.1. Design Choices

We represent a meeting with its transcript and summary in Fig. 1. The transcripts are long documents consisting of multi-party dialogues (refer to the left side of the tool window). The meeting summary is a structured document. We decided to break down the meeting summary into separate *summary points*. A summary point roughly represents a line in a summary document (refer to the right part of the tool window). The meeting usually has more versions of transcripts (e.g., generated by ASR and a manual one) and more versions of summaries (e.g., supplied by meeting participants created during the meeting and others provided by an annotator). We add drop-down lists to select a specific version of the transcript and summary. If the user changes the version of one, the program loads the appropriate version automatically.

We segment the transcript into dialogue acts (DAs). A DA represents the meaning of an utterance at the level of illocutionary force Austin (1975). In the context of our tool, a DA represents a continuous portion of a transcript uttered by one speaker on a single topic. We believe that for better readability, the DA might be further broken down into smaller utterances.

Optionally, the meeting might have an audio or video recording. The meeting recording is helpful during the meeting annotation (i.e., creating/editing the meeting transcript and summary). The tool offers an embedded player. Then, the annotator does not have to switch between the annotation tool and a media player. Also, if the transcripts come with timestamps, the annotator can easily skip by double-clicking to the particular DA.

Many annotation tools we reviewed in Section 2.1 provide automated suggestions. We decided not to include this feature as we believe it would bias the annotators.

ALIGNMEET is designed with two modes: Annotation and Evaluation. We further elaborate them in Sections 3.2 and 3.3.

3.2. Annotation

The annotation task consists of several sub-tasks. We envision the following sub-tasks: (1) transcript annotation, (2) summary annotation, and (3) alignment.

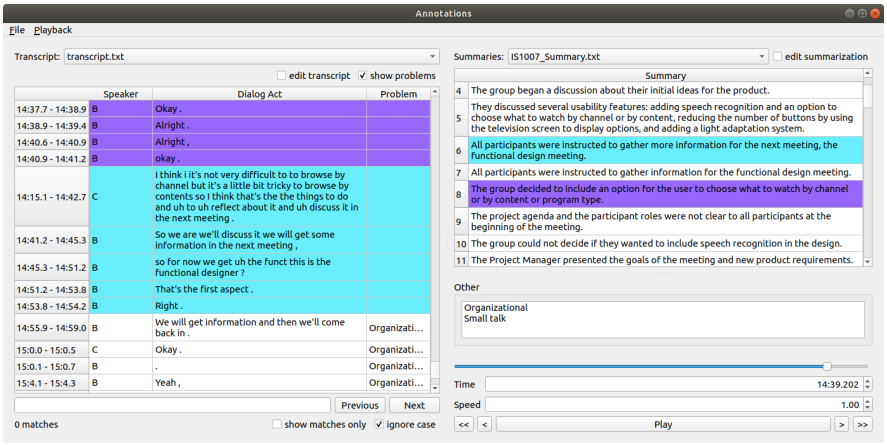


Figure 1: The ALIGNMEET in the annotation mode. The left column contains the meeting transcript broken down to dialogue acts. The right column contains a summary, and the player. The alignment between dialogue acts and the summary point is shown using colors.

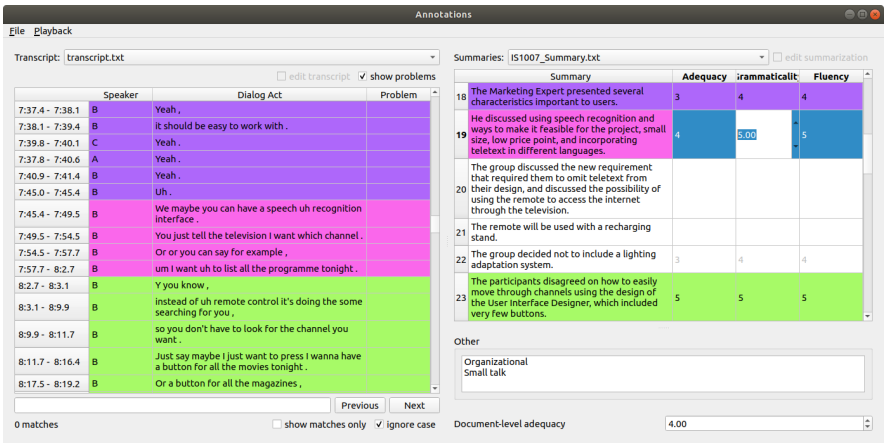


Figure 2: The ALIGNMEET main view in evaluation mode. The left column contains the meeting transcript broken down to dialogue acts. The right column contains a summary, problems, and player. Evaluation, assignment of scores to the particular summary point, is enabled only for the summary points where corresponding DAs are visible in the transcript view.

3.2.1. Transcript Annotation

Transcripts may be either generated by an ASR or manually created. The tool supports both scenarios, i.e., transcribing the recording, post-editing, splitting the transcript into dialogue acts, and speaker annotation.

We introduce a set of keyboard shortcuts that make simple tasks like creating/deleting or even splitting DAs very quickly. Additionally, we offer a search toolbar supporting regular expressions.

3.2.2. Summarization Annotation

Summarization annotation involves the creation or possible modification of an existing meeting summary. The tool provides a convenient platform to add more points to an existing summary by simply clicking "add" or "delete".

Except for summary points, we intentionally do not enforce any precise summary structure and provide users with the flexibility to design their summary. Though, we support indentation as a form of horizontal structur-

ing (with a user-defined indentation symbol).

3.2.3. Alignment

The alignment captures which dialogue acts are associated with a particular summary point. We call a set of DAs belonging to a summary point a *hunk*. DAs which do not correspond to a summary point may be assigned meta-information (i.e., marked as small talk or organizational).

We believe aligning multiple summary points to a DA would further increase the difficulty of the alignment task. It would also cause a "summary point fragmentation", as the annotator might address the same information in separate summary points. When a DA includes more information that should create more summary points, we suggest splitting the DA accordingly.

The matching background color of a hunk and a summary point represents a single alignment (see Fig. 1). To make the interface more clean and readable for the annotator, we color only summary points whose hunks are currently visible in the transcript view.

Aligning DA(s) to a particular summary point or meta-information item is very intuitive:

1. *Select DA(s)* in the transcript view. Selection can be continuous and also discontinuous. Standards GUI gestures are supported (i.e., dragging over items, `[Ctrl]`/`[Shift]` + clicking/dragging).
2. *Select a summary point* by double-clicking on an item in the summary view or the meta-information list.

Resetting alignment is also possible by selecting DA(s) or a summary point and selecting an action in the context menu or keyboard shortcut.

In this way, alignment facilitates the annotation and mitigates potential errors. The annotator has a clear overview of which parts of a meeting are already annotated and makes any revisions straightforward.

3.3. Evaluation Mode

We reviewed several quality criteria for a summary evaluation in Sections 2.2 and 2.3 based on which we formulate a novel manual evaluation scheme. We integrated the evaluation into the tool (see Fig. 2).

For the evaluation, we utilize *adequacy*, *grammaticality* and *fluency*. We think that evaluating these criteria on a document level is challenging and error-prone. Therefore, we propose the evaluation on two levels: (1) a hunk (based on alignment) and (2) a document level. At the hunk level, the evaluation is based only on the aligned part of the transcript and a corresponding summary point.

At the hunk-level, we evaluate *adequacy, grammaticality* and *fluency* using 5-star scale (Likert, 1932) with 1 being the worst and 5 the best. At the document level, aggregate the hunk-level criteria. Using the alignment, we further compute *coverage*, i.e., the number of aligned DAs divided by the total number of DAs.

Aside from averaging hunk-level adequacy across the document, we also independently ask annotators to report the overall accuracy of the minutes. We call this score 'Doc-level adequacy' in the following.

4. Use Case and Pilot Study

In this section, we present a use case, and we conduct a small-scale pilot study.

4.1. Use Cases

We conducted the First Shared Task on Automatic (Ghosal et al., 2021) on creating minutes from multi-party meetings. As a part of the shared task, we made available a minuting corpus. ALIGNMEET was created during the annotation process. We have started with a modified NITE (Kilgour and Carletta, 2006) tool, but the annotators faced many issues, including the need to make changes to the transcript and minutes. Hence, we decided to create a new tool to meet the annotators' requirements. We used agile development, i.e., we constantly improved ALIGNMEET following the annotators' comments.

Before annotation, each meeting consisted of a recording, ASR-generated transcript, and meeting minutes assembled by the meeting participants (often incomplete). First, we asked the annotators to revise the ASR transcript. Later, we asked the annotators to provide minutes and alignment. We have observed different styles of minuting among the annotators. Therefore, most of the meetings have two versions of minutes provided by different annotators.

We employed 27 annotators with a mainly non-technical background. Finally, we have collected 113 meetings in English with 103 hours of recordings and 53 Czech meetings totaling 53 hours. In total, the annotation of the corpus took 2208 hours of work.

4.2. Pilot Study

To assess ALIGNMEET, we conduct a simple experiment similar to Collins et al. (2019) for both modes of tool: (1) annotation and (2) evaluation. We evaluate all the results across two different meeting corpora, AMI (McCowan et al., 2005) for English and AutoMin for Czech. We considered one meeting per language from each corpus (the selected English meeting has 205 DAs and the selected Czech meeting has 153 DAs; both are approximately 16 minutes long). The task was to create an abstractive summary, align the transcript with corresponding parts of the reference summary, and finally evaluate the reference summary relying on the constructed alignment. Each of the three annotators had a different experience level and report their timings in Table 3. The summarization stage took on average 40.7 minutes and 33.0 minutes for English and Czech, respectively. The alignment took on average 16.0 and 19.7 minutes and evaluation on average of 11.7 and 17.7 minutes for English and Czech data, respectively.

	English			Czech		
Annotator	E1	E2	E3	C1	C2	C3
Experienced	✗	✓	✓	✗	✓	✓
Summarization	37	45	40	23	45	31
Alignment	5	23	20	18	30	11
Evaluation	10	15	10	25	15	13
Total time	52	83	70	66	90	55

Table 3: Pilot study: experience and time in minutes an annotator spent on each task.

In other words, this particular meeting needed about 2–3 times its original time to summarize, about its duration to align and finally somewhat less than its duration to evaluate. Based on this minimal study, a factor of 4 or more has to be expected when processing meetings by annotators who have not taken part in them. The evaluation results are in Table 4. Additionally, we report the inter-annotator agreement (IAA). Our definiton of IAA is rather strict, we count the number of DAs that were aligned to the same summary point by all annotators divided by the total number of DAs.

If we consider the recorded pace of our annotators, the AMI meeting corpus consisting of 137 meetings and 45,895 DAs in total (see Table 1). It would take 9,105 minutes to summarize, 3,582 minutes to align, and 2,613 minutes to evaluate using our tool, or 255 hours in total. We infer from Table 3 that the time spent on the task does not necessarily depend on the annotator experience but rather the personal preferences and thoroughness of the annotator. Despite the limited size of the experiment, we believe that the results suggest the tool is intuitive and facilitates fast annotation.

5. Conclusion

We presented ALIGNMEET, an open-source and intuitive comprehensive tool for meeting annotation. The main feature is to perform alignment between parts of a transcript with the corresponding part of the summary. We also integrate the proposed evaluation strategy of meeting summaries in the tool.

In the future, we will add the support for automatic transcript generation with timestamps, user-defined problems in the list of explicit problem labels, and a quick onboarding tutorial integrated into the user interface. Finally, we hope ALIGNMEET will generally improve as annotators will provide their feedback.

Acknowledgements

This work has received funding from START/SCI/089 (Babel Octopus: Robust Multi-Source Speech Translation) of the START Programme of Charles University, the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460 (ELITR), and 19-26934X (NEUREM3) of the Czech Science Foundation.

	English			Czech		
Annotator	E1	E2	E3	C1	C2	C3
Experienced	✗	✓	✓	✗	✓	✓
#Summary points	15	11	19	23	14	21
#Alignments	378	378	203	282	282	282
IAA		0.21*			0.63	
Avg. adequacy	3.71	3.71	3.17	3.67	4.93	4.67
Avg. grammaticality	3.86	4.21	4.08	5.00	4.13	4.67
Avg. fluency	4.71	4.07	4.92	5.00	4.53	4.53
Doc.-level adequacy	3.00	4.00	4.00	4.00	4.00	4.00
Coverage	1.00	0.94	0.54	0.64	0.54	0.30

Table 4: Pilot study: annotator experience, number of produced summary points and alignments, and evaluation score.
∗ If we remove the second annotator, we obtain agreement 0.59.

6. Bibliographical References

Antognini, D. and Faltings, B. (2020). Gamewikisum: a novel large multi-document summarization dataset. *arXiv preprint arXiv:2002.06851*.

Arumae, K. and Liu, F. (2019). Guiding extractive summarization with question-answering rewards. *arXiv preprint arXiv:1904.02321*.

Austin, J. L. (1975). *How to do things with words*, volume 88. Oxford university press.

Bhatia, S., Biyani, P., and Mitra, P. (2014). Summarizing online forum discussions–can dialog acts of individual messages help? In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2127–2131.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018). Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Catherine, L., Carletta, J., and Renals, S. (2013). Detecting summarization hot spots in meetings using group level involvement and turn-taking features. In *INTERSPEECH 2013 14th Annual Conference of the International Speech Communication Association*, pages 2723–2727, Lyon, France. ICSA.

Collins, E., Rozanov, N., and Zhang, B. (2019). Lida: lightweight interactive dialogue annotator. *arXiv preprint arXiv:1911.01599*.

Cucurnia, D., Rozanov, N., Sucameli, I., Ciuffoletti, A., and Simi, M. (2021). Matilda-multi-annotator multi-language interactivelight-weight dialogue annotator. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 32–39.

D'Haro, L. F., Banchs, R. E., Hori, C., and Li, H. (2019). Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics. *Computer Speech & Language*, 55:200–215.

Erol, B., shyang Lee, D., and Hull, J. (2003). Multimodal summarization of meeting recordings. In *In*

Proceedings of the IEEE International Conference on Multimedia & Expo, Baltimore, MD, July.

Fang, C., Mu, D., Deng, Z., and Wu, Z. (2017). Word-sentence co-ranking for automatic extractive text summarization. *Expert Systems with Applications*, 72:189–195.

Frefel, D. (2020). Summarization corpora of wikipedia articles. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6651–6655.

Ghosal, T., Singh, M., Nedoluzhko, A., and Bojar, O. (2021). Overview of the first shared task on automatic minuting (automin) at interspeech 2021. In *In print*.

Gillick, D., Riedhammer, K., Favre, B., and Hakkani-Tur, D. (2009). A global optimization framework for meeting summarization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4769–4772. IEEE.

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. (2019). Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Goo, C. and Chen, Y. (2018). Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742, Athens, Greece, Dec. IEEE Xplore.

Gruenstein, A., Niekrasz, J., and Purver, M. (2005). Meeting structure annotation: Data and tools. In *6th SIGdial Workshop on Discourse and Dialogue*.

Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Heeman, P. A., Yang, F., and Strayer, S. E. (2002). Dialogueview-an annotation tool for dialogue. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 50–59.

Hermann, K. M., Kočiskỳ, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. *arXiv preprint arXiv:1506.03340*.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003). The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Kilgour, J. and Carletta, J. (2006). The nite xml toolkit: Demonstration from five corpora. In *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*.

Kipp, M. (2001). Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., and Socher, R. (2019). Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Lee, D., Shin, M., Whang, T., Cho, S., Ko, B., Lee, D., Kim, E., and Jo, J. (2020). Reference and document aware semantic evaluation methods for korean language summarization. *arXiv preprint arXiv:2005.03510*.

Li, M., Zhang, L., Ji, H., and Radke, R. J. (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.

Libovický, J., Palaskar, S., Gella, S., and Metze, F. (2018). Multimodal abstractive summarization of open-domain videos. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL). NIPS*.

Likert, R. (1932). A technique for the measurement of attitudes. volume 140, pages 5–55. Archives of Psychology.

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

Liu, F. and Liu, Y. (2009). From extractive to abstractive meeting summaries: Can it be done by sentence compression? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 261–264.

Liu, C., Wang, P., Xu, J., Li, Z., and Ye, J. (2019). Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1957–1965.

Liu, X., Xue, W., Su, Q., Nie, W., and Peng, W. (2020). metacat: A metadata-based task-oriented chatbot annotation tool. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–25.

Ma, S. and Sun, X. (2017). A semantic relevance based neural network for text summarization and text simplification. *arXiv preprint arXiv:1710.02318*.

McBurney, P. W. and McMillan, C. (2014). Automatic documentation generation via source code summarization of method context. In *Proceedings of the 22nd International Conference on Program Comprehension*, pages 279–290.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., et al. (2005). The ami meeting corpus. In *Proceedings of the 5th International Confer-*

ence on Methods and Techniques in Behavioral Research, volume 88, page 100. Citeseer.

Mehdad, Y., Carenini, G., Tompa, F., and Ng, R. (2013). Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 136–146.

Murray, G., Carenini, G., and Ng, R. (2010). Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of the 6th International Natural Language Generation Conference*.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Nihei, F., Nakano, Y. I., and Takase, Y. (2018). Fusing verbal and nonverbal information for extractive meeting summarization. In *Proceedings of the Group Interaction Frontiers in Technology*, GIFT'18, New York, NY, USA. Association for Computing Machinery.

Novikova, J., Dušek, O., Cercas Curry, A., and Rieser, V. (2017). Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark, September. Association for Computational Linguistics.

Oya, T., Mehdad, Y., Carenini, G., and Ng, R. (2014). A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 45–53.

Oya, T. (2014). Automatic abstractive summarization of meeting conversations. Master's thesis, University of British Columbia, Vancouver, Canada.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Plüss, B. (2012). Annotation study materials.

Rameshkumar, R. and Bailey, P. (2020). Storytelling with dialogue: A critical role dungeons and dragons dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134.

Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, September.

Riedhammer, K., Favre, B., and Hakkani-Tur, D. (2008). A keyphrase based approach to interactive meeting summarization. In *2008 IEEE Spoken Language Technology Workshop*, pages 153–156. IEEE.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July. Association for Computational Linguistics.

Shirafuji, D., Kameya, H., Rzepka, R., and Araki, K. (2020). Summarizing utterances from japanese assembly minutes using political sentence-bert-based method for qa lab-poliinfo-2 task of ntcir-15. *arXiv preprint arXiv:2010.12077*.

Sonjia, W. and Gina-Anne, L. (2008). Topic summarization for multiparty meetings.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Wang, L. and Cardie, C. (2013). Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405.

Wang, A., Cho, K., and Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Weisser, M. (2003). Spaacy–a semi-automated tool for annotating dialogue acts. *International journal of corpus linguistics*, 8(1):63–74.

Zechner, K. (2001a). Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *IN PROC. ACM SIGIR*, pages 199–207, New Orleans, USA. ACM.

Zechner, K. (2001b). *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA.

Zhang, J. J. and Fung, P. (2012). Automatic parliamentary meeting minute generation using rhetorical structure modeling. *IEEE transactions on audio, speech, and language processing*, 20(9):2492–2504.

Zhang, X., Zhang, R., Zaheer, M., and Ahmed, A. (2020). Unsupervised abstractive dialogue summarization for tete-a-tetes. *arXiv preprint arXiv:2009.06851*.

Zhu, X. and Penn, G. (2006). Summarization of spontaneous conversations. In *Ninth International Conference on Spoken Language Processing*.

Zhu, C., Xu, R., Zeng, M., and Huang, X. (2020a). End-to-end abstractive summarization for meetings. *CoRR*, abs/2004.02016.

Zhu, C., Xu, R., Zeng, M., and Huang, X. (2020b). A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online, November. Association for Computational Linguistics.

Zhu, C., Liu, Y., Mei, J., and Zeng, M. (2021). Mediasum: A large-scale media interview dataset

for dialogue summarization. *arXiv preprint arXiv:2103.06410*.

C Automatic Minuting Baseline Experiments (accepted)

An Empirical Analysis of Text Summarization Approaches for *Automatic Minuting*

Muskaan Singh, Tirthankar Ghosal, and Ondřej Bojar
Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics,
Charles University, Czech Republic
(last-name)@ufal.mff.cuni.cz

Abstract

A significant portion of the working population has their mainstream interaction virtually these days. Meetings are being organized and recorded daily in volumes likely exceeding what can be ever comprehended. With the deluge of meetings, it is important to identify and jot down the essential items discussed in the meeting, usually referred to as the *minutes*. The task of minuting is diverse and depends on the goals, style, procedure, and category of the meeting. *Automatic Minuting* is close to summarization; however, not exactly the same. In this work, we evaluate the current state-of-the-art summarization models for automatically generating meeting minutes. We provide empirical baselines to motivate the community to work on this very timely, relevant yet challenging problem. We conclude that off-the-shelf text summarization models are not the best candidates for generating minutes which calls for further research on meeting-specific summarization or minuting models. We found that Transformer-based models perform comparatively better than other categories of summarization algorithms; however, they are still far from generating a good multi-party meeting summary/minutes. We release our experimental code at https://github.com/ELITR/Minuting_Baseline_Experiments.

1 Introduction

With the world adapting to the *new normal* in the pandemic and virtual interactions going mainstream, meeting are held and recorded daily in volumes likely exceeding what can be ever perceived. With the deluge of meetings, it is essential to record the key points of the discussions during the meeting to take stock and identify action items for the future, usually referred to as the *minutes* (see Figure 1). However, not all meetings have the same goal. Some are general meetings, some are topic-focused, while some are informal. According to a certain study,[1] there are six major categories of working meetings: status update, information sharing, decision making, problem-solving, innovation, and team-building meetings. Each meeting has a different set of agenda items and objectives expected to appear in its minutes.

To deal with the flooded information from multiple meetings, which sometimes results in severe cognitive overload, it is essential to provide minutes of the meeting to the participants. Without a meaningful note-taking scribe, it is challenging to correctly remember the contents of a meeting, even for the participants. Not only to the participants, but minutes also help the non-participants (e.g., absentees) to quickly understand what was being discussed, decisions-made, or action items proposed. However, the task is not straightforward, it is sometimes difficult even for meeting participant to take notes on the fly. With the great progress of NLP in almost all areas of speech and text processing, an automatic minuting assistant would be a valuable addition to the meeting workflow. However, the task of *Automatic Minuting* is challenging due to a variety of other reasons, which include: comprehending the goal of the meeting, identifying the crux of the discussion while

[1] http://meetingsift.com/the-six-types-of-meetings/

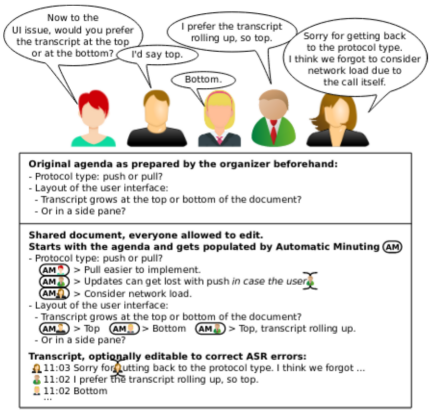


Figure 1: A Proposal for Automatic Minuting

eliminating small talk and redundancies, identifying topics and drifts, etc.

1.1 Relation of Text Summarization and Automatic Minuting

Automatic minuting is close to text summarization, but the goals of these two tasks are somewhat different. *Text summarization* intends to sum up the central concepts of the text, preserving fluency and coherence of the output summary, while *minuting* is a kind of a slot-filling task; it is motivated more towards topical coverage and churning out action points. Thus, the resulting minutes are expected to contain a bulleted list where fluency or coherence may be less critical. An example highlighting the subtle difference between a meeting summary and a minute is in Figure 2. Comprehending multi-party dialogues is itself challenging, so is automatically producing a text summary. Hence, the problem grows more intense when these two problems come together.

1.2 Contribution

Our work is an attempt towards this complex task of automatic minuting while exploring the performance of existing state-of-art text summarization techniques. Our contributions in this work are:

- We implement 13 different summarization methods (extractive, abstractive) an test them on

(A) Meeting Transcript segment:

ME: . . . I've done some research. We have we have been doing research in a usability lab where we observed users operating remote controls. we let them fill out a questionnaire. Remotes are being considered ugly.and an additional eighty percent indicated that they would spend more money on a fancy-looking remote control. Fifty percent of the people indicated they only loo used about ten percent of the buttons on a remote control . . .

ID: I've got a presentation about the working design. first about how it works. It's really simple. Everybody knows how a remote works. The user presses a button. The remote determines what button it is, uses the infrared to send a signal to the TV ... they only use about ten percent of the buttons, we should make very few buttons . . .

UI: But Got many functions in one remote control, you can see, this is quite simple remote control. few buttons but This re remote control got a lot of buttons. people don't like it, so what I was thinking about was keep the general functions like they are.

PM: Extra button info. that should be possible as. let's see what did we say. More. Should be fancy to, fancy design,easy to learn. Few buttons, we talked about that. Docking station, LCD. general functions And default materials. . . And we have to be very attent in putting the corporate image in our product. So it has to be visible in our design, in the way our device works. . .

PM: . . . I will put the minutes in the project document folder. . . And we have a lunch-break now.

(B) Meeting minutes segment:

- Discussion about the research performed on usability of remote controls and talked about the docking station, LCD, and general functions.
- Eighty percent indicated that they would spend more money on a fancy-looking remote control while ten percent use very few buttons.
- Working of a remote was explained and decided to make few buttons.
- It should have a fancy design which is easy to learn with few buttons on the right places.
- A lot of functions of the remote control should be put in a simple manner.
- Pricing needs to be decided and should be a great deal to people. Survey indicated that an LCD screen in the remote control would be preferred.

(C) Meeting summarization segment:

The Project Manager stated the agenda and the marketing expert discussed what functions are most relevant on a remote,what the target demographic is, and what his vision for the appearance of the remote is.The Marketing Expert also brought up the idea to include a docking station to prevent the remote from getting lost and the idea to include an LCD screen.The User Interface Designer pushed for a user interface with large buttons, a display function, a touchscreen, and the capability of controlling different devices.The team then discussed teletext, the target demographic, the buttons the remote should have, the idea of marketing a remote designed for the elderly, an audio signal which can sound if the remote is lost, LCD screens, and language options · · · whether to include teletext in the design despite the new requirement which indicates that the team is not to work with teletext.The buttons are generally used, but the main feature is ugly and ugly.The remote will only have a few buttons.The remote will feature a small LCD screen.The remote will have a docking station.

Figure 2: A meeting of AMI dataset with (a) transcript, (b) minutes, and (c) summarization. Notations: PM -project manager, ME -marketing expert, ID -industrial designer, UI -user interface designer are roles of the speakers.

three different meeting datasets: AMI (Mccowan et al., 2005) and ICSI (Zechner, 2001) and AutoMin[2].

- We evaluate the output minutes using five automatic evaluation metrics along with expert, crowd-sourced human evaluations on criteria like adequacy, coverage, fluency, and grammaticality.

2 Related Work

Text and speech summarization are widely popular NLP tasks, and there is a lot of literature describing their methods and results. However, in this work, we focus on summarizing multi-party dialogues in a meeting setup, and for this task, the amount of prior work is not so extensive.

The majority of the existing meeting summarization experiments are conducted on the AMI (Carletta, 2007) or ICSI corpus (Janin et al., 2003a). In our work, we do not provide a novel method for the task; instead, we evaluate the performance of text summarization methods to attempt the novel task of *Automatic Minuting*. Most of the prior work in summarization are on newspaper texts (Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016; See et al., 2017; Celikyilmaz et al., 2018; Chen and Bansal, 2018; Zhong et al., 2019; Xu and Durrett, 2019; Liu and Lapata, 2019; Lebanoff et al., 2019; Cho et al., 2019; Wang et al., 2020; Xu et al., 2019; Jia et al., 2020) using the standard CNN-daily mail (Hermann et al., 2015) or Newsroom (Grusky et al., 2018) corpora.

Although comparatively lesser, meeting summarization is explored in the works of Chen et al. (Chen and Metze, 2012), Wang et al. (Wang and Cardie, 2013). Some investigations to generate meeting summaries explore with leveraging entailment graphs and ranking strategy by (Mehdad et al., 2013),decisions, action items and progress by (Wang and Cardie, 2013), template generation by (Oya et al., 2014),multi-sentence compression by (Shang et al., 2018), incorporation of multi-modal information by (Li et al., 2019). Recently, a very promising model was proposed by (Zhu et al., 2020) to generate meeting summarization utilizing the word and turn level hierarchical structure.

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

Symbol	Representation
$\tau = (\tau_1, \tau_2, \ldots, \tau_\nu)$	Transcript (Meeting recordings)
$\mu_i = (s_1, \ldots, s_{M_i})$	Minutes
$\rho_j \in \mathrm{P}$	Speakers
α	Agenda of the meeting
s_k	Minute Item
N_i	Total utterances
δ_j	Individual utterances
η	Neural network parameters
$P(\mu\|\tau;\eta)$	Conditional probability (minute/transcript)

Table 1: Problem Description Notations

3 Problem Description

Each meeting consists of multiple participants where every person participates with some utterance or conversation represented by δ. Formally, $\tau_i = ((\rho_1, \delta_1), (\rho_2, \delta_2), \ldots, (\rho_{N_i}, \delta_{N_i}))$ where $\rho_j \in \mathrm{P}$ are the speakers, N_i is the number of utterances in the transcript τ_i and δ_j are the individual utterances (sequences of words; $1 \leq j \leq N_i$).

The minutes formed by human annotators for meeting τ_i is denoted by μ_i, which is a sequence of segments (think items in bulleted list). Formally, $\mu_i = (s_1, \ldots, s_{M_i})$, where s_k is the given minutes item, i.e. a sequence of words and punctuation.

The goal is to automatically generate the minutes $\mu_i = (\mu_1, \ldots, \mu_n)$ given the transcript $\tau = ((\rho_1, \delta_1), (\rho_2, \delta_2), \ldots, (\rho_{N_i}, \delta_{N_i}))$ for a specific agenda α of meeting (see Table 1).

4 Methods

Here we cover details of the end-to-end summarization models with the goal to maximize the conditional probability $P(\mu|\tau;\eta)$ of minute μ given a meeting transcript τ and neural network parameters η.

4.1 Extractive Methods

Given a transcript, extractive methods are supposed to select a subset of the words or sentences which best represent the discussion of the meeting. In this section, we study these extractive methods to generate minutes for a meeting automatically.

- **TF-IDF** (Christian et al., 2016) receives the input transcript for pre-processing and removes all the stopwords, stemming, and word tagging. Further, calculates their TF-IDF value and cumulate across each sentence, highest-scoring top-n selected as minutes.

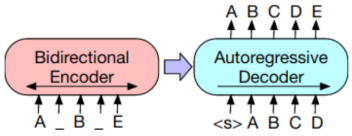


Figure 3: A systematic diagram from BART (Lewis et al., 2019)

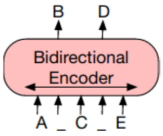


Figure 4: A schematic comparison of BART in Figure 3 with BERT from (Devlin et al., 2018)

- **Unsupervised**, is a heuristic approach, where we use different hand-crafted features (such as word frequency, cue words, numeric data, sentence length, and proper nouns) to rank the sentences. Sentences above a given threshold are selected into the minutes.
- **TextRank** (Mihalcea and Tarau, 2004) is a text summarization technique based on a graph algorithm. The input transcript has individual sentences, each represented by a vector embeddings. The similarity (refer to PageRank algorithm (Xing and Ghorbani, 2004)) between each sentence vector is stored in a matrix and converted into a graph. The graph represents sentences as vertices and similarity score as edges. The top-ranked sentences formulate the minutes for a particular transcript.
- **LexRank** (Erkan and Radev, 2004) is another text summarization technique based on a graph algorithm. It is similar to TextRank, but the edges between the vertices have a score obtained from the cosine similarity of sentences represented as TF-IDF vectors. A threshold takes only one representative of each similarity group (sentences similar enough to each other) and derives the resulting minute for the given transcript.
- **Luhn Algorithm** (Luhn, 1958) is one of the oldest algorithms proposed for summarization based on the frequency of words. It is a naive approach based on TF-IDF and focussing on the "window size" of non-important words between words of high importance. It also assigns higher weights to sentences occurring near the beginning of a document.
- **LSA: Latent Semantic Analysis (LSA)** (Gong and Liu, 2001) algorithm derives the statistical relationship of words in a sentence. It combines the term frequency in a matrix with singular value decomposition.

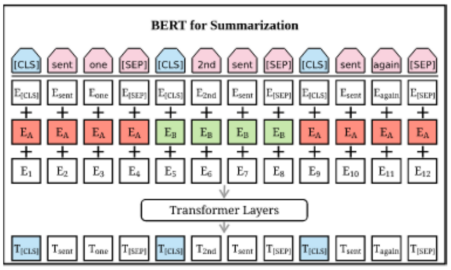


Figure 5: An illustration of BERTSUM from (Liu and Lapata, 2019)

4.2 Abstractive Methods

Given a transcript, the task is to generate a concise minute that captures the salient notions of the meeting. The generated abstractive minute potentially contains new phrases and sentences that have not appeared in the meeting transcript.

- **BART** (Lewis et al., 2019), uses the basic seq2seq architecture with bidirectional encoder as in BERT (refer to Figure 4) with additional left-to-right denoising autoencoder (refer to Figure 3). The pretraining of seq2seq tasks involves a random shuffling of the original transcript and a novel in-filling scheme, where text spans are replaced with the mask token value. It exhibits a significant performance gains when fine-tuned for text generation and comprehension tasks.
- **BERTSUM** (Liu and Lapata, 2019) is an extension to BERT (Devlin et al., 2018) with novel document-level encoder which has multiple [CLS] symbols injected to input document

sequence for memorizing sentence representations. Additionally, it applies interval segmentation embedding (illustrated in Figure 5 with red and green color) to distinguish multiple sentences. These embeddings are summed and given as input to several bidirectional transformer layers, generating contextual vectors and further decoding. Additionally, there is new fine-tuning schedule which adopts different optimizers for the encoder and decoder for alleviating the mismatch(as the encoder is pre-trained while decoder is not).

- **BERT2BERT** (Rothe et al., 2020) uses BERT checkpoints to initialize encoder-decoder to provide a better understanding of input, mapping of input to context, and generation from context while the attention variable initialize randomly. While in this paper, we tokenize our data using WordPiece[3] to match the pre-training vocabulary for BERT as well as for noise consistency training and maintaining copy to protect gradient propagation through it.

- **Longformer Encoder-Decoder (LED)** (Beltagy et al., 2020) is another variant for longformer which supports long document generative seq-2-seq task. This encoder-decoder model has its attention mechanism, combining local window attention with task-motivated global attention that supports larger models (with thousands of tokens).

- **Pegasus** (Zhang et al., 2020) uses transformer-based encoder-decoder model for sequence-to-sequence learning. In PEGASUS, important sentences are removed/masked from an input document and are generated together as one output sequence from the remaining sentences, similar to an extractive summary.

- **Roberta2Roberta** (Liu et al., 2019) is an encoder-decoder model, meaning that both the encoder and the decoder are RoBERTa models. In this work, we initialize the Roberta-large model with checkpoints. It involves pre-training with the Masked Language Modeling (MLM) objective, where the model randomly masks 15% of the words in an input sentence and predicts them back based on other words in that sentence.

- **T5** (Raffel et al., 2019) is also an encoder-decoder transformer model. It can be easily pre-trained on a multi-task mixture of unsupervised and supervised, with each task converted in text-to-text format. In this work, we pre-train T5 by fill-in-the-blank-style with denoising objectives while using similar hyperparameters and loss functions.

5 Experiments

In this section, we describe the experimental details for off-the-shelf text summarization models for automatic minuting. We describe the hyperparameter setting for different models in Table 2.

5.1 Dataset

We base our experiments on two popular and one new dataset.

AMI For our experiments, we use the popular AMI dataset (Mccowan et al., 2005), which contains 100 hours of meeting discussions with their abstractive and extractive summaries. The audio recordings of all the meetings are provided with manually corrected transcripts. The AMI corpus contains a wide range of annotations such as dialogue acts and topic segmentation, named entities, and manually written meeting minutes. The AMI corpus consists of 138 meeting instances with their corresponding summaries.

ICSI corpus (Janin et al., 2003b) are mostly from regular meetings of computer science working teams. The corpus contains 70 hours of recordings in English (for 75 meetings collected in Berkeley during the years 2000-2002). The speech files range in length from 17 to 103 minutes and involve from 3 to 10 participants. Interestingly, the corpus contains a significant portion of non-native English speakers, varying in fluency from nearly-native to challenging-to-transcribe. All audio files are manually transcribed. ICSI consists of 75 meeting instances.

AutoMin [4] dataset is from the first shared task on

[3] https://github.com/google-research/bert/blob/master/tokenization.py

[4] https://elitr.github.io/automatic-minuting/cfp.html

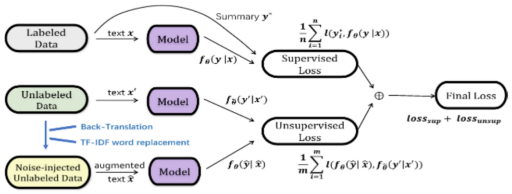


Figure 6: Illustration of BERT2BERT model for noised consistency training from (Liu et al., 2021)

Table 2: Hyperparameter settings and parameters of execution for the examined models

	Models						
Hyperparameter	**BART**	**BertSum**	**BERT2BERT**	**LED**	**Pegasus**	**Roberta2Roberta**	**T5**
learning rate	1e-5	1e-5	1e-5	1e-5	1e-5	1e-5	1e-5
weight decay	0.001	0.1	0.001	0.1	0.01	0.001	0.01
max. grad. norm	1.0	1.0	1.0	1.0	1.0	1.0	1.0
warmup steps	1300	500	300	500	1200	1400	500
batch size	24	32	32	24	48	32	32
max epochs	4	10	4	4	4	10	4
Runtime Parameters							
Python	3.7.3	3.7.3	3.7.3	3.7.3	3.7.3	3.7.3	3.7.3
GPU: GeForce RTX	2080 Ti	2080 Ti	2080 Ti	2080 Ti	3090	2080 Ti	2080 Ti
GPU count	1	1	1	1	1	1	1
GPU RAM (GB)	11	11	11	11	25	11	11
Machine RAM (GB)	248.8	248.8	248.8	248.8	183.0	248.8	248.8

automatic minuting at Interspeech 2021. It consists of manually created minutes from multiparty meeting transcripts. This dataset contains real project meetings in two different settings: technical project meetings (both in English and Czech) and parliamentary proceedings (English). We only use English data for our experiments which consists of 123 meetings with multiple minutes. For evaluation on AutoMin, we average out our scores.

5.2 Quantitative Evaluation

In this section, we evaluate all the generated outputs from different models described in Section 4 and show them in Table 3. We use the popular automatic summarization metrics like ROUGE (1, 2, L, WE) (Lin, 2004), BERTScore(Zhang et al., 2019) and BLEU (Papineni et al., 2002) which are lexical to evaluate the quality of the summary. The scores are averaged across the datasets. We see that in the abstractive methods, T5 performs best in terms of the metrics we took. It is based on Transfer learning, where a model is first pre-trained on "Colossal Clean Crawled Corpus" a data-rich task before being fine-tuned on a downstream task. It has been shown to achieve state-of-the-art results on many benchmarks covering summarization. The extractive summarization algorithms LSA performs best in the extractive methods as we analyzed. LSA algorithm exhibits the statistical relationship of words in a sentence, combining the term frequency in a matrix with singular value decomposition and therefore performs state-of-art results for AutoMin. However, these quantitative metrics indicate the quality of the generated summary by these various models across the different datasets. Along with the quantitative evaluation, we vouch for the qualitative assessment of the generated minutes.

5.3 Qualitative Evaluation

To assess the quality of the automatically generated minutes, we conduct a qualitative evaluation of those by human assessors. We evaluate the qualitative performance of both the extractive and abstractive methods that we employ for the meeting summarization task (see Section 4). We

Table 3: Quantitative Analysis of Baseline Abstractive and Extractive Summarization Methods. The highest score have been highlighted for a particular model across AMI, ICSI and AutoMin

Abstractive Approaches							
	Dataset	ROUGE_1	ROUGE_2	ROUGE_L	ROUGE_WE	BERTScore	BLUE
BART (Lewis et al., 2019)	AMI	18.29	3.42	9.95	3.33	29.47	20.63
	ICSI	6.68	00.28	3.58	0.00	43.91	20.26
	Automin	24.88	6.36	14.09	6.22	32.08	15.24
BERTSUM (Liu and Lapata, 2019)	AMI	13.25	1.73	7.42	2.19	29.59	25.37
	ICSI	5.01	0.17	2.87	0.062	4.87	20.89
	Automin	20.73	3.67	11.28	4.95	28.94	22.80
BERT2BERT (Rothe et al., 2020)	AMI	12.95	2.04	6.75	2.50	14.56	21.90
	ICSI	5.97	0.15	2.93	0.22	24.59	21.22
	Automin	23.51	5.19	12.03	6.22	19.42	15.54
LED (Beltagy et al., 2020)	AMI	5.51	0.52	4.09	0.46	43.57	34.66
	ICSI	1.45	0.03	1.12	0.02	22.34	**35.31**
	Automin	9.24	1.28	6.96	0.51	35.80	26.21
Pegasus (Zhang et al., 2020)	AMI	14.56	2.51	8.10	2.75	24.85	21.43
	ICSI	5.76	0.18	3.12	0.06	42.82	19.67
	Automin	22.72	4.55	11.97	4.66	29.12	16.68
Roberta2Roberta (Liu et al., 2019)	AMI	13.50	2.16	7.82	2.11	26.29	21.30
	ICSI	6.27	0.18	3.22	0.07	42.66	17.45
	Automin	16.67	3.12	9.48	3.13	28.09	28.90
T5 (Raffel et al., 2019)	AMI	16.14	2.70	9.00	2.92	34.34	22.44
	ICSI	5.99	0.21	3.32	0.02	**49.64**	20.77
	Automin	**27.01**	**6.71**	**14.63**	**7.59**	33.30	16.79
Extractive Approaches							
TF-IDF (Christian et al., 2016)	AMI	11.36	1.59	5.72	2.62	16.07	25.29
	ICSI	3.65	0.06	2.18	0.02	**48.71**	21.14
	Automin	19.06	3.29	8.45	3.63	25.30	22.43
Unsupervised	AMI	11.98	1.76	7.13	1.81	36.87	24.60
	ICSI	5.91	0.17	3.08	0.06	32.29	21.58
	Automin	23.45	5.04	12.96	2.68	29.93	22.60
TextRank (Mihalcea and Tarau, 2004)	AMI	10.12	1.56	5.33	2.22	8.62	24.74
	ICSI	5.94	0.12	2.85	0.05	19.28	21.64
	Automin	22.96	**5.45**	11.94	7.19	17.92	18.32
LexRank (Erkan and Radev, 2004)	AMI	10.81	1.52	5.97	2.45	12.56	25.39
	ICSI	5.03	0.11	2.82	00.03	31.62	19.72
	Automin	22.55	4.14	12.21	5.13	24.94	16.09
Luhn Algorithm (Luhn, 1958)	AMI	10.11	1.57	5.35	2.24	7.92	**26.16**
	ICSI	6.14	0.13	2.95	0.07	17.99	20.75
	Automin	22.55	4.14	12.21	5.13	24.94	19.05
LSA (Gong and Liu, 2001)	AMI	10.34	1.78	5.44	2.25	7.35	23.46
	ICSI	6.48	0.15	3.16	0.05	26.33	20.90
	Automin	**23.52**	**7.73**	**13.29**	**8.90**	14.61	22.43

ask our annotators to evaluate each automatically generated minute/meeting summary in terms of their *adequacy*, *fluency*, *grammaticality*, and *coverage* using the 5-star Likert rating scale (Likert, 1932). The annotators assign an integer from 1 (worst) to 5 (best) against each criterion to assess the *goodness* of the minutes. We had three annotators for the task evaluating a sample of randomly selected minutes from each of our three datasets generated by the different text summarization methods. We show our human evaluation of the automatically generated summaries in Table 4 by both abstractive and extractive methods. For each method, we average out the evaluations by our annotators on the sample instances. Kindly find the output samples in `https://anonymous.4open.science/r/minuting-baselines-AB22/README.md` We provide our annotators with the transcripts of the meetings and the corresponding minutes. Our annotators have at least a Master's degree and education in English. For *adequacy*, we ask our annotators to judge if the minute adequately sums up the main contents of the meeting. *Fluency* would refer to how fluent, coherent, and readable is the output minute text. *Grammaticality* would mean the grammatical correctness of the minute. Finally, by

Table 4: Qualitative Analysis of Baseline Abstractive and Extractive methods. The highest score have been highlighted for a particular model across AMI, ICSI and AutoMin

Abstractive Methods					
	Dataset	Adequacy	Fluency	Grammaticality	Coverage
BART (Lewis et al., 2019)	AMI	2.66	**3.33**	**4**	**3.33**
	ICSI	2.66	3	3.66	2.33
	Automin	**3**	3	3.33	3.33
BERTSUM (Liu and Lapata, 2019)	AMI	2.33	3.33	**4**	2.66
	ICSI	2	3	3	3
	Automin	**2.66**	**3.33**	3.66	**3**
BERT2BERT (Rothe et al., 2020)	AMI	**3.33**	3	**4**	3
	ICSI	3	**3.33**	3.33	3
	Automin	2.33	2.66	3.66	**3**
LED (Beltagy et al., 2020)	AMI	1	1	1	1
	ICSI	1	1	1.33	1
	Automin	**1.33**	**1.66**	**1.66**	**1.33**
Pegasus (Zhang et al., 2020)	AMI	2.66	**3.66**	**5**	3
	ICSI	3	2.66	3.33	**3.33**
	Automin	**3**	3	3.66	2.66
Roberta2Roberta (Liu et al., 2019)	AMI	2	2.66	3	2.33
	ICSI	**2**	**3**	**3.33**	1.66
	Automin	2	2.66	2.66	**2.33**
T5 (Raffel et al., 2019)	AMI	1.66	2	3.66	1.33
	ICSI	2	**3.33**	3.66	2.33
	Automin	**2.66**	3	**3.66**	**3**
Extractive Methods					
TF-IDF (Christian et al., 2016)	AMI	**1.66**	**2.33**	**2.66**	**2.33**
	ICSI	1.33	2	2.33	2
	Automin	1.66	2	2.66	2
Unsupervised	AMI	2	**3**	**3**	**2.33**
	ICSI	1.66	3	3	2
	Automin	**2.33**	2.66	3.33	2.33
TextRank (Mihalcea and Tarau, 2004)	AMI	**2.66**	2.66	**3**	**3.66**
	ICSI	1.33	**2.66**	2.85	2
	Automin	2	2.66	2.33	2.66
LexRank (Erkan and Radev, 2004)	AMI	**2.66**	**2.33**	**2.66**	**2.33**
	ICSI	1.66	2.33	2.33	2.33
	Automin	1.33	2.33	2.66	2.33
Luhn Algorithm (Luhn, 1958)	AMI	1	**2.66**	**2.66**	**3**
	ICSI	2	2.33	2.33	2.33
	Automin	**2.66**	2.66	3	3
LSA (Gong and Liu, 2001)	AMI	**2.33**	**3**	**3**	**3.66**
	ICSI	2.33	3	2.66	3
	Automin	1.66	2	2	2.66

coverage we ask the annotators to rate if the minutes cover the major topics in the meeting transcript.

We can see from Table 4 that the BERT-based models yield output that our annotators found better in terms of *Adequacy*, *Fluency*, and *Coverage*. BART, Pegasus, T5 score better in *Grammaticality*. Overall the scores are low for the *AutoMin* dataset as it is the only dataset that has minutes in the form of bulleted points; semantic coherence of texts is not a major priority there. However, AutoMin simulates the human minuting behavior on the fly during actual meetings. Output from the extractive methods scores comparatively less w.r.t. that of abstractive methods in human evaluation. The reason being that these extractive methods extract texts from the transcripts without regard to coherence, readability, or grammar; hence are not well ranked by our evaluators. However, we see that *TextRank* and LSA provide comparable coverage w.r.t. the deep neural-based abstractive algorithms. Each algorithm is motivated towards achieving a different objective in the generated summary, and hence there is no *one shoe fits all* algorithm for the minuting task. Hence it definitely calls for more fine-tuned algorithms towards this specific task.

6 Conclusion and Future Work

In this work, we perform an empirical analysis of several *off-the-shelf* text summarization models when applied in the task of automatic minuting. We see that automatic minuting is challenging and could not be well-addressed with the existing summarization models. Both our quantitative and qualitative evaluation reveals that the extractive models perform better than the abstractive ones. However, they are still far from being acceptable. To sum up, we intend to provide baseline evaluations to the community for this challenging task with this paper. As future work, we would want to explore a template-based extractive method to generate the meeting summary from the transcripts. Our investigation indicates that leveraging on BERTSum could be a plausible direction to probe next. In future we would try, if possible, speaker segmentation embedding (i.e. EA, EB, EC, ED ...) for BERTSUM model to reflect different speakers in multi-party dialogue, instead of interval segmentation embedding (i.e. EA, EB, EA, EB ...).

Acknowledgements

This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460 (ELITR) and the grant 19-26934X (NEUREM3) of the Czech Science Foundation.

References

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

Jean Carletta. 2007. Unleashing the killer corpus: Experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. *arXiv preprint arXiv:1803.10357*.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

Yun-Nung Chen and Florian Metze. 2012. Integrating intra-speaker topic modeling and temporal-based inter-speaker topic modeling in random walk for improved multi-party meeting summarization. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Sangwoo Cho, Logan Lebanoff, Hassan Foroosh, and Fei Liu. 2019. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. *arXiv preprint arXiv:1906.00072*.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

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.

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*.

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.

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.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Karl Moritz Hermann, Tomáš Kočiskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *arXiv preprint arXiv:1506.03340*.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003a. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003b. The icsi meeting corpus. pages 364–367.

Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. 2020. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3622–3631.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring sentence singletons and pairs for abstractive summarization. *arXiv preprint arXiv:1906.00077*.

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*.

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.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

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*.

Junnan Liu, Qianren Mao, Bang Liu, Hao Peng, Hongdong Zhu, and Jianxin Li. 2021. Noised consistency training for text summarization. *arXiv preprint arXiv:2105.13635*.

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*.

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

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.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

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

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Jean-Pierre Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. *arXiv preprint arXiv:1805.05271*.

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. *arXiv preprint arXiv:2004.12393*.

Wenpu Xing and Ali Ghorbani. 2004. Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pages 305–314. IEEE.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. *arXiv preprint arXiv:1902.00863*.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Discourse-aware neural extractive text summarization. *arXiv preprint arXiv:1910.14142*.

Klaus Zechner. 2001. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA.

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

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.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. *arXiv preprint arXiv:1907.03491*.

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

D Proposed Automatic Minuting Method (under review)

A Pipeline Method for Generating Minutes from Multi-Party Meeting Proceedings

Kartik Shinde†, Tirthankar Ghosal‡, Muskaan Singh‡, and Ondřej Bojar‡
†Indian Institute of Technology Patna, India
‡Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics,
Charles University, Czech Republic
†kartik_1901ce16@iitp.ac.in
‡(last-name)@ufal.mff.cuni.cz

Abstract

Automatically generating meeting minutes is a challenging yet time-relevant problem in natural language processing. With the manifold rise in online meetings nowadays, meeting minutes seem more important than ever. However, *automatic minuting* is not straightforward due to a variety of reasons: low-quality ASRs, summarizing long dialogue discourse, retaining topical relevance and coverage, handling redundancies and small talks, etc. In this paper, we document our investigations on a pipelined approach to automatically generate meeting minutes standing on the shoulders of BART trained on multi-party dialogue summarization datasets. We achieve comparable results with our simple yet intuitive method with previous state-of-the-art models. We make our codes available at https://anonymous.4open.science/r/automatic-minuting-using-BART.

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 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 (*coverage is probably more significant than coherence and conciseness*). Summarizing spoken multi-party dialogues comes with its own set of challenges: incorrect/noisy automated speech recorder (ASR) outputs, long discourse, topical shifts, the dialogue turns, redundancies and small talks, etc. Hence we deem automatic minuting to be more difficult than text summarization. Due to the variety of sub-problems associated with this task, we adopt a pipelined approach. Our method encompasses (i) pre-processing the ASR-generated meeting transcripts for redundancies and noises, followed by (iii) unsupervised topical segmentation, finally (iii) summarizing with BART (Raffel et al., 2019) trained on a large-scale dialogue summarization dataset. Our initial investigation yields encouraging results with resultant minutes resembling the human gold standard. Qualitatively, our system performance is comparable to models like HMNet (Zhu et al., 2020) and DialogLM (Zhong et al., 2021), the current state-of-the-art systems, which also are extensively large and resource-expensive. Our main contribution in this work is in establishing that one can develop a lightweight, easy-to-implement, efficient pipelined automatic minuting method by leveraging pre-trained language models on large-scale dialogue summarization datasets while also generating readable and adequate meeting minutes.

2 Related Work

There have been several recent contributions to the meeting and dialogue summarization literature. Early studies like (Chen and Metze, 2012) used intra-speaker topic modeling to improve summarizing multi-party meetings. Later in 2019, several approaches from Zhao et al. (2019); Liu and Chen (2019); Liu et al. (2019) brought to our attention the efficacy of hierarchical methods to learn the inherent structure of conversations. Li et al. (2019) demonstrates a multi-modal hierarchical attention mechanism across the topic, utterance, and word levels. However, it depends on manual annotation of topical segments and visual focus in meetings which are not commonly available. Zhu et al. (2020) introduces a hierarchical network 'HMNet' for an end-to-end training with cross-domain flexibility, which is one of the state-of-the-art models

for meeting summarization. Liu and Chen (2021) proposes a dynamic sliding window strategy for abstractive summarization that helps achieve close to state-of-the-art numbers. Zhong et al. (2021) presents a novel pre-training framework for long dialogue understanding and summarization with window-based denoising. Very recently Zhang et al. (2021) introduced a flexible multi-stage framework for longer input texts, combining a multi-stage greedy transcript segmentation into a simultaneous end-to-end training.

Most of the aforementioned end-to-end deep neural models involve high-cost computing, are resource-intensive, and run time-consuming complex algorithms. Our proposed pipelining approach is by far simple and consists of separate stages for each sub-task: preprocessing, segmentation, redundancy elimination, and summarization. Each stage has a unique problem to address, with specified target outputs, culminating in the final objective, i.e., minutes generation. Our pipelined approach also allows the user to monitor outputs at every stage, and have increased control over what parameters/hyperparameters to tune for subsequent phases in the pipeline. We would also like to point that the earlier methods do not aim for automatic meeting minutes, rather they strive to generate coherent meeting summaries in form of paragraphs. Our motivation is to generate meeting minutes in form of bullet points that adequately capture the principal components of the meeting. The AutoMin shared task[1] at Interspeech 2021 resembles our investigation objective.

3 Methodology

With an extensive analysis of various meeting, dialogue, and document summarization corpora, we design a pipeline that performs preprocessing of the input text, redundancy elimination from the processed data, segmentation, and abstractive summarization with an underlying summarization module. The outputs are again filtered using unsupervised redundancy elimination methods based on several factors to obtain the final summaries/minutes.

3.1 Preprocessing

Redundancy Elimination. Since current summarization models are not trained to eliminate such redundancies, alongside capped to certain input lengths for precise generation, they struggle to process a long sequence of multi-speaker utterances and the dispersed information that comes with them. The training approach in BART can handle noisy inputs efficiently, but the problem persists across all other models. We leverage some preprocessing methods and employ utterance cleaning and elimination based on some thresholds to tackle this issue.

Consider a transcript with Speaker-Utterance pairs, $X^0 = \{(p^0{}_1, U^0{}_1), (p^0{}_2, U^0{}_2), ..., (p^0{}_L, U^0{}_L)\}$, where $p^0{}_j \subset P$, $1 \leq j \leq L$, is a participant and $U^0{}_j = (w^j{}_1, w^j{}_2..., w^j{}_{l_j})$ is the tokenized utterance from p_j. For i-th utterance, $U^0{}_i = (w^i{}_1, w^i{}_2..., w^i{}_{l_i})$ in the transcript, we generate a cleaned sequence, $U^c{}_i = (W^i{}_1, W^i{}_2..., W^i{}_{L_i})$, by eliminating repetitions, pauses and masks like {vocalsounds}, {disfmarkers}, {unintelligible} and similar disruptions. These utterances are then filtered using a custom stopwords set, $S = \{s_1, s_2, ...s_n\}$, that we define from various meeting transcripts from currently available corpus like AMI (McCowan et al., 2005), ICSI (Janin et al., 2003), and the dataset from AutoMin 2021 shared task. Following this, we obtain the compressed utterance, $U^x = U^c \cap S'$ and corresponding context ratio, R as shown.

$$R = L(|U^x|)/L(|U^c|) \tag{1}$$

here, L(|Al) is the cardinality of set 'A'. Thus, a processed transcript is obtained by appending the utterances after applying a threshold over all the obtained context ratios.

$$X' = \sum_{i=1}^{L}[(p_i, U^c{}_i)|R_i \geq \alpha] \tag{2}$$

Linear Segmentation. As current summarization models limit the length of input sequences they can process, they cannot take a full-length transcript in our data as input. We adopt a brute-force approach here and employ a linear segmentation by slicing the transcripts into blocks of segments with a uniform token length. We choose to adopt varying token-lengths of (i)512, (ii)768, and (iii)1024 tokens, respectively, in pipeline configurations for capping the segments and subsequently append the inputs for inference.

Topical Segmentation. In order to avoid the issues caused due to absence of context in

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

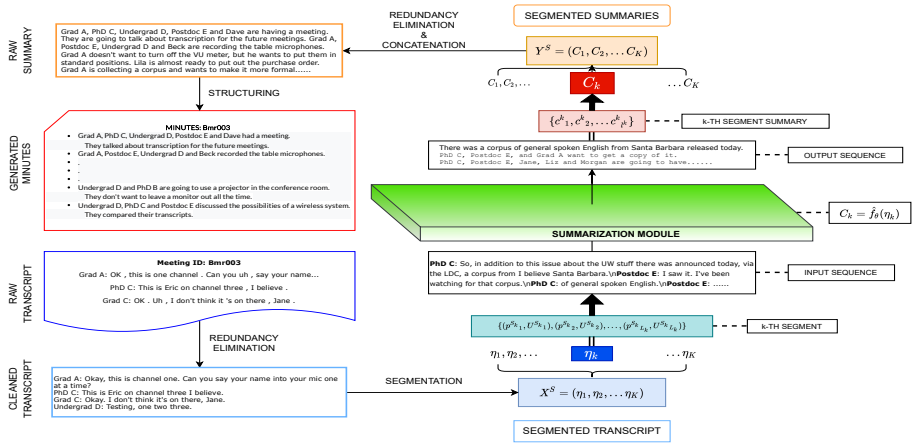


Figure 1: Architectural Representation of the proposed pipeline.

segments (like in linear segmentation), we employ several topical segmentation methods alongside Depth-Scoring and TextTiling algorithm (Hearst, 1993).

We inherit the depth-scoring method from (Solbiati et al., 2021). We use a sentence window of ($k_w = 10$), with an average segment-length cap of ($\hat{L} = 60$) and topic change threshold of ($\tau = 0.5$) (all tunable hyperparameters) for our experiments. For a transcript with N-turns, we apply maxpooling on the list of utterances from the windows - $(k - k_w, k)th$ turns, and $(k, k + k_w)th$ turns; and obtain the cosine similarities between all such subsequent pair of windows as k ranges from $(k_W, n - k_W)$. For a series of windown-similarity scores $\hat{s} = (sim_{k_w}, ...sim_{n-k_W})$, the depth scores are computed using: $dp_k = \frac{hl(k)+hr(k)-2sim_k}{2}$ where $hl(k)$ and $hr(k)$ are the highest similarity score on the left and right of interval k. The topic change indices are computed with the help of the obtained window-similarity scores and depth scores. Following are the variations used while determining the topic change indices.

- **Segment-window capping.** Following this, the topic change indices are computed as shown.

$$T = \{i \in [0, M] | sim_{k_w+i} \leq \mu_s - \sigma_s\} \quad (3)$$

where, μ_s and σ_s are the mean and variance of the sequence, $M = n - k_W$ of block similarities $\hat{s}$.

- **TextTiling.** Following this, after computing the window similarity scores, we use the plain TextTiling method for computing the segments in a transcript. For a depth scores series $D = (d_1, d_2, ...d_{N-k_w})$, the topic change indices are determined as shown.

$$i_{topic} = \sum_{i=1}^{n-k_W} [(i) | d_i \geq \tau] \quad (4)$$

After implementing the segmentation, we obtain the a segmented transcript $X^S = (\eta_1, \eta_2, ...\eta_K)$ where $\eta_k = \{(p^{S_k}{}_1, U^{S_k}{}_1), (p^{S_k}{}_2, U^{S_k}{}_2), ..., (p^{S_k}{}_{L_k}, U^{S_k}{}_{L_k})\}$ is the set of speaker-utterance pairs belonging to that segment.

For each segment, we concatenate the speaker role strings with the utterances in a line-wise manner to simulate the structure of a dialogue. This step is done before passing the inputs to root summarization modules since we fine-tune the models on dialogue summarization datasets before incorporating them into the pipeline.

3.2 Summarization

We choose the pretrained BART model, from (Lewis et al., 2019), as the primary summarization module in our pipeline (best-performing one). We also experiment with summarization models using T5 (Raffel et al., 2019), Pegasus (Zhang et al., 2020), RoBERTa2RoBERTa (Rothe et al., 2020), etc. We fine-tune all the models, available

via Huggingface [2] [3], on well-known dialogue summarization datasets before integrating them into our pipeline. The hyperparameters, models configurations, and fine-tuning approaches used in our experiments are discussed in further sections.

BART is a denoising autoencoder for pretraining sequence-to-sequence models. The model is trained by corrupting text in an arbitrary noising function and then teaching it to reconstruct the original text. BART's ability to use bi-directionalism when operating on sequence generation tasks bolsters its use for text summarization. While BERT cannot adopt a bidirectional mechanism for sequence generation, BART exploits the GPT-2 (Radford et al., 2019) architecture for predicting the following words with the help of words encountered previously in the current sequence. Hence, we primarily test the pipeline with various BART-based setups.

We pass the input sequence obtained from the preprocessing module through the summarization module. Again, for k-th segment, it returns a summary $C_k = \{c^k{}_1, c^k{}_2, ...c^k{}_{l^k}\}$, where $c^k{}_i$ is the i-th summary line of the k-th segment. We rejoin all the segment summaries $Y^S = (C_1, C_2, ...C_K)$ to get the raw summary text.

3.3 Post-processing

To eliminate the redundancies in the outputs, we use a similar process as in section 3.1. Since the models are fine-tuned to provide fluent outputs, another elimination procedure seems needless. Instead, we use sentence compression methods, including swapping in shortened phrases, pronouns, and splitting longer sentences into two for more readability. Following this, for each summary line, we filter out a set of special entities (speaker names, project/corporations names, location details) and use a token-length threshold of ($\tau_{token} = 10$) to include only those summary sentences which are quantitatively informative enough (i.e., consisting a minimum of τ_{token} number of tokens).

4 Dataset Description

Our experiments comprise various abstractive summarization datasets throughout their course.

Fine-tuning of Summarization Module Here, we choose from some of the popular abstractive summarization datasets. Primarily, We use the dialogue summarization corpora like SAMSum (Gliwa et al., 2019) and DialogSum (Chen et al., 2021). These datasets are made-up of multi-party spoken dialogues with annotated abstractive summaries. SAMSum consists real-life messenger conversation, while DialogSum is derived from different sources like (i) DailyDialog (Li et al., 2017) (ii) DREAM (Sun et al., 2019) and (iii) MuTual (Cui et al., 2020), and an English speaking practice websites. These conversations can be formal/informal and may contain slang phrases, emoticons, and typos, and are usually short.

Alongside these two corpora, we also include MediaSum (Zhu et al., 2021), which consists of media interview transcripts sourced from the broad range of domains and their associated summaries/topics. These transcripts are longer and have more complex dialogues than the rest two corpora. Since interviews generally follow a set of pre-defined structures and topical discourse, we believe that such data would highly benefit the models.

Target Datasets - We use the well-known meeting datasets from the domain of meeting summarization: AMI and ICSI, as well as the AutoMin corpus. These datasets are sourced from staged product design meetings in companies, academic group meetings in schools, and similar meetings. Each instance has a transcription of the entire dialogue and is annotated with a meeting summary and human-identified topic boundaries (except in AutoMin). These meeting transcripts have an extremely long, turn-based structure and are rich in redundant information. Relevant and essential information, however, is dispersed throughout the transcript. Table 1 shows all the relevant statistics of the dialogue summarization and meeting summarization datasets used in our experiments.

A big challenge while processing spoken dialogues is the difference in their information flow as compared to a monologic text, which is intuitively reflected in the dialogue discourse structures (Wolf and Gibson, 2005). For example, two utterances can be closely related even with a significant distance between them. Due to the unique structure of the spoken dialogue, important information is rather dispersed. Naive sequence-to-sequence generation methods subsequently prove to be useless in the case of such datasets, which is reflected by the poor ROUGE scores of the LEAD baseline

[2] https://huggingface.co/models
[3] https://github.com/huggingface/transformers

Datasets	# Dialogues	# Turns	# Speakers	# Avg. Turn Len.	# Len. of Dialogue	# Summary Len.	%-Compression
SAMSum	16.4K	11.2	2.4	9.1	124	23.4	81.12%
DialogSum	13.5K	9.5	2	15.8	168.5	25.8	84.7%
MediaSum	463.6K	30	6.5	49.6	1553.7	14.38	99%
AMI	137	535.6	4.0	10.4	5,570.4	321	94.24%
ICSI	59	819.0	6.3	10.5	8,567.7	576	93.28%
AutoMin	124	254.4	5.8	9.7	8,890.8	387	95.65%

Table 1: Statistics of the dialogue and meeting datasets being used.

Datasets	Examples	Doc. Len.	Summ. Len.	% Comp.	% novel unigrams
XSum	226K	488	27	94.5%	37.8%
CNN/DM	311K	906	63	93%	16.9%
R-TIFU	7.9K	641	65	89.9%	43.84%

Table 2: Statistics of the document summarization datasets being used.

on SAMSum. Besides, interruptions too appear frequently in the middle of conversations, making the speakers' utterances incomplete and potentially destroying coherent discourse structures. Pragmatics and social common sense indirectly give a unique challenge in spoken language understanding and significantly impact summarization. For instance, humans can understand that the "Here you are" is actually "make a payment" and "Goodbye" indicates that the event "check out" is finished. It requires commonsense knowledge to understand such dialogues fully. Hence, these dialogue datasets are more challenging than the conventional document corpora.

Some of the evaluated models in our experiments also go through a finetuning phase with standard document summarization datasets like XSum (BBC articles) (Narayan et al., 2018), CNN/DailyMail (News Articles) (Nallapati et al., 2016) and Reddit-TIFU (informal subreddit posts) (Kim et al., 2018) prior to that on dialogue summarization task. Such a finetuning/warm-starting has been observed to benefit the abstractiveness and language understanding of the model as compared to plain finetuning. Figure 2 shows all the relevant statistics of these document summarization datasets used in our experiments. We can see the extremely abstractive nature of datasets like XSum and CNN/DM, which indicates that these datasets can potentially train the models to generate sequences more selectively, thus automatically eliminating redundancies.

5 Experimentation

Most of the summarization models are trained on a single Tesla K80 GPU. Few larger models like BART-large, T5-large require multi-GPU training on NVIDIA GTX 1050 Ti, or single GPU training on the NVIDIA A100-PCIE-40GB variant. Training for individual fine-tuning procedures takes less than 3 hours, while warm-starting takes approximately 1.5-2 hours, depending on the dataset used. The hyperparameters and model configurations are consistent with the default values used during the pretraining of respective models. We provide the hyperparamaters and model configurations on an anonymized repository: https://anonymous.4open.science/r/automatic-minuting-using-BART.

6 Evaluation

We experiment our pipeline with the different summarization modules, fine-tuned on combinations of abstractive summarization datasets, and report our performance on AMI, ICSI, and AutoMin meeting summarization/minuting data.

6.1 Automatic Evaluation

For automatic evaluation, we make use of popular text summarization evaluation metrics. We report in terms of ROUGE-1, ROUGE-2, ROUGE-SU4, which measure the overlap of unigrams, bigrams, and unigrams plus skip-bigrams (with max. skip of 4), respectively. We also provide the METEOR(Banerjee and Lavie, 2005) scores which reward matching stems, synonyms, and paraphrases within.

6.2 Human Evaluation

To evaluate the quality of our output, we carry out a human evaluation of our minutes and compare it with the best-performing model outputs from the AutoMin 2021 shared task. We contacted the AutoMin organizers and human evaluators from the AutoMin shared task, who have then rated our minutes in terms of Adequacy, Grammaticality, and Fluency scores on a Likert scale of 5. We attach more importance to human evaluation than automatic evaluation in this task as automatic metrics for text summarization evaluation have various shortcomings and are not apt to judge the quality of meeting minutes (Ghosal et al., 2021).

	AMI			ICSI		
Model	R-1	R-2	R-SU4	R-1	R-2	R-SU4
Random	35.13	6.26	13.17	29.28	3.78	10.29
Cluster Rank (Garg et al., 2009)	35.14	6.46	13.35	27.64	3.68	9.77
Extractive Oracle	39.49	9.65	13.20	34.66	8.0	10.49
PGNet (See et al., 2017)	40.77	14.87	18.68	32.0	7.7	14.46
HMNet (Zhu et al., 2020)**	53.02	18.57	24.85	46.28	10.6	19.12
DialogLM (Zhong et al., 2021)	53.7	19.6	-	49.5	12.5	-
Summ^N (Zhang et al., 2021)	53.4	20.3	-	48.8	12.2	-
bert2bert-cnndm-samsum	40.72	10.1	**27.13**	35.03	7.35	**24.48**
bart-xsum-dialogsum	42.4	10.34	17.67	36.95	6.94	13.68
t5-dialogsum	42.71	11.05	18.34	37.01	7.48	13.68
bart-xsum-samsum*	**45.17**	**13.3**	20.33	**38.75**	**8.51**	14.98

Table 3: ROUGE-1, ROUGE-2, ROUGE-SU4 scores of generated summary in AMI and ICSI datasets. The first partition separates the baselines from the expensive systems. The second partition separates our setups from all other previous models. *Our best-performing setup incorporated using the pipeline approach. ('bart-xsum-samsum' stands for a model finetuned on the XSum corpus (Narayan et al., 2018), followed by further finetuning on the SAMSum corpus(Gliwa et al., 2019) .) **Scores of highly cost-expensive models in comparison with our approaches.

	Automatic Evaluation			Human Evaluation		
Model	R-1	R-2	R-L	Adequacy	Grammatical	Fluency
Ours-bart-xsum-samsum	**40.45**	**11.27**	**18.24**	**4.46/5.00**	**4.45/5.00**	**4.18/5.00**
AutoMin system #2	23.94	9.19	15.84	4.25	4.34	3.93
AutoMin system #3	22.19	3.66	11.97	2.88	2.84	2.94
AutoMin system #4	20.93	5.46	12.61	2.32	2.64	2.52

Table 4: Performance of our pipeline in comparison with other participating systems at the AutoMin Shared Task.

Model	Pk	WinDiff	ROU-1	MTR
Random	0.61	0.75	-	-
TextTiling	0.39	0.41	43.4	18.1
Capped	**0.34**	**0.35**	42.5	16.7
Linear (768)	0.44	0.5	**45.17**	**20.6**

Table 5: Comparison of score with different segmentation methods

Model	R-1	R-2	R-SU4	BERT*	MTR**
bart-xsum-samsum	**45.2**	**13.3**	**20.3**	**0.60**	**20.6**
bart-xsum-dialogsum	42.4	10.3	17.7	0.59	18.6
bart-base-samsum	39.9	11.2	16.1	**0.60**	15.1
bart-base-mediasum	33.2	7.0	11.3	0.55	14.0

Table 6: Comparison of the BART-based setups with different fine-tuning datasets. *BERT stands for BERT-score. **MTR stands for METEOR.

Furthermore, we also carry out ablation experiments over our best-performing pipeline to check the effectiveness of segmentation methods used in prior experiments. We report the performance on the AMI dataset, which comes with reference segmentations of transcripts. We use the Pk (Beeferman et al., 1999) and WinDiff (Pevzner and Hearst, 2002) methods to evaluate the segmentation accuracy and report ROUGE-1 and METEOR scores on the model-segmented AMI transcripts due to their relevance in the task.

6.3 Results and Analysis

We evaluate the pipeline with more than 20 combinations of summarization models, segmentation approaches, hyperparameters, and different warm-starting setups (as discussed in Section 4). We calculate and compare the results of our pipeline with the various baselines and comparing systems mentioned in previous sections. We show the results of the best-performing pipeline configurations in the tables below. For more extensive results (i.e. including the all the configurations of our pipeline method), kindly visit the given repository : https://anonymous.4open.science/r/automatic-minuting-using-BART

Table 6 shows the performance of our pipeline when embedded with different summarization modules. Several setups show a fairly higher score across all the datasets in the experiments. Our best performing model outscores the previous similar approach by almost 5 points on ROUGE-1, and other setups perform close to the previous approaches. Table 3 compares the ROUGE scores of earlier models with our best setup (bart-XSum-samsum + linear). Our pipeline outperforms every earlier approach, except the HMNet (an exceptionally heavier model than the other models used). Table 4 compares the human evaluation score of models and baselines on the AMI dataset. Our model outperforms others by a significantly great margin on each indicator, proving the effectiveness and validity of the components in the pipeline. Such high scores on quality indicators clearly show how effective some rule-reinforced, unsupervised methods can generate standard meeting minutes.

By studying the outputs and scores of different combinations of setups used, we notice some case

Input: Transcript
Project Manager : Right uh . So um . So where's the PowerPoint presentation ? Sorry ? Microsoft PowerPoint , right . Right , okay . So . Right . Okay , so we've got uh so we've got new project requirements . Um . So basically we've got three things , and we've got forty minutes in which to uh {disfmarker} for this meeting to uh to discuss the various options . Um . Three presentations . **Industrial Designer** : We have a {disfmarker} I guess we have a presentation each , 'cause I've got one . Um . **User Interface**: Yeah . **Marketing** : Yeah . **Project Manager** : That's nice to know , one from each of you . Um new project requirements . Um so do we want to do the presentation rst ,or do we want to um {disfmarker} W I I got um {gap} or or three things basically , um relating to the remote being only for T_V_ . We discussed that last time **Industrial Designer** : Mm-hmm . **Project Manager** : and in actual fact that was pr pretty well what we came up with anyway . . . . **Industrial Designer** : Kay . **Marketing** : Okay . {vocalsound} {vocalsound} **Project Manager** : Come on .
Output: Minutes
• Industrial Designer explains to Project Manager how the remote works. They have their own logo, color scheme and they have to incorporate it on the product. The current expansion of channels in the process of taking place, certainly the button up, down, but how many channels do we have to prepare for and how many buttons do we need? Marketing, Project Manager, Industrial Designer, and User Interface discussed the design of a new TV remote control. • Marketing, Project Manager, Industrial Designer and Project Manager discussed the design of a remote control for the television. The Project Manager, Industrial Designer, Marketing and User Interface discussed the remote control functions. • Industrial Designer will put delay to allow multiple numbers. • Project Manager, User Interface, Marketing and Industrial Designer discussed the design of the remote control. • Marketing, Industrial Designer, Project Manager and User Interface discussed the design of remotes for older televisions. There are 16 buttons on the front of the device: power button, standby, channel, volume, up, down, brightness, contrast, and a mute. On the back it will have the channel tuner, brightness and contrast. • The project manager did the minutes of the meeting.

Table 7: Sample minute generation instance from AMI corpus: Meeting Id-ES2014b

patterns that almost every model conformed to. 1.) The experiments reveal that BART-based models usually show the best performance for any fine-tuning dataset used. Given the versatility of its architecture and noise-handling ability, the BART-based setups are most likely to perform better on this task than other models. 2.) The models fine-tuned on the SAMSum corpus offered a better generation quality than the other datasets. We attribute this to the fact that the dialogues from the SAMSum dataset are relatively simplistic in nature than the dialogues from datasets like DialogSum and MediaSum. The conversations are short and artificially curated, which overall helps the models to understand the semantics effectively.

On analyzing the human evaluation scores, we also notice the differences caused by the training datasets used before finetuning phase. Datasets like XSum demonstrate an extremely abstractive nature of summarization. Although the source text in XSum (Narayan et al., 2018) is longer than the dialogue instances from datasets like SAMSum (Gliwa et al., 2019), the summaries are relatively concise. A similar difference is observed when the model is trained on the XSum dataset as compared to other datasets like the Reddit-TIFU (Kim et al., 2018) and the CNN/DailyMail(Nallapati et al., 2016). The obtained minutes are relatively short with more novel word % and paraphrased sentences - qualities that are vital in relation to abstractive summarization.

Another observation is the effect of segmentation approaches on the minutes. We find the ROUGE(Lin, 2004) and METEOR(Banerjee and Lavie, 2005) scores surprisingly higher with linear segmentation (segments capped with 768 tokens). Table 5 shows the performance of segmentation approaches along with the corresponding pipeline summaries. Although the employed segmentation methods perform satisfactorily, they show low scores on ROUGE when used with the pipeline. A possible explanation for this can be that the fine-tuned model tends to elaborate when the fed dialogues are concise. Moreover, due to the topical segmentation strategy, several segments had snippets of chitchats and irrelevant information from the meeting. These otherwise ignored small talks subsequently reflect in the generated summaries, thereby affecting the overall scores of their setups.

Table 7 shows a sample generation instance from the pipeline. The transcript corresponds to 'ES2014b' from the AMI dataset. As it stands, the above generated minute is perfectly coherent with the discussions from the meeting, brief and concise like a human-annotated minute. It has a relevant topic structure with bullets.

Case-1: Made-up entities
Instance - *"PhD A PhD F, PhD C and PhD F are discussing the encoding of things with time and data."*
Explanation - It seems like as a normal summary line with correct grammar and readability. Although, on tracking the transcript, we find out the 'PhD C' is not a real speaker, 'Grad C' is the real speaker here. Hence, this is an error due to anonymization.
Instance - *"Marketing, Project Manager, Industrial Designer and Project Manager are meeting to...."*
Explanation - Here, since the Project Manager was a part of some segment and did not conform to any human-name form, the model mentions them twice in the same summary line. Again, an error due to anonymization.
Case-2: Absences of topical context
Instance - *"PhD D discovered that on the wireless ones, you can tell if it's picking up breath noises..."*
Explanation - Although the pipeline manages to capture the context of every discussion in a transcript most of the time, some cases like this persist. On reading the given summary, one may not understand discussed wireless device during the meeting. Hence, the error conforms to the second type - an absence of topical context.
Case-3: Incomplete phrases
Instance - *"they don't match well with the operating behavior of the | Marketing, Industrial Designer, Project Manager are discussing the design of the remote control"*
Explanation - Due to interruptions during utterances, the transcripts sometimes fail to capture the entire line in one utterance - often continued with a hyphen. This reflects in the model outputs as shown with a '|' separator.
Instance - *"They have decided to start with the black and white version. They will use double A or triple A batteries, rubberized buttons, a plastic casing for the plastic shell, a variety of designs, | Marketing Project Manager, Industrial Designer, User Interface and Project Manager are discussing the design of a keychain."*
Explanation - This instance is another example of the error type explained above.

Table 8: Error instances from the pipeline-generated summaries corresponding to each error case discussed in section 6.4

6.4 Error Analysis

Although the proposed approach shows a satisfactory performance on the task, we qualitatively examine and find the outputs conforming to several types of errors enlist below. Table 8 shows several instances for each error case discussed below.

- **Made-up entities**. Anonymization of discrete entities in transcripts (e.g., LOCATION7, PERSON4, Marketing Manager) is consistent in most organizations. Since no such anonymization methods are used during curation of the SAMSum dataset, this sometimes results in the generation of made-up entities that are initially not part of that transcript.
- **Rare absences of topical contexts**. The pipeline allows us to repair this absence by varying the token intake length of the underlying summarization module. However, this issue did not appear in generated minutes as the pipeline efficiently captures the discussions' topics from previous segments.
- **Incomplete phrases**. We also notice scarce occurrences of some incomplete sentences. These generally belong to those parts of the transcripts where the utterances either had missing punctuation or hesitations and interruptions on the speaker's part.

7 Conclusion

In this paper, we explore how we could use LLMs trained on dialogue summarization datasets to generate meeting minutes automatically. We evaluate our proposed BART-based pipeline approach on several multiparty meeting summarization datasets. Our initial performance is promising and certainly puts up a case for further investigations to employ large language models for this challenging task. In future work, we would like to optimize our existing pipeline by replacing extractive filtering and utterance-level topic segmentation with an end-to-end method.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1):177–210.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Yun-Nung Chen and Florian Metze. 2012. Integrating intra-speaker topic modeling and temporal-based inter-speaker topic modeling in random walk for improved multi-party meeting summarization. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. *arXiv preprint arXiv:2004.04494*.

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.

Nikhil Garg, Benoit Favre, Korbinian Reidhammer, and Dilek Hakkani Tür. 2009. Clusterrank: a graph based method for meeting summarization. Technical report, Idiap.

Tirthankar Ghosal, Muskaan Singh, Anja Nedoluzhko, and Ondřej Bojar. 2021. Report on the sigdial 2021 special session on summarization of dialogues and multi-party meetings (summdial). In *ACM SIGIR Forum*, volume 55, pages 1–17. ACM New York, NY, USA.

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

Marti A Hearst. 1993. Texttiling: A quantitative approach to discourse segmentation. Technical report, Citeseer.

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

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2018. Abstractive summarization of reddit posts with multi-level memory networks. *arXiv preprint arXiv:1811.00783*.

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*.

Manling Li, Lingyu Zhang, Richard J Radke, and Heng Ji. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *57th Conference of the Association for Computational Linguistics*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

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

Zhengyuan Liu and Nancy Chen. 2019. Reading turn by turn: Hierarchical attention architecture for spoken dialogue comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5460–5466.

Zhengyuan Liu and Nancy F Chen. 2021. Dynamic sliding window for meeting summarization. *arXiv preprint arXiv:2108.13629*.

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

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. In *Proceedings of the 5th international conference on methods and techniques in behavioral research*, volume 88, page 100. Citeseer.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

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*.

Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

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

Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Cali. 2021. Unsupervised topic segmentation of meetings with bert embeddings. *arXiv preprint arXiv:2106.12978*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational linguistics*, 31(2):249–287.

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.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed H Awadallah, Dragomir Radev, and Rui Zhang. 2021. Summˆ n: A multi-stage summarization framework for long input dialogues and documents. *arXiv preprint arXiv:2110.10150*.

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*, pages 3455–3461.

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

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. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. *arXiv preprint arXiv:2004.02016*.

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

EVENT REPORT

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 crowd-sourcing 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.

F Automin Overview Paper (to be published soon)

Overview of the First Shared Task on Automatic Minuting (AutoMin) at Interspeech 2021

Tirthankar Ghosal, Ondřej Bojar*, Muskaan Singh and Anja Nedoluzhko*

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Czech Republic

last-name@ufal.mff.cuni.cz

Abstract

In this article, we report the findings of the First Shared Task on **Automatic Minuting (AutoMin)**. The primary objective of the AutoMin shared task was to garner community participation to automatically create minutes from multi-party meetings. The shared task was endorsed by the International Speech Communication Association (ISCA) and was also an Interspeech 2021 satellite event. AutoMin was held virtually on September 4, 2021. The motivation for **AutoMin** was to bring together the Speech and Natural Language Processing (NLP) community to jointly investigate the challenges and propose innovative solutions for this timely yet important use case. Ten different teams from diverse backgrounds participated in the shared task and presented their systems. More details on the shared task can be found at `https://elitr.github.io/automatic-minuting`.

Index Terms: automatic minuting, meetings, multi-party dialogues, speech, summarization

1. Introduction

The COVID-19 pandemic has forced a substantial part of working population go virtual, especially those from Information Technology (IT), IT-enabled services, academia, etc. Among the many other challenges while adapting to the new normal, one crucial challenge was to enable smooth coordination among the employees (remote/hybrid), students, etc. By all means, meetings are the most vital component to ensure collaborative work and efficient to-and-fro communications. Hence, virtual meetings became more frequent and seamlessly got integrated to our daily routine. Thanks to the various remote meeting tools, in spite of the changed way of person-to-person interactions, people could still continue their collaborative work activities (at least to some extent). However, this also gave rise to a completely different set of problems, of which frequent meetings and unsettled work-life balance stands tall. Continual meetings and frequent context switching create an exorbitant information overload on the meeting participants. It is difficult to remember and recollect all the key information, decisions, action points, etc. from the meetings and more so if they are back-to-back or recurring. Hence writing minutes, or *minuting* for short, is an important activity in meetings (be it in-person or virtual).

Usually there is a designated person who jots down the *minutes of the meeting*, an external scribe or a participant from the meeting. However, taking running notes in parallel while being attentive to the meeting proceedings is a difficult job, and sometimes can distract attention from the meeting or waste other participant's time when waiting for the note-taker. Hence automated solutions to assist humans to efficiently jot down the meeting notes, action points, decisions, etc. would be a very useful NLP application. We are intrigued with the possibility of an AI system automatically generating the minutes of the meeting and sending them to the participants after the meeting. Or more realistically, such an AI system could create an initial minutes draft that would assist the participants to collaboratively revise and generate the final minutes. *How convenient would it be to just hover over past calendar invites to get the automatically generated summary of the meeting?* Such an application would also help the late joiners or those who missed the meeting to stay abreast with what happened in the meeting when they were not there. Hence, *Automatic Minuting* would be a super helpful NLP application for the working population. Our **AutoMin** shared task is a first step in this direction.

Minuting as an NLP task is closely related to summarization, however, they are not exactly the same. While text summarization is motivated towards generating a coherent, precise summary of the given textual content (news articles [1], scientific documents [2], dialogues [3], etc.), minuting is exclusively for meetings. Meeting minutes are usually free-form texts, often structured into bullet points lists, with probably less emphasis on textual coherence but more on coverage [4].

It is desirable that minutes capture the important aspects of the meeting in a concise way but it is more important not to leave out any topic of significance that was discussed in the meeting (obviously, small talk or casual chat that are unrelated to the meeting topic or agenda should be left out and should not be a part of the minutes). Hence, coverage and readability are perhaps the more important aspects in minuting.

Also, it is desirable that minutes include speaker names and possibly selected significant utterances from the central person or participants in the meeting. For instance, utterances from the project lead would probably be more salient than those of a new intern in a project meeting to appear in the minutes; with obvious exceptions.

Automatic minuting will also depend on the quality of the transcripts produced by automatic speech recognition (ASR). Although ASR quality has seen great improvements in recent years [5], still there are several sore points, such as handling speech from non-native speakers, multilingual speech, noise and artifacts of noise cancellation methods, etc. [6]. All these speech-related phenomena make minuting different from and likely more challenging than written text summarization.

Designing methods of automatic minuting is further complicated by the fact that there is no universal framework for creating minutes even by humans and desired outputs vary across

*equal contribution

different types of meetings, subjects, and objectives. Minuting is also a very subjective exercise and depends on the perspective of the note-taker. Two persons taking minutes can arrive at significant differences in content [4].

Furthermore, different participants in a meeting would have different information needs (a project leader vs. a team member vs. an administrative person). Also, the quality of minutes significantly varies depending on whether they are taken by an active participant or later by a non-participant [4]. A non-participant or an external scribe who would jot down the minutes after listening to the meeting recording can easily miss the context which is essential to comprehend the meeting content.

In terms of evaluation, there is no agreed upon framework via which we can measure the *goodness* of the minutes. People are still using conventional text summarization evaluation measures which are not meeting-specific and are also found to be not very effective in evaluating spontaneous speech summaries [7].

With all these challenges in mind, we launched our First Shared Task on Automatic Minuting, AutoMin 2021. The goal was to involve the community to take up this important challenge, make the first step, and ignite research interest in this problem.

Our AutoMin shared task consisted of one main task and two supporting tasks, relying on a dataset of transcripts and minutes from mostly technical meetings in English and Czech. A closely related special session on Speech Summarization was carried out in 2006[1]. With the increased dominance of deep learning and large language models in Speech and Natural Language Processing (NLP), we thought that probably it is a right time to launch a shared task effort on this important problem.

Some unique features of AutoMin 2021 were:

- the first shared task on generating minutes from real multi-party meetings,
- a meeting dataset on a language (Czech) other than English,
- multiple reference minutes created by different annotators, to allow observing the variance of outputs when humans are carrying out the task,
- source-based manual evaluation, to avoid evaluation bias which would be induced by a particular reference minute.[2]

With the first AutoMin and its proposed successive iterations, we aim to bring the interested NLP community in one platform and also rejuvenate the common interest in the topic of automatic generation of minutes from multi-party meetings.

2. Earlier Efforts and Related Literature

Meeting summarization as a problem came into light in the early 2000's. The AMI [8] and ICSI [9] datasets were the first publicly available datasets for research on multi-party meetings which also included summarization. The AMI Meeting corpus [8] contains 100 hours of meeting discussions, two thirds of which are, however, meetings acted artificially according to a scenario. The open-source corpus contains audio/video recordings, manually corrected transcripts, and a wide range of annotations such as dialogue acts, topic segmentation, named entities, extractive and abstractive summaries. The ICSI corpus [9] contains 70 hours of regular computer science working teams meetings in English. The speech files range in length from 17 to 103 minutes and involve from 3 to 10 participants. Interestingly, the corpus contains a significant portion of non-native English speakers, varying in fluency from nearly-native to challenging-to-transcribe. Other meeting collections are substantially smaller (e.g., NIST Meeting Room [10] or ISL [11]), unprocessed (e.g., various official meetings or recorded debates), or do not represent well the "project meetings" domain (e.g., proceedings of parliaments or city councils).

Klaus Zechner's seminal thesis on summarization of meeting speech and dialogues helped to shape the investigations in this topic further [12]. However, the NLP community did not witness much efforts in this problem after that, especially in terms of resource creation. The difficulty in resource creation can be majorly attributed to the several privacy issues including sensitive, personal information discussed in meetings [4]. More recently, there have been efforts towards developing large-scale multi-party dialogue/speech summarization datasets which can be leveraged for meeting summarization, e.g., MediaSum [13], SAMSum [14], CRD3 [15], MultiWOZ [16], Spotify podcast [17], doctor-patients conversations [18], DialogSum [19], etc. The *public meetings* [20] corpus is another recent resource for summarizing multi-party meetings in French.

Shared tasks and challenges played an important role to help evolve the present thriving text summarization community over the years. These campaigns or leaderboards leveraged on joint community efforts to solve a multitude of problems. For a success, the task has to be well-defined and backed by training and test data, allowing to compare the latest state-of-the-art techniques on a common platform. The summarization tasks in Document Understanding Conferences (DUC, 2001-2007) [21], several scientific document summarization challenges [22] in the Scholarly Document Processing (SDP) [23, 24, 25] workshops, the more recent DialogSumm challenge [26], the Financial Narrative Summarization challenge [27] are several examples of such activities in closely related areas.

Our AutoMin challenge is motivated along similar lines. We envisage AutoMin to evolve as a platform for community investigation into tasks pertaining to automatically generating minutes from multi-party meetings.

3. Task Descriptions and Evaluation Procedure

In AutoMin 2021, we proposed one main task (A) and two subsidiary tasks (B, C). The subsidiary tasks were optional but encouraged and their goal was to study the subjectivity associated with taking minutes (different people produce different minutes). Along with English, participants were encouraged to submit their system runs for the Czech portion of the data, which we made available for all the three tasks.

The provided dataset is detailed in Section 4 below.

3.1. Task A

*The **main task** consisted of automatically generating minutes from multiparty meeting conversations* provided in the form of transcripts. The objective was to generate minutes as bulleted lists, summarizing the main contents of the meeting, as opposed to usual paragraph-like text summaries.

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

[2] We use the common English word "minutes" to refer to a meeting summary in general. In cases where we need to highlight the existence of multiple such summaries for a given meeting, we also use the non-standard singular "a minute" to refer to one of them.

3.2. Task B

Given a pair of a meeting transcript and a manually-created minute, the task was 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 which 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. Evaluation Procedure

We evaluated the participant system-generated minutes (Task A) manually against the input transcript and also automatically against manually-created reference minutes via automatic text summarization metrics. Human evaluation should be treated as the primary one, because we agree that automatic text summarization metrics are not suitable to evaluate the quality of the candidate minutes.

On purpose, we do not provide any final ranking of systems in a form of leader board. We see AutoMin as a forum to encourage an inclusive community participation, exchange of ideas to stimulate the research rather than as a competition in a particular evaluation measure.

3.4.1. Human Evaluation of Task A

For the **manual evaluation** of Task A, we used three quality criteria which are common for evaluating text samples produced by automatic language generation systems. Our human evaluation metrics were: *adequacy*, *fluency*, and *grammatical correctness*.

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.

In each of these criteria, the evaluators rated the minutes on a Likert Scale [28] of 1 to 5 where 1 signifies the worst and 5 signifies the best output. Furthermore, we asked the evaluators to try to assess each of these qualities as independently of the other ones as possible.

Unlike usual summaries, we put less emphasis on paragraph-like continuous text because we believe meeting minutes are more practical in the form of lists.

The manual evaluation was carried out by our several external evaluators ensuring that each minute was evaluated independently by two of them.

To summarize the multiple evaluations of a given minute, we report both the averaged score as given by multiple evaluators as well as the maximum score the candidate minute has received.

We provided the evaluators with only the meeting transcript, not any of the reference minutes. Our manual evaluation is thus *reference-free*.

3.4.2. Automatic Evaluation of Task A

For our automatic evaluation of Task A, we relied on the widely popular text summarization metric ROUGE [29] in its three variants: ROUGE-1, ROUGE-2, ROUGE-L.

ROUGE metrics are based on n-gram similarities with a given reference. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It works by comparing an automatically produced summary against a reference summary (usually generated by a human). Different references thus inevitably lead to different ROUGE scores against each of them.

Recall in the context of ROUGE reflects how much of the reference summary the candidate summary is recovering or capturing:

$$\text{ROUGE}_{\text{Recall}} = \frac{\text{\# Overlapping n-grams}}{\text{Total n-grams in Reference Summary}} \quad (1)$$

Precision in the context of ROUGE means how much of the candidate summary was in fact relevant or needed:

$$\text{ROUGE}_{\text{Precision}} = \frac{\text{\# Overlapping n-grams}}{\text{Total n-grams in Candidate Summary}} \quad (2)$$

Despite the name ("Recall-Oriented..."), ROUGE actually commonly combines recall and precision using the harmonic mean to F-score. In our evaluation, we use ROUGE F1 scores for all ROUGE variants.

ROUGE-1 refers to the overlap of unigrams, ROUGE-2 refers to the overlap of bigrams, and ROUGE-L measures longest matching sequence of words using Longest Common Subsequence (LCS).

As discussed in Section 4, for many meetings, we had several reference minutes created by different annotators. We report both the average and also the maximum ROUGE-* score obtained by a candidate minute across the multiple references.

As we mention earlier, proper evaluation metrics for meeting summarization are severely needed [4] and text summarization metrics like ROUGE are only a poor alternative. Hence, we plan to launch an evaluation metric challenge in subsequent iterations of AutoMin.

3.4.3. Task B and C Evaluation

For the evaluation of Task B and Task C (which were basically classification tasks), we use F1 score (specifically that of YES-class) and Accuracy as our evaluation metrics. For Task B, YES class indicates that the minute belongs to a given meeting transcript. For Task C, the YES class signifies that two minutes belong to the same meeting.

Our F1 score is calculated as:

$$F1 = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

where Precision and Recall are for the YES class, in other

words:

$$\text{Precision} = \frac{\text{\# Correct YES Predictions}}{\text{Total \# of YES Predictions}} \quad (4)$$

$$\text{Recall} = \frac{\text{\# Correct YES Predictions}}{\text{Total \# of Actual YES Instances}} \quad (5)$$

Our Accuracy is:

$$\text{Accuracy} = \frac{\text{\# Correctly Answered Items}}{\text{\# Total Items}} \quad (6)$$

F1 score can be seen as more important because items in Task B and C datasets are predominantly 'NO' instances, see Sections 7.6 and 7.7.

4. Dataset Description

We prepared the dataset for the shared task mostly from various technical project meetings conducted either in English (EN) or in Czech (CS). Our data went through a series of pre-processing steps as described below. One of the goals to remove any personal data from the texts, so that the dataset does not interfere with the EU GDPR regulation.

When processing the data, we noticed that it can sometimes contain further potentially sensitive parts, beyond the requirements of GDPR (e.g., personal affairs discussed in small talk at the meetings). We thus decided to postpone full publication and provided the dataset only to registered participants of AutoMin upon signing a form of a non-disclosure contract. A modified version of the dataset called ELITR Minuting Corpus[3] is also being made publicly available, after additional de-identification checks and manual removal of these potentially sensitive parts of the text.[4]

The full processing sequence was this:

1. A preliminary consent to use meeting data was negotiated with meeting organizes and meeting participants. The consent was to record and process the full meeting data at Charles University with the foreseen publication of de-identified transcripts and minutes. We had a second round of explicit consents with publication from the meeting participants which we discuss later in this section.
2. The sound recordings from the online meetings were obtained.
3. The recordings were automatically transcribed using our own Automatic Speech Recognition (ASR) systems for English [30] and Czech [31].
4. We provided our annotators with the audio recording and the raw transcripts from the ASR. It demanded some manual effort to correct the ASR-generated transcript. We asked our annotators to carry out the following tasks:
   (a) Clean and correct the raw ASR-generated transcript (e.g., spelling corrections, mispronunciations, typos, etc.).
   (b) Break the transcript into smaller segments at natural linguistic points in the speech such as sentence or phrase boundaries, speech vs. silence/pauses, or speaker change. We warned that segmentation of spontaneous speech into sentences is often difficult and we expect that different transcribes could arrive at different segmentations. We nevertheless hope that with a fixed sequence of uttered words, the different segmentations should not affect minuting very often.
   (c) Diarize the transcripts, i.e., add speakers' codes ("PERSON1" etc.) at the beginning of each speaker's utterance in round brackets.
   (d) Format the transcript according to the agreed guidelines (in short: one sentence per line, focus on recognizing the sequence of words, preserve colloquial speech and speech errors, including errors in grammar, add punctuation and letter casing).

   Sometimes it required multiple rounds of communication between meeting participants and the annotators to resolve ambiguities. Since the annotators were not part of the actual meetings, hence sometimes they missed prior context in the meetings.
5. **Creating Minutes** After the transcript was manually corrected, the next step was to create the minutes. In most cases, it was external annotators who provided minutes for the meetings. For some meetings, we also had the minutes created by some of the meeting participants; these variants of minutes were more precise in the points they mentioned but they seemed less reliable in terms of coverage and structural match with the actual meeting.

   As mentioned earlier, we prepared multiple minutes for the same meeting to address the subjectivity associated with the problem.

   Our guidelines for annotators on how to prepare the minutes were rather broad, to get as realistic minutes as possible. We provided our annotators with examples of minutes and they were free to consult existing web resources on the topic. Our guidelines included general recommendations on creating minutes, such as being concise, concrete, avoid overusing person names, and focusing on topical coverage, action points, and decisions. We also asked our annotators to generate *bullet-point* minutes instead of a coherent textual summary.

   From the formal point of view, meeting minutes in our dataset mostly have some metadata, such as the name, date, and purpose of the meeting, the list of attendees, and the minuting author's name. The metadata are however directly included at the top of the text of the minute and their form is not fully standardized.

   The first versions minutes were mainly generated by the same annotator who corrected the transcript for the given meeting.

   Due to our free-form instructions, the human-generated minutes vary in length and type. Shorter minutes contain just a few action items (less than half a page). Longer minutes may be up to two (occasionally even more) pages.

   Examples of the minutes from our dataset are provided in Appendix A.
6. **De-identification** To avoid any personally-identifiable information or sensitive project-specific information getting leaked out, we took care to de-identify the entire

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

[4] A consequence of this difference in releases is that AutoMin 2021 results can be *exactly* replicated only by the task participants but the same experiments will be possible on the slightly different fully public version of ELITR Minuting Corpus as soon as it is released.

	Lines	Words
Transcript (EN)	731.1 ± 399.9	7101.4 ± 3851.1
Minutes (EN)	36.5 ± 32.9	371.2 ± 457.1
Transcript (CS)	1213.6 ± 476.6	8620.2 ± 3091.7
Minutes (CS)	25.2 ± 12.1	238.7 ± 160.5

Table 1: *Summary of AutoMin Shared Task Dataset. The figures correspond to mean±standard deviation.*

	Train	Dev	Test-I	Test-II
Task A				
EN	85	10	18	10
CS	33	10	10	6
Task B				
EN	565	280	972	-
CS	720	320	300	-
Task C				
EN	555	378	1431	-
CS	782	496	435	-

Table 2: *Number of instances in AutoMin Shared Task data across the tasks and dataset splits.*

dataset. We replaced person, project, and organization mentions with identifier strings: [PERSON*number*], [ORGANIZATION*number*] and [PROJECT*number*], respectively. We kept the identifiers stable throughout our dataset, so whenever the annotators were able to establish the identity of a given person, the same identifier was used. We note that in practice, this was complicated by unclear speech, unknown spelling of various foreign names, and lack of knowledge of people's voices.

Before releasing the corpus, we shuffled these identifiers within each meeting. In other words, the transcript and all its minutes share the same codes, but different meetings use different randomization in the released version.

7. After de-identifying the transcript and the minutes, we ran again a round of consent collection from meeting participants. All our participants were invited to review the de-identified transcripts and minutes, to validate for themselves that the de-identification is sufficient; some further concerns were raised and led to another round of identification of small problematic elements in the texts. The participants provided their explicit consent to use the data for a public release, after the last problematic parts will be removed.

Table 1 shows a summary statistic of our shared task dataset. We report the average number of lines and words in the dataset transcript and minutes.

4.1. Training, Development and Test Sets for AutoMin

Table 2 shows our *train-dev-test* splits for the AutoMin tasks. We had two separate releases of the test data for Task A, called Test-I and later Test-II.

For Task A, one instance corresponds to one meeting transcript and all its reference minutes. Please note that the test set minutes were not provided to the participants.

For Task B, one instance is a pair of a meeting transcript and a minute. For Task C, it is a pair of minutes. For Task B and C, we kept meeting instances separate for train, test, and dev sets. In other words, Task B as well as C test instances are created only from meetings that appeared in Task A test set. We randomly paired the minute-transcript (Task B) and minute-minute (Task C) to generate the task-specific instances. When selecting the random pairs, we did not consider the entire dataset to generate the instances. We used our knowledge of meetings source and selected only some of these sources. We particularly preferred sources where the meetings were recurring, so that Task B and C are more challenging. Another advantage of this subsetting is that we can use other portions of the dataset in next iterations of AutoMin.

5. Shared Task Timeline and Procedure Overview

AutoMin followed this timeline:

- Trial Data Available: March 22, 2021
- Training Data Available: May 15, 2021
- Test Data Release (1st set): June 15, 2021
- Test Data Release (2nd set): July 1st week, 2021
- System Output Submission Deadline: July 15, 2021
- Result Announcement and Notification: August 16, 2021
- System Report Due: August 23, 2021
- Review Notification: September 1, 2021
- Event Date: September 4, 2021

Since our data were still only confidential when we were running the task, we first invited expression of interest from the Speech and Language Processing Community to take part in this challenge via several forums[5] and provided "trial data" to illustrate the tasks. The trial data can be accessed here:

`https://github.com/ELITR/automin-2021`

The participating teams were required to sign an agreement of data confidentiality with us. Once we had the agreements signed, we invited the participants to access our private Github repository to access the shared task data and the participating instructions.

Since the number of instances in AutoMin were generally insufficient for the training of end-to-end deep learning models, we encouraged task participants to use related data from dialogue summarization datasets, other meeting summarization corpora, or general summarization datasets to (pre-)train their models.

6. Participating Teams and Approaches

Of the 27 teams who registered for AutoMin, 10 teams eventually took part in the shared task. We had participating teams from academia as well as industry from all around the world including Japan, India, Germany, Switzerland, Russia, and UK.

We briefly discuss the approaches of our participating teams (ordered alphabetically):

[5] Corpora list, ISCA Web, SIGIR list, SIGDial list, Twitter, LinkedIn and others.

- **Team ABC [32]** participated in all tasks (A, B, and C) for EN data. They employed a BART-based [33] minuting architecture trained on the SAMSum corpus [14] with certain pre-processing (e.g., a simple rule-based transcript segmentation) to generate the bullet-point minutes. For Task B and Task C, the authors employed a feature-based approach with Support Vector Machines and Random Forest classifiers. They used the same set of features for both tasks: cosine similarity between the vectors, ROUGE-1, ROUGE-L, Jaccard similarity, Sequence Matcher, named entity match, ratio of the most common words to the total number of unique words, etc.
- **Team Auto Minuters [34]** participated in all tasks again only for EN meetings. The authors use a pre-trained T5-base model for summarization and fine-tuned the model on the shared task dataset for the minuting Task A. For Tasks B and C, they use several similarity scores (Jaccard and cosine similarity in particular) which they use as input to a K-Nearest Neighbour (KNN) classifier.
- **Team Hitachi [35]** participated in all the tasks for both EN and CS data. They did not use the provided reference minutes for training of Task A. Instead, they topically segmented the transcript and used a BART-based summarizer trained on SAMSum dialogue summarization dataset. In addition, they applied argumentation mining techniques on the generated minutes to improve their coherence and internal structured. The authors resolve Task A in Czech cross-lingually: they use mBART [36] to translate the Czech transcripts to English, process English and then translate the generated minutes back from English to Czech.

  For Tasks B and C, team Hitachi used multiple relevance scores and trained several machine learning models such as SVM, Logistic Regression, Random Forests, and Multi-Layer Perceptron for subsequent classification. They used Optuna[6] for hyperparameter optimization to select the best model for Task B and C.
- **Team JU_PAD [37]** participated in Task A (EN). They stacked pre-trained models for extractive (TextRank [38]) and abstractive summarization (BART trained on CNN/Daily Mail [39]) to generate the minutes. They followed these steps to generate their minutes: Pre-processing (speaker identification, speaker-dialogue separation), Part-of-Speech tagging (each word is attached with its POS), Sentence/Dialogue Processing (dialog act tagging), Extractive Summarization, Abstractive Summarization, and Minute generation (generating the bullet-point minutes from flat paragraph-like summaries). They used a pre-trained a CRF-based model on Switchboard Dialog Act Corpus [40] for dialogue act tagging.
- **Team Matus_Francesco [41]** participated in Task A (EN). They base their minuting system on the PEGASUS [42] summarization model (Pre-training with Extracted Gap-sentences for Abstractive Summarization). They perform certain pre-processing steps including removal of filler words and small talk, co-reference resolution (replacing pronouns like *you*, *I* with the corresponding named-entities) and dialogue partitioning (segment the longer transcripts into shorter chunks) prior to the summarization model.

  The in-time submission was M/F (baseline). The authors also made two late submissions M/F (coref) and M/F (final) where they further fine-tune their model on the AutoMin dataset followed by decoder optimization and add a certain post-processing (removal of non-important and irrelevant information via TF-IDF scoring with a user-tunable threshold).
- **Team MTS [43]** made four different submissions with different approaches for Task A (EN). They used a pipeline system for speech recognition (on AMI and ICSI corpus where the audio is available) and summarization. Their four submissions made use of PreSumm (MTS (P/S)) [44], Google Text-to-Text Transfer Transformers [45] (MTS (T5)), Pegasus [42] (MTS (Pegasus)) and a customized clustering and vectorization approach (MTS (customized)), resp., to generate the minutes. The authors use off-the-shelf pre-trained transformer-based models for the summarization part. The customized approach included steps like syntactic phrase extraction, deletion of redundant words, a vectorization step in combination with TF-IDF scores and Universal Sentence Encoder [46] followed by the final clustering (Affinity Propagation clustering [47]) step.
- **Team Symantlytical [48]** participated in all tasks for EN meetings. For Task A, they made use of Generative Pre-trained Transformer (GPT-2) [49] model to generate the meeting minutes. For Tasks B and C, they used sentence vector representations: BERT [50] trained on SNLI [51] and Paraphrase RoBERTa [52]) with cosine similarity. Finally they used a thresholding scheme on the similarity values to determine the classes for the two tasks (the final threshold value for both the tasks was 0.65).
- **Team Turing TESTament [53]** participated in all tasks for EN data. For Task A, the team employed a feature-based approach (sentence length, unigram frequency, presence of numerical entities, topics from LDA, proper nouns, number of affirmative utterances) with the ranker method TOPSIS[7] to extract the most significant statements from the transcripts with a rule-based heuristic, and finally simply concatenating them as minute items in the end.

  For Tasks B and C, they used sentence representations from BERT trained on the SNLI dataset [51] with cosine similarity to find the similarity of the transcripts and the minutes. Finally, they used a similarity threshold (0.75) for the classification.
- **Team UEDIN [54]** participated in Task A (EN). They developed a minuting system that combines BERT-based extractive summarization with logistic regression-based filtering and certain rule-based pre- and post-processing steps. They leveraged *lecture summarizer*[8] which was originally designed to summarize transcripts of university lectures.
- **Team Zoom [55]** participated in AutoMin, making a late submission to Task A (EN). They used the MediaSum [13] corpus to train their transformer-based summarization model SEAL [56], and fine-tuned on AutoMin, AMI, and ICSI datasets.

[6] https://www.preferred.jp/en/projects/optuna/

[7] https://en.wikipedia.org/wiki/TOPSIS

[8] https://github.com/dmmiller612/lecture-summarizer

	Lines	Words
Transcripts	712.4±322.8	6765.5±2498.7
Ref. Minutes	34.9±17.5	334.3±189.3
ABC	33.9±7.1	433±113.2
Auto Minuters	58.2±29.3	740.7±310.9
Hitachi	99.7±40.2	1822.0±776.1
JU_PAD	18.4±3.8	721.9±125.5
M/F (baseline)	38.4±10.5	1105.9±319.6
M/F (co-ref)†	34.0±13.3	589.2±289.7
M/F (final)†	17.6±12.5	434.6±310.4
MTS (customized)	13.9±3.9	237.9±76.5
MTS (Pegasus)	1±0	108.6±38.4
MTS (P/S)	1±0	634.3±372.7
MTS (T5)	1±0	61.9±14.9
Symantlytical	31.8±40.1	951.4±370.4
Turing TESTament	147.6±165.6	3239.1±982.2
UEDIN	11.3±4.8	160.4±76.3
Zoom†	1±0	22.3±15.9

Table 3: *Basic properties of manual transcripts, reference minutes and all participating team submissions of test set meetings (EN only). We report the average±standard deviation values for the number of lines and words. † marks late submissions.*

Team	Adequacy	Fluency	G/C
ABC	**3.98±0.73**	**4.27±0.55**	**4.45±0.37**
Auto Minuters	2.32±0.60	2.52±0.50	2.64±0.52
Hitachi	**4.25±0.46**	**3.93±0.57**	**4.34±0.41**
JU_PAD	2.86±0.58	2.95±0.61	2.84±0.51
M/F (baseline)	2.55±0.63	2.27±0.63	2.91±0.49
M/F (co-ref)†	2.68±0.65	2.73±0.49	3.18±0.33
M/F (final)†	2.82±0.96	3.09±0.97	3.50±0.85
MTS (customized)	1.86±0.48	1.91±0.46	2.30±0.57
MTS (Pegasus)	1.25±0.31	1.78±0.65	2.61±0.57
MTS (P/S)	1.48±0.40	1.39±0.39	1.96±0.51
MTS (T5)	1.11±0.21	1.73±0.57	2.57±0.74
Symantlytical	2.46±0.51	2.64±0.49	2.98±0.69
Turing TESTament	2.91±0.72	2.46±0.56	2.93±0.66
UEDIN	2.12±0.69	3.34±0.56	3.86±0.62
Zoom†	1.05±0.22	2.32±1.65	3.52±1.77
Overall	2.37±1.09	2.62±1.01	3.09±0.99

Table 4: *Average human evaluation scores (1: worst, 5: best) for English meetings. G/C means Grammatical Correctness. The top score and all scores that fall within its std. dev. bounds are bolded. † marks late submissions.*

7. Evaluation

In this section, we detail the evaluation campaign we carried out for AutoMin. As mentioned earlier, we performed both automatic and human evaluation, treating human evaluation measures as the primary one for Task A. Tasks B and C were only evaluated automatically via classification measures: *F1-score* and *Accuracy*.

We carried out our human evaluation on the participant minutes: for EN data we had two human evaluators assess each of the submissions, for CS submissions we had one native speaker to evaluate the participant minutes (Hitachi was the only team who submitted their system run for CS meetings).

Kindly note that the human evaluation was *reference-less*. In other words, our evaluators had access to only the transcript of the meeting to evaluate the candidate minutes (participant submissions). We purposely did this to eliminate the bias of our human evaluators towards the reference minutes.

As intended from the beginning, we did not rank our participants, but we have takeaways from the best as well as the relatively poorer system outputs.

Kindly refer to Appendix B to get a glimpse of some participant minutes in Task A (EN).

During the evaluation, we didn't find any significant difference in performance between the two test sets (Test-I and Test-II), so we merge them for the rest of the analysis.[9] Note that some submissions arrived late (marked with † in the tables) and some MTS submissions are treated as additional ones; these are discussed in a separate Section 7.5.

[9]If you want to compare the two test sets, please refer to the detailed tables in Appendix C and Appendix D for English and Tables 5 and 7 for Czech. English Test-I consisted of meetings *en_test_001–018* and Test-II consisted of meetings *en_test_019–028*. Czech Test-I consisted of meetings *cs_test_001–010* and Test-II consisted of meetings *cs_test_011–016*.

7.1. Basic Statistics

We report basic test set statistics in Table 3: the average number of lines, words in each transcript and reference minutes as well as for the participant submission (candidate minutes). This provides a first useful comparison of the participant minutes with respect to the reference minutes and transcripts.

We can see that there is a wide variation in the length of the reference minutes as well as those generated by the different participants. This variance in part comes from the different length/duration of the meetings, which is almost directly proportional to the length of the transcript. The variance of minutes lengths depends on the meeting duration, amount of discussed content but also on the minuting behavior of the human scribe (some make detailed minutes, some prefer doing shorter ones).

Some participant minutes were not in the form of bulleted list like we intended to have. Instead, they produced flat summaries; some even generated one long single-line summary which was difficult to interpret in manual evaluation, see Zoom or some of MTS models. Zoom and T5 by MTS have also produced by far the shortest outputs, suggesting that their Transformer-based models may suffer from the length overfitting issue [57].

7.2. Task A Manual Evaluation Results

For human evaluation, we had multiple evaluators evaluating each candidate minute in the three criteria: Adequacy, Fluency and Grammatical Correctness (G/C).

Table 4 shows the summary of our human evaluation for the test meetings on EN data. Manual scores for individual meetings can be found in Appendix C. For Czech, only the Hitachi team provided minutes and the detailed scores of individual meetings as well as the average are provided in Table 5.

Note that although we kept the identity of the teams hidden to the human assessors and reshuffled the order of the submissions, we realize that it is often not difficult to make an educated guess looking at the pattern of the candidate minutes and identify minutes produced by one system. We acknowledge that this unintended human bias may have affected the evaluation.

Test Meeting	Adequacy	Fluency	G/C
cs_test_001	4	2	1
cs_test_002	3	2	1
cs_test_003	3	2	1
cs_test_004	3	2	1
cs_test_005	3	2	1
cs_test_006	4	3	2
cs_test_007	4	3	2
cs_test_008	3	1	2
cs_test_009	2	2	1
cs_test_010	2	2	1
cs_test_011	2	2	1
cs_test_012	2	2	1
cs_test_013	2	1	1
cs_test_014	2	2	1
cs_test_015	2	3	2
cs_test_016	2	2	1
Average	2.69±0.79	2.06±0.57	1.25±0.45

Table 5: *Adequacy, Fluency and Grammatical Correctness (G/C) of Team Hitachi. Only Team Hitachi Participated in Task A for Czech meetings. In this case, only one Czech evaluator (native speaker) did the human evaluation.*

We see that for English, adequacy overall received the lowest scores, fluency was deemed better and grammatical correctness was the highest. Arguably, annotators were free to use the 1–5 range on the Likert scale as they liked but we still assume that each of them used it comparably across the three scales.

This general tendency is apparent in many submissions, with MTS (T5 and Pegasus) and Zoom being the most striking examples: their adequacy is close to the lowest possible value of 1 but their grammatical correctness is in the middle range, 2.5–3.

We are of the opinion that for practical usability, adequacy should be the most important criterion. However, promoting adequacy is apparently not easy in system design, only Turing TESTament, Hitachi, M/F and PreSumm model by MTS managed to score higher in adequacy than in fluency.

For Czech meetings, Hitachi as the only participating team received better scores for adequacy than for fluency and grammaticality. This is surely promising, but the result can be affected by the fact that the final Czech was the output of a machine translation system. We see it as more likely that the minutes suffer from a lower fluency and grammaticality rather than assuming that when applied cross-lingually, the underlying system manages to produce better (more adequate) outputs.

In Table 4, we bolded the best score and all scores that fall within its reported standard deviation. Proper significance testing for our purpose has yet to be selected. We see that ABC and Hitachi scored best in all three criteria. As hinted above, Hitachi seems to be somewhat better in adequacy.

A great result is that Hitachi in adequacy, ABC in fluency and both of them in grammatical correctness score close to 5, the highest value of the scale. In this first year of AutoMin, we have however too little experience with manual evaluations to know whether the annotators tend to use the scale as a relative measure (5 meaning the best of all but still mediocre), or if they use it as an indicator of an absolute, acceptable, quality.

Teams	R-1	R-2	R-L
ABC	**0.33±0.08**	**0.08±0.04**	**0.19±0.06**
Auto Minuters	**0.25±0.06**	**0.06±0.03**	**0.14±0.04**
Hitachi	**0.26±0.09**	**0.08±0.03**	**0.14±0.05**
JU_PAD	**0.27±0.07**	**0.06±0.03**	**0.15±0.04**
M/F (baseline)	0.21±0.07	**0.05±0.02**	0.11±0.04
M/F (co-ref)†	**0.25±0.08**	**0.06±0.03**	**0.14±0.05**
M/F (final)†	0.21±0.06	**0.05±0.03**	0.12±0.04
MTS (customized)	0.20±0.04	**0.05±0.02**	0.11±0.03
MTS (Pegasus)	0.08±0.05	0.01±0.01	0.06±0.03
MTS (P/S)	0.16±0.09	0.03±0.03	0.09±0.05
MTS (T5)	0.06±0.04	0.01±0.01	0.05±0.03
Symantlytical	**0.26±0.07**	**0.06±0.03**	**0.13±0.04**
Turing TESTament	0.20±0.08	**0.06±0.04**	0.12±0.06
UEDIN	0.21±0.04	**0.05±0.03**	**0.14±0.03**
Zoom†	0.05±0.03	0.00±0.01	0.03±0.02

Table 6: *Average of the maximum automatic evaluation scores for each team against test-set reference minutes (EN only). The top score and all scores that fall within its std. dev. bounds are in **bold**. † marks late submissions.*

7.3. Task A Automatic Evaluation Results

For automatic evaluation, we took the usual text summarization metric ROUGE [29] in its three variants (ROUGE-1, ROUGE-2, and ROUGE-L). As described, each meeting has multiple reference minutes to allow for at least partial reflection of the fact that minuting styles across people differ.

For each candidate minute, we calculate ROUGE (F1) scores across all available references and report the average and also the maximum. When taking the maximum, we essentially allow each team to "use their particular style" of the minute and score it with the reference "closest to this style".

Tables 13 to 15 in Appendix D show the ROUGE-1, ROUGE-2, and ROUGE-L evaluations for individual English meetings, respectively. In Table 6 here, we again summarize them by reporting the average of maximum ROUGE scores obtained by each participant against the different reference minutes.

Best scores are in bold, again with all other scores that fall within the std. dev. band of the best one. Compared to manual evaluation, more systems reach this top band. ABC scores best in ROUGE-1 and ROUGE-L but this advantage is not visible in ROUGE-2. As stated earlier, ROUGE-1 and ROUGE-2 measures are motivated with uni-gram and bi-gram overlap respectively, whereas ROUGE-L is inspired with the overlap in the longest common sub-sequence between the candidate and reference summaries. Systems with higher ROUGE scores signify that the n-grams/strings in their candidate summaries match with the reference summaries to a higher extent than others.

It is also worth mentioning that systems which produced a one-line summary all except MTS (PreSumm) receive ROUGE-2 of flat zero and generally the lowest R-1 and R-L scores. It is quite obvious that the systems with lower number of n-grams/strings in their summaries would be penalized when evaluated via ROUGE.

Interestingly, we see that it is not the case that the teams which produced longer summaries (i.e., have more information) necessarily have higher ROUGE scores, despite the fact that longer outputs could lead to more matches. ROUGE is a lexical measure which relies on word overlap with the reference. Both extractive and abstractive methods for creating minutes

	ROUGE-1		ROUGE-2		ROUGE-L	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max
cs_test_001	0.20	0.26	0.04	0.05	0.08	0.10
cs_test_002	0.13	0.20	0.02	0.05	0.06	0.10
cs_test_003	0.16	0.18	0.03	0.04	0.09	0.10
cs_test_004	0.15	0.22	0.02	0.03	0.06	0.08
cs_test_005	0.06	0.09	0.01	0.01	0.03	0.05
cs_test_006	0.12	0.21	0.02	0.04	0.06	0.10
cs_test_007	0.11	0.21	0.02	0.03	0.05	0.08
cs_test_008	0.15	0.23	0.02	0.04	0.07	0.09
cs_test_009	0.15	0.20	0.02	0.04	0.06	0.08
cs_test_010	0.16	0.20	0.02	0.03	0.08	0.10
cs_test_011*	0.29	0.29	0.03	0.03	0.11	0.11
cs_test_012*	0.15	0.15	0.03	0.03	0.07	0.07
cs_test_013*	0.05	0.05	0.01	0.01	0.03	0.03
cs_test_014*	0.24	0.24	0.04	0.04	0.08	0.08
cs_test_015*	0.19	0.19	0.03	0.03	0.09	0.09
cs_test_016*	0.24	0.24	0.04	0.04	0.10	0.10

Table 7: *ROUGE-1,2, and L scores of Team Hitachi against the CS test set reference minutes. Only Team Hitachi Participated in Task A for Czech meetings. Meetings marked with $*$ have only one reference minute.*

can suffer from mismatch in case the reference uses different words than the transcripts. Without some more explicit form of alignment between the candidate and the reference, and some technique of handling paraphrases, ROUGE is not likely to well reflect the true quality of the minute.

7.4. Correlation between Automatic and Human Evaluation

As a first type of meta-evaluation, we check how our two evaluation strategies correlate. We plot the Pearson correlation between the automatic and the human evaluation scores for each of the teams. Kindly refer to Figure 1 for correlation between average scores and Figure 2 for correlation between the average manual score across the two assessments and the maximum automatic scores across the several references.

Across the teams, the correlation heatmaps indicate a high correlation among the different versions of ROUGE but generally a low correlation between manual scores and ROUGE scores. Higher correlations between manual and automatic scores were found only for MTS and Zoom, which suffer the unsegmented output problem and do not score well in adequacy.

For some teams, we also see a high correlation between adequacy and fluency. While these two scores are known to be often correlated when evaluating the quality of machine translation, we are surprised to see such an effect here. We were hopeful that adequacies would reflect the level to which the minute is an adequate summary of the meeting – which in turn would hopefully boil down to some form of coverage. Obviously, the manual evaluation method deserves some refinement. It is possible that *some* evaluators failed to separate adequacy and fluency, despite our instructions to do so, but it still does not explain why it would affect only some systems because we were allocating annotation tasks to evaluators by meetings: all candidate minutes for a given meeting were assessed by one evaluator, so that they would have the complete picture.

Interesting negative correlations are observed in some situations: ABC, Hitachi, and to a lesser extent UEDIN and JU_PAD show slight negative correlations between fluency and ROUGE scores. ROUGE is not primarily geared towards fluency, so it needs to align with it. Unfortunately, the correlation of ROUGE with adequacy for these systems is little or none, either. We have to conclude that ROUGE is a problematic automatic measure for this task.

The observations are similar for the manual maxima (Figure 2), except that the high correlations between manual fluency and adequacy have generally disappeared, which is a good sign.

7.5. Additional Submissions to EN Task A

Team Matus_Francesco made two further late submissions (marked with †) and team MTS submitted another three runs (additional runs) to Task A. We included their summary results already in Tables 4 and 6, for an easy comparison with others. The detailed results (automatic and human evaluation) are in separate Tables 16 to 21 in Appendix E.

We can clearly see that Teams MTS and Team Matus_Francesco significantly improved their performance both in terms of automatic and human evaluation. Tables 4 and 6 summarize the improvement across the test set.

For team MTS, the best performing submission turns out to be the customized clustering-based approach both in terms of automatic and human scores. We can also see that the customized approach from MTS yields more lines in the minutes (Table 3). Although other submissions from MTS generated summaries of comparable length (in terms of number of words), they were not split into sentences, and hence suffered in readability.

For Matus_Francesco, it proved helpful in the late submission to fine-tune the Pegasus-large model and run a decoder optimization step, preventing the decoder from generating personal pronouns and repeating n-grams. Also, their initial baseline submission consisted of Pegasus-base model whereas in their late submissions they used Pegasus-large which probably contributed to their enhanced performance. An interesting difference is in automatic and manual evaluation: automatic scores prefer the Pegasus-large model with co-reference resolution (M/F (co-ref)) whereas human evaluation prefers the decoder optimization variant (M/F (final)).

7.6. Task B Evaluation Results

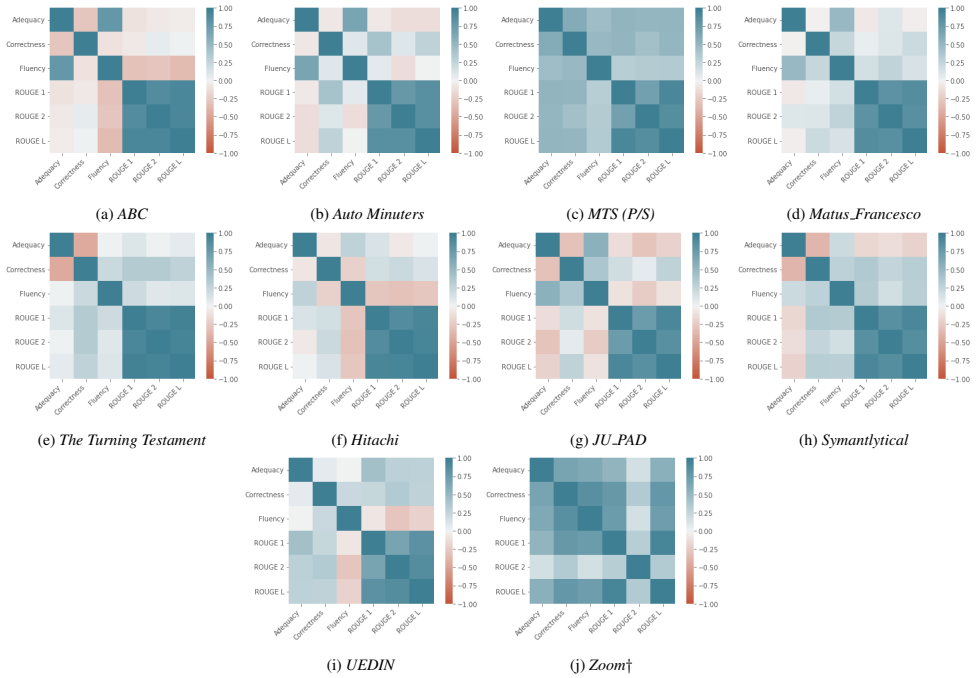
For Task B, five teams participated with their methods for the EN meetings while only Team Hitachi participated for the CS meetings.

Since, Task B is essentially a classification problem, we use Accuracy and F1 scores to evaluate the submissions.

We are more interested in finding out if the submissions can detect the minute-transcript pairs that belong to the same meeting. Hence we report the participant performance (F_1 score) for the YES class, indicating how often the system does not miss a YES pair (the underlying recall) as well as does not suggest many false positives (the underlying precision).

Please note that the proportion of the NO class instances is higher than that of the YES class in the train, test, and dev sets making it further difficult to predict the YES class. The train set had 15.4% and the dev set had 10% of YES-class instances only. Participants were encouraged to make use of external datasets to mitigate the class imbalance.

Task B results in Table 8 shows that out of the five participating teams, ABC, Auto Minuters, and Hitachi fared well in terms of accuracy. However, since there is a strong class imbalance (only 5.5% of test set instances have the answer YES),



(a) *ABC* (b) *Auto Minuters* (c) *MTS (P/S)* (d) *Matus_Francesco*

(e) *The Turning Testament* (f) *Hitachi* (g) *JU_PAD* (h) *Symantlytical*

(i) *UEDIN* (j) *Zoom†*

Figure 1: *Correlation between manual and automatic evaluation scores of the participating teams (taking the average scores, EN meetings only). "Correctness" here denotes grammatical correctness. † marks a late submission.*

Team	Accuracy	F1
ABC (EN)	87.6%	0.08
Auto Minuters (EN)	94.8%	0.37
Hitachi (EN)	97.7%	0.82
Symantlytical (EN)	42.6%	0.11
Turing TESTament (EN)	41.1%	0.10
Hitachi (CS)	95.7%	0.75

Table 8: *Task B Evaluation, F1-scores are for the YES class. Only Team Hitachi participated in the CS portion of the dataset.*

accuracy fails to depict the merit of the systems in identifying the YES-class instances.

Considering F1 instead of Accuracy, it is only the Hitachi team that maintains a good performance, in both English and Czech.

Based on system description papers, we see that Team Hitachi used multiple similarity and relevance features (tf-idf, cosine similarity, named entity overlap ratio, date consistency, and BERTScore [58]) with adequate hyperparameter optimization. Team ABC, too, used several features (as mentioned in Section 6) but apparently, they missed to properly weigh the contribution of their features in the task. Other teams like Symantlytical and The Turing TESTament used manual threshold-based schemes on their features which may have resulted in their poorer performance. Team AutoMinuters used simple features like cosine and Jaccard similarity between the transcript and minute pairs and fed those to a kNN classifier. They performed second to Hitachi both in terms of Accuracy and F_1 score.

7.7. Task C Evaluation Results

We evaluate Task C similarly to Task B. Five teams participated for the EN meetings and one for the CS meetings. Please note that Task B and C were optional for our participants.

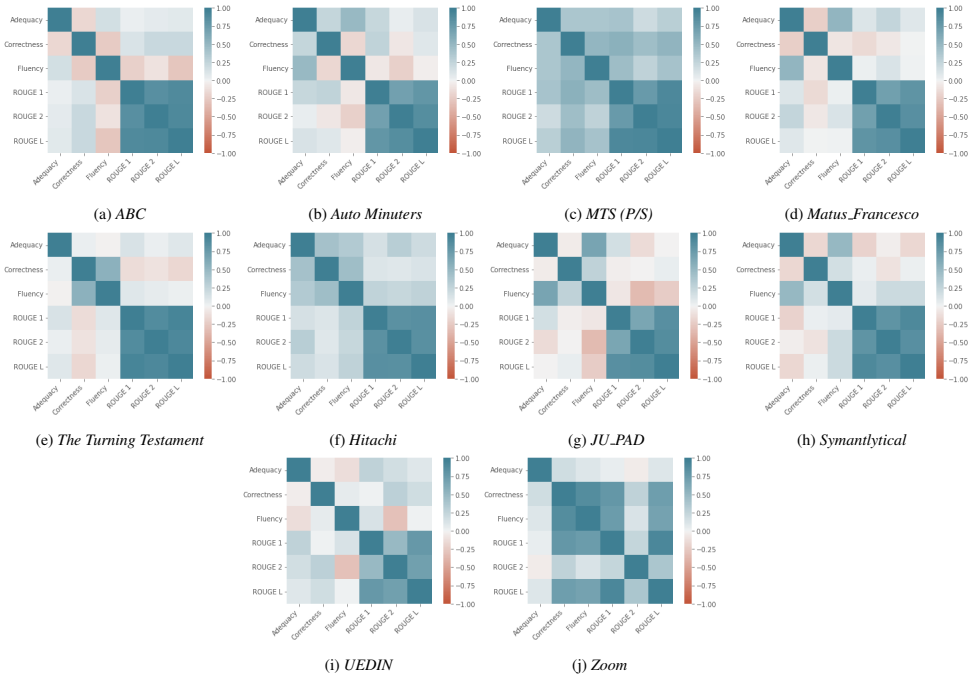
Table 9 shows the performance of the participating teams in Task C.

As in Task B, NO instances are prevalent in Task C. Only 12.8% of the train instances, 11.9% of the dev set and 6.4% of the test set are the YES classes.

In Task C, almost all systems have a good Accuracy (above 80% or even 90%) but again, it is only the Hitachi team that performs well (.66 or .90) in F1, too. Almost all the teams used the same set of features/approaches which they used in Task B which is not surprising given the similarity of the tasks.

Team Symantlytical and The Turing TESTament used pre-trained deep model representations with cosine similarity and thresholds for the classification, however still they did not succeed to produce good results.

Team ABC shows the biggest discrepancy here: with Accuracy of 84.3%, its F1 score is only 0.03. A detailed look reveals that ABC suffers from both a low recall (0.03) as well as a low



(a) *ABC* (b) *Auto Minuters* (c) *MTS (P/S)* (d) *Matus_Francesco*

(e) *The Turning Testament* (f) *Hitachi* (g) *JU_PAD* (h) *Symantlytical*

(i) *UEDIN* (j) *Zoom*

Figure 2: *Correlation between Automatic and Human Evaluation Scores of the participating teams (taking the maximum scores, EN meetings only). † marks a late submission.*

Team	Accuracy	F1
ABC (EN)	84.3%	0.03
Auto Minuters (EN)	92.3%	0.39
Hitachi (EN)	93.8%	0.66
Symantlytical (EN)	80.0%	0.28
Turing TESTament (EN)	52.3%	0.14
Hitachi (CS)	98.4%	0.90

Table 9: *Task C Evaluations, F1-scores are for the YES class. Only Team Hitachi participated in the CS portion of the dataset.*

precision (0.02) on the YES class. Hence, ABC's system misses to classify most of the YES class instances and is biased towards predicting all instances as NO. They still get away with higher accuracy since majority of the test set instances are NO here.

8. Findings of AutoMin

Organizing AutoMin was a fulfilling experience for us starting from the novelty and uniqueness of the task, developing the dataset, coming up with the baselines, promoting the event, building the community ensuring their participation, working closely with the participants, evaluating and analyzing the submissions and eventually writing this paper. We summarize our findings and recommendations for the main task:

1. The best-performing systems in AutoMin suggest: BART-based deep neural models perform comparatively better than other transformer models to generate readable minutes.
2. As is evident from the several submissions (ABC, Hitachi, JU_PAD, Matus_Francesco, etc.), segmentation of the long meeting transcripts (either topically or via simple segmentation schemes) is crucial to the performance of the subsequent summarization modules in the proposed systems. Existing summarization models apparently have certain limitations in the number of tokens they can process to produce a good output. It would be interesting to know where the limitation effectively comes from: the inability to capture all the necessary information from a long input, the inability to produce longer output [57], or both.
3. Considering the current non-availability of large-scale domain datasets on multi-party meeting summarization (even AutoMin dataset is small-scale), the best recipe that evolved out for Task A looks like: train a deep neural model on available dialogue summarization datasets (SAMSum [14], DialSum [19], etc.) and further fine-tune it on the minuting or meeting summarization datasets (AMI [8], ICSI [9], AutoMin).
4. Simply using off-the-shelf text summarization models trained on text summarization datasets from other do-

mains (newswire, speech, etc.; see submissions by MTS or Zoom) does not seem to work satisfactorily, emphasizing the need for pre-processing and post-processing on this task. Also as discussed in the previous point, dialogue-summarization-specific training of the deep neural models proved to be helpful for summarizing the multi-party speech. Meetings usually have a specific theme/agenda and involve multiple parties which may not be the case for dialogues. However, structurally, meetings and dialogues are closer than meetings and regular texts. Hence in the absence of large-scale meeting summarization/automatic minuting datasets, dialogue summarization datasets are probably the best alternative we have to train the deep neural models.

5. Resource scarcity is a major hindrance for research on this particular topic. There is a need to develop large-scale datasets to enable end-to-end training and leverage the power of large language models for this problem. Our experience says that the major reason behind the non-availability of meeting datasets are the privacy and ethical concerns in professional meetings. People are not comfortable in sharing their meeting discussions which may contain sensitive and personal information in free-flow conversations.

   It took us a lot of time and effort to de-identify the named-entities in the meeting conversations and also to further "censor" the transcripts, removing information which is no longer protected by GDPR but which is still potentially sensitive, as suggested by the meeting participants. Prior to this additional "censorship", we released the data to the shared task teams only after they signed a non-disclosure agreement with us.

   We procured consent for publication from meeting participants after showing them the de-identified version of the meeting transcripts and we find this two-stage consents (1. consent to record and process internally, 2. consent to publish the processed, de-identified data) an ideal strategy. It is much easier for the participants to realize what is being released from a full preview compared to some generic description.

6. Although the cross-lingual submission by Hitachi worked reasonably for Czech according to our evaluation, a further verification on other languages is needed. We thus see a need for efforts to develop multilingual datasets because many meetings are conducted in languages other than English.

7. The AMI [8] and ICSI [9] were the only dedicated datasets on meeting summarization until AutoMin. Organizations (academia/industry) need to come forward to donate their meetings and minutes (overcoming the ethical and privacy limitations) to create a large-scale dataset. We put up a similar call in our ELITR project blog.[10]

8. There exists a large variety of meetings with different scope and goals, and a large variety in minuting styles. A "one size fits all" approach to generate a meeting minute probably would not work here. In addition to data collection across meeting types as advocated above, meeting notes (minutes) taken by different people from different perspectives and expectations are required to train a model to avoid biases towards certain styles of minuting.

9. As discussed in the previous point, minuting is a subjective activity. Different note-takers/participants would have different perspectives/expectations on what are the best possible minutes. Hence, effort towards personalized minutes generation is a worthy research direction. Generating a query-focused summary from meetings [59] is a nice example of this kind.

10. The community acknowledges a dire need for better evaluation metrics for text summarization including meeting summarization [4]. As we documented for ROUGE in our correlation results, the current automatic metrics (ROUGE, BERTScore [58], etc.) are not a good estimator of the quality of the summaries. We see here a large room for improvement from further research.

11. Human evaluation of the generated minutes using simple Likert scales was possible, but further improvements of the procedure should be sought for, and a larger-scale evaluation of inter-annotator and intra-annotator agreement is desirable. While human evaluation is likely to remain inevitable when comparing the quality of the generated output from different models, some evaluation support tools could speed up the process and increase agreement at the same time. We anticipate that a semi-automatic *human-in-the-loop* evaluation scheme would be the best fit for this problem.

12. To maintain the acquired motivation of the community in joint and focused investigations on automatic minuting, it calls for further shared tasks/challenges like AutoMin, DialogSum [26], etc.

[10] https://elitr.eu/recipe-for-miracles-to-happen/

9. Conclusions and Future Plans

We reported on AutoMin 2021, the first shared task on automatic construction of meeting summaries, "minutes". We received submissions from 10 teams and observed an interesting variance in approaches as well as final output quality.

Our observations confirm that automatic evaluation for minuting is unreliable, that the training data are small and that off-the-shelf models like Transformer do not lead to good results. At the same time, very promising outputs were obtained from BART-based models that followed some meeting segmentation strategy. One open concern here is the adequacy of the summaries, which we evaluated only with a simple score, not via a careful scrutiny matching summary points and utterances from the transcript.

The final writeup of AutoMin overview took us longer than desired, but finally, we have this concise picture. We are already starting preparatory steps for the next iteration of AutoMin, hoping to attract a similar or larger attention of the NLP and speech community.

10. Acknowledgement

First of all, we would like to thank the participants for their enthusiastic participation in AutoMin and also bearing with our delays. We would like to thank the registrants who unfortunately could not participate but still became part of the discussion. We extend our gratitude to the reviewers from ISCA for their positive feedback on AutoMin, the satellite chairs of Interspeech 2021 to support AutoMin. We thank our program committee for reviewing and helping us improve the system description papers.

We organized AutoMin 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. Ondřej Bojar would also like to acknowledge the support from the grant 19-26934X (NEUREM3) of the Czech Science Foundation.

Finally, we thank our data partners for contributing their meeting data. We acknowledge the involvement of our ELITR colleagues for their inputs on designing the shared task. We thank Peter Polák, Kartik Shinde, Ahmad Haji Mohammadkhani, James Levell for their inputs and help in various phases of the baseline experiments and data preparation. Lastly and more importantly we cannot thank enough our annotators and scribes for helping us develop our AutoMin shared task dataset. We are highly looking forward to the next iteration of AutoMin.

11. References

[1] Z. Li, J. Tang, X. Wang, J. Liu, and H. Lu, "Multimedia news summarization in search," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 3, pp. 1–20, 2016.

[2] A. Cohan and N. Goharian, "Scientific document summarization via citation contextualization and scientific discourse," *International Journal on Digital Libraries*, vol. 19, no. 2, pp. 287–303, 2018.

[3] X. Feng, X. Feng, and B. Qin, "A survey on dialogue summarization: Recent advances and new frontiers," *arXiv preprint arXiv:2107.03175*, 2021.

[4] T. Ghosal, M. Singh, A. Nedoluzhko, and O. Bojar, "Report on the sigdial 2021 special session on summarization of dialogues and multi-party meetings (summdial)," in *ACM SIGIR Forum*, vol. 55, no. 2. ACM New York, NY, USA, 2021, pp. 1–17.

[5] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 3465–3469. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-1873

[6] R. Hsiao, D. Can, T. Ng, R. Travadi, and A. Ghoshal, "Online automatic speech recognition with listen, attend and spell model," *IEEE Signal Processing Letters*, vol. 27, pp. 1889–1893, 2020.

[7] Y. Zhang, A. Ni, T. Yu, R. Zhang, C. Zhu, B. Deb, A. Celikyilmaz, A. H. Awadallah, and D. R. Radev, "An exploratory study on long dialogue summarization: What works and what's next," in *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 4426–4433. [Online]. Available: https://doi.org/10.18653/v1/2021.findings-emnlp.377

[8] I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, "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*, 2005.

[9] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The icsi meeting corpus," 2003, pp. 364–367.

[10] M. Michel, J. Ajot, and J. G. Fiscus, "The NIST Meeting Room Corpus 2 Phase 1," in *Machine Learning for Multimodal Interaction, Third International Workshop, MLMI 2006, Bethesda, MD, USA, May 1-4, 2006, Revised Selected Papers*, 2006, pp. 13–23. [Online]. Available: https://doi.org/10.1007/11965152_2

[11] S. Burger, V. MacLaren, and H. Yu, "The isl meeting corpus: The impact of meeting type on speech style," 01 2002.

[12] K. Zechner, "Automatic summarization of spoken dialogues in unrestricted domains," 2001. [Online]. Available: https://isl.anthropomatik.kit.edu/downloads/Zechner_Klaus_thesis.pdf

[13] C. Zhu, Y. Liu, J. Mei, and M. 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*. Online: Association for Computational Linguistics, Jun. 2021, pp. 5927–5934. [Online]. Available: https://aclanthology.org/2021.naacl-main.474

[14] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization," in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 70–79. [Online]. Available: https://aclanthology.org/D19-5409

[15] R. Rameshkumar and P. Bailey, "Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 5121–5134. [Online]. Available: https://aclanthology.org/2020.acl-main.459

[16] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 5016–5026. [Online]. Available: https://aclanthology.org/D18-1547

[17] A. Clifton, S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. Bonab, M. Eskevich, G. Jones, J. Karlgren, B. Carterette, and R. Jones, "100,000 podcasts: A spoken English document corpus," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5903–5917. [Online]. Available: https://aclanthology.org/2020.coling-main.519

[18] K. Krishna, S. Khosla, J. Bigham, and Z. C. Lipton, "Generating SOAP notes from doctor-patient conversations using modular summarization techniques," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 4958–4972. [Online]. Available: https://aclanthology.org/2021.acl-long.384

[19] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, "DialogSum: A real-life scenario dialogue summarization dataset," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5062–5074. [Online]. Available: https://aclanthology.org/2021.findings-acl.449

[20] P. Tardy, D. Janiszek, Y. Estève, and V. Nguyen, "Align then summarize: Automatic alignment methods for summarization corpus creation," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 6718–6724. [Online]. Available: https://aclanthology.org/2020.lrec-1.829

[21] H. T. Dang, "Overview of duc 2005," in *Proceedings of the document understanding conference*, vol. 2005, 2005, pp. 1–12.

[22] M. K. Chandrasekaran, G. Feigenblat, E. Hovy, A. Ravichander, M. Shmueli-Scheuer, and A. de Waard, "Overview and insights from the shared tasks at scholarly document processing 2020: CL-SciSumm, LaySumm and LongSumm," in *Proceedings of the First Workshop on Scholarly Document Processing*. Online: Association for Computational Linguistics, Nov. 2020, pp. 214–224. [Online]. Available: https://aclanthology.org/2020.sdp-1.24

[23] M. K. Chandrasekaran, G. Feigenblat, D. Freitag, T. Ghosal, E. Hovy, P. Mayr, M. Shmueli-Scheuer, and A. de Waard,

"Overview of the first workshop on scholarly document processing (SDP)," in *Proceedings of the First Workshop on Scholarly Document Processing*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1–6. [Online]. Available: https://aclanthology.org/2020.sdp-1.1

[24] I. Beltagy, A. Cohan, G. Feigenblat, D. Freitag, T. Ghosal, K. Hall, D. Herrmannova, P. Knoth, K. Lo, P. Mayr, R. M. Patton, M. Shmueli-Scheuer, A. de Waard, K. Wang, and L. L. Wang, Eds., *Proceedings of the Second Workshop on Scholarly Document Processing*. Online: Association for Computational Linguistics, Jun. 2021. [Online]. Available: https://aclanthology.org/2021.sdp-1.0

[25] I. Beltagy, A. Cohan, G. Feigenblat, D. Freitag, T. Ghosal, K. Hall, D. Herrmannova, P. Knoth, K. Lo, P. Mayr *et al.*, "Overview of the second workshop on scholarly document processing," Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), Tech. Rep., 2021.

[26] Y. Chen, Y. Liu, and Y. Zhang, "DialogSum challenge: Summarizing real-life scenario dialogues," in *Proceedings of the 14th International Conference on Natural Language Generation*. Aberdeen, Scotland, UK: Association for Computational Linguistics, Aug. 2021, pp. 308–313. [Online]. Available: https://aclanthology.org/2021.inlg-1.33

[27] M. El-Haj, A. AbuRa'ed, M. Litvak, N. Pittaras, and G. Giannakopoulos, "The financial narrative summarisation shared task (FNS 2020)," in *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*. Barcelona, Spain (Online): COLING, Dec. 2020, pp. 1–12. [Online]. Available: https://aclanthology.org/2020.fnp-1.1

[28] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[29] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[30] T.-S. Nguyen, S. Stüker, and A. Waibel, "Super-human performance in online low-latency recognition of conversational speech," in *22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021) : Brno, Czech Republic, 30 August-3 September 2021*, vol. 6. Curran Associates, Inc., 2021, pp. 4131–4135.

[31] J. Kratochvíl, P. Polák, and O. Bojar, "Large corpus of czech parliament plenary hearings," in *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, France: European Language Resources Association, 2020, pp. 6363–6367.

[32] K. Shinde, N. Bhavsar, A. Bhatnagar, and T. Ghosal, "Team abc @ automin 2021: Generating readable minutes with a bart-based automatic minuting approach," in *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021, pp. 1–10. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-2

[33] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. 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*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[34] P. Mahajan and H. Singh, "Team autominuters at automin 2022: Fine-tuning t5 to generate minutes," in *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021, pp. 1–5. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-3

[35] A. Yamaguchi, G. Morio, H. Ozaki, K. ichi Yokote, and K. Nagamatsu, "Team hitachi @ automin 2021: Reference-free automatic minuting pipeline with argument structure construction over topic-based summarization," in *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021, pp. 1–8. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-4

[36] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 11 2020. [Online]. Available: https://doi.org/10.1162/tacl_a_00343

[37] S. Pan, P. Nandi, and D. Das, "Team ju_pad @ automin 2021: Mom generation from multiparty meeting transcript," in *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021, pp. 1–4. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-5

[38] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 404–411. [Online]. Available: https://aclanthology.org/W04-3252

[39] R. Nallapati, B. Zhou, C. N. dos Santos, c. Gülçehre, and B. Xiang, "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*. The Association for Computer Linguistics, 2016, pp. 280–290. [Online]. Available: http://aclweb.org/anthology/K/K16/K16-1028.pdf

[40] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13," University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO, Tech. Rep. 97-02, 1997.

[41] M. Žilinec and F. I. Re, "Team matus and francesco @ automin 2021: Towards neural summarization of meetings," in *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021, pp. 1–6. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-6

[42] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: pre-training with extracted gap-sentences for abstractive summarization," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 11 328–11 339. [Online]. Available: http://proceedings.mlr.press/v119/zhang20ae.html

[43] O. Iakovenko, A. Andreeva, A. Lapidus, and L. Mikaelyan, "Team mts @ automin 2021: An overview of existing summarization approaches and comparison to unsupervised summarization techniques," in *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021, pp. 1–6. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-7

[44] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3730–3740. [Online]. Available: https://aclanthology.org/D19-1387

[45] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[46] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil, "Universal sentence encoder for English," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 169–174. [Online]. Available: https://aclanthology.org/D18-2029

[47] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[48] A. Garg, "Team symantlytical @ automin 2021: Generating readable minutes with gpt-2 and bert-based automatic minuting approach," in *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021, pp. 1–6. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-8

[49] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[50] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[51] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 632–642. [Online]. Available: https://aclanthology.org/D15-1075

[52] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019.

[53] U. Sharma and H. Singh, "Team the turing testament @ automin 2021: Feature engineering approach to creating meeting minutes using topsis," in *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021, pp. 1–5. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-9

[54] P. Williams and B. Haddow, "Team uedin @ automin 2022: Creating minutes by learning to filter an extracted summary," in *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021, pp. 1–4. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-10

[55] F. Schneider, S. Stüker, and V. Parthasarathy, "Team zoom @ automin 2021: Cross-domain pretraining for automatic minuting," in *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021, pp. 1–3. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-11

[56] Y. Zhao, M. Saleh, and P. J. Liu, "Seal: Segment-wise extractive-abstractive long-form text summarization," *ArXiv*, vol. abs/2006.10213, 2020.

[57] D. Varis and O. Bojar, "Sequence length is a domain: Length-based overfitting in transformer models," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8246–8257. [Online]. Available: https://aclanthology.org/2021.emnlp-main.650

[58] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. 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. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[59] M. Zhong, D. Yin, T. Yu, A. Zaidi, M. Mutuma, R. Jha, A. H. Awadallah, A. Celikyilmaz, Y. Liu, X. Qiu, and D. 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*. Online: Association for Computational Linguistics, Jun. 2021, pp. 5905–5921. [Online]. Available: https://aclanthology.org/2021.naacl-main.472

A. Sample Reference Minutes created by our Annotators

```
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 3: *A sample minute taken by our external annotators*

B. Sample Minutes from AutoMin participants

We present some minute samples from our participants' submissions to show the variety of automatically generated minutes by the various methods. One can easily see the quality of the minutes in terms of detailedness (coverage) and readability (grammatical correctness and fluency). For fair comparison we include the participant's generated minutes from the same meeting.

```
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 4: *A sample minute from Team ABC*

```
Summary:

* The Czech Republic government has already lifted the restrictions on commercial activity. From the
    4th of May, some commercial activity can already be open, but the free circulation of people
    can't start until June.
* Starting from June, PERSON5 and PERSON13 will be allowed to reach their families in the same
    region.
* PERSON5 went to the park yesterday.
  - Disc: The rules have changed since this Monday and now people can't go out even if they need to
      buy groceries.
  - Fact: People can go to the forest if they want, but they can't do anything else than that.
* PERSON6, PERSON5, and PERSON12 will have to postpone one of their regular meetings. They will meet
    in September instead of in May or June as they don't know if they will be able to travel
    somewhere.
* They need to vote in a Doodle to get the dates.
* PERSON5 has sent a link to a Google sheet with a couple of deliverables due in June.
  - Disc: They haven't decided who will be the internal review person for any of them, so it's a
      good opportunity to choose them now.
* If you would like to read one of the deliverables, you need to tell PERSON5.
* PERSON6 wants ORGANIZATION6 to review the summarization of the ASR system.
  -- Disc: PERSON15 will do the review.
* PERSON6 thinks they should ask for an extension for this deliver because Congress has not taken
    place and they contribute everything to the system.
* PERSON1 is looking for someone to review three points one.
  - Fact: ORGANIZATION4 is developing their own system in this area.
  - Disc: ORGANIZATION4 will do it. PERSON1, PERSON6, PERSON13, and PERSON16 will sign the form.
* PERSON6 wants to put PERSON16's name on the list, but he will let them know if he wants to be put
    there or not, so they will put him with two question marks.
* PERSON5 is writing a project management guide. It will be a lot of words.
  - Disc: The best position would be ORGANIZATION6, because they have coordinated a number of
      projects.
  - Disc: PERSON1 would rather read about ASR systems than read the guide.
* PERSON5 is looking for a project management guide for the new deliverable.
  - Fact: He doesn't have it. It's the first version of the deliverable and there's no description.
  - Disc: The project could have been managed without a guide for 15 months, if they don't have a
      guide.
* QT21 has period reports and data management plans, but not the project management guides.
  - Disc: PERSON6 thinks it probably was copy-pasted from something from somewhere. PERSON1 doesn't
      remember such a thing.
* PERSON5 will write something and someone will review it.
* The person who would like to coordinate future projects should have some incentive to read it.
  - Disc: ORGANIZATION8 could be asked to review the project because they don't know what EU
      projects are about yet.
* PERSON6 wants the internal reviews to be ready by the 8th of June.
  - Disc: The review is the first draft.
  - Fact: The reviewer has two weeks to fix it and a week for no further than one week with no more
      than one more week spare for final tracks from the coordinator.

.......
.......
------ Note ------
*: A topic or important point of the discussion.
Fact: An objective statement.
Disc: A subjective discussion such as an opinion and claim.
-: An item related to the topic or point.
--: A supplemental or supporting statement for its previous sentence.
```

Figure 5: *Excerpt of a sample minute from team Hitachi. The minutes from this team were usually longer.*

```
we have to wait until june for free circulation of people starting from fourth of june we are
    allowed to reach our family
people should have little incentive to read it is partner planning to do start anyone
things slept eighth of june sounds fine should be end of review
main responsible for deliverable is organization6 will not be confused from layout of test set
we do include latency sltf does include delay latency wasted effort there is two measures of wasted
    effort
i have strong preference not to submit my model to ehmm to organizers to run it for one unpublished
    code
i wanted to mention is forced alignment finds words in sound is not reliable for us it is shifted
ty bu zticha jo ty neru co potebuje coe ekni povleen jo potom prosim t to zvldnem pozdjc zkus to ty
    organization2 has experience with defending their approach to users
slt includes transform models in new generation there is not internal slt internal slt end - to -
    end asr
you could recover from to preforma to preserve stability reintroduce kind of correction
it 's better we final proposal informaly as doodle ask person4
she seen seminary em sub organization4 subtitles in future of page on projector in class
it could be low like french watching session to asr domains was challenging
it was hard to follow have to met 's safer way of selling what
we should sent email to person4 to truce his date would be towards end of next week
we should run it for ourselves mm - so sorry for not managing hour as
i need to peel potatoes thanks for joining will be in close touch for demo in coming days thank you
    thank you thank you
```

Figure 6: *Sample Minute from Team MTS (customized)*

```
[ORGANIZATION1] has announced that it will work with [ORGANIZATION2] to create a new field of data
    for the next generation of the group. [PERSON1] has been working on the project since the [
    ORGANIZATION2] introduced the project in July.
```

Figure 7: *Sample Minute from Team Zoom*

C. Detailed Manual Scores of English Minutes

Teams→	ABC		Auto Minuters		Hitachi		JU_PAD		M/F (baseline)		MTS (P/S)		Symantlytical		The Turing TESTament		UEDIN		Zoom†	
Test Meetings↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	5	5	2.5	3	4.5	5	3	3	2.5	3	1.5	2	2	2	2	2	2.5	4	1	1
en_test_002	4.5	5	2	2	5	5	3.5	4	2.5	3	1	1	2.5	3	3	3	1	1	1	1
en_test_003	4.5	5	2.5	3	4.5	5	3.5	4	2.5	3	1.5	2	2.5	3	3	3	1.5	2	1	1
en_test_004	5	5	3	3	4.5	5	3.5	4	3	4	1	1	3.5	4	3	3	3.5	4	1	1
en_test_005	4.5	5	2	2	4.5	5	3	3	1.5	2	1	1	3	3	3	4	1.5	2	1	1
en_test_006	4.5	5	2.5	3	4.5	5	3.5	4	2	2	1.5	2	2	2	3	3	1.5	2	1	1
en_test_007	4.5	5	4.5	5	4	5	4	5	3.5	4	2	2	3.5	5	4	4	1	1	1	1
en_test_008	5	5	2	2	4	5	3.5	4	2.5	3	2	2	2.5	3	3.5	4	1	1	1	1
en_test_009	5	5	3	3	4.5	5	3	3	3	4	1.5	2	3	3	3.5	4	1	1	1	1
en_test_010	4.5	5	3	4	4.5	5	3.5	4	2.5	3	1	1	3	3	4	4	2.5	3	1	1
en_test_011	4	5	2.5	4	3.5	4	3	3	2.5	3	2	2	2.5	3	2.5	3	2.5	3	1	1
en_test_012	4	5	2	3	3.5	4	3.5	4	3.5	4	2	2	2	3	2.5	3	1.5	2	1	1
en_test_013	3.5	5	2	3	4	4	3	3	3.5	4	1.5	2	2	3	2	3	2.5	3	1	1
en_test_014	3	4	2	3	4	4	2.5	3	2.5	3	2	2	2.5	3	3	3	3	3	1	1
en_test_015	3	4	2	3	4	4	2.5	3	3.5	4	2	3	2	3	2	3	2	3	1	1
en_test_016	4	5	2.5	4	5	5	3	3	2.5	3	2	2	2	3	3	3	2.5	3	1	1
en_test_017	3	4	1.5	2	4.5	5	3	3	3	3	1.5	2	2	3	3.5	4	2.5	3	1	1
en_test_018	3	5	2.5	4	4.5	5	3	3	2.5	3	2	3	2	3	3.5	4	2.5	3	1	1
en_test_019	4.5	5	2	3	3.5	4	2.5	3	2.5	3	1.5	2	3	3	3.5	4	2.5	3	1	1
en_test_020	3.5	5	2.5	3	5	5	2	3	2.5	3	1.5	2	3	3	4	5	2	2	1	1
en_test_021	4	4	2	3	4	4	2.5	3	1	1	1	1	2.5	3	2	3	1.5	2	1	1
en_test_022	4.5	5	1.5	2	4	5	2	3	2.5	3	1	1	2	3	2	3	3	3	1	1
en_test_023	4	4	1.5	2	3.5	4	2	2	2.5	3	1	1	2	3	2	3	3	3	1	1
en_test_024	3	3	2	3	4.5	5	2	3	2	3	1	1	2	3	2	3	2.5	3	1	1
en_test_025	3.5	4	2	3	3.5	4	2.5	3	2	3	1.5	2	2	3	2	3	2	2	1	1
en_test_026	3.5	4	2.5	4	4.5	5	2	3	1.5	2	1.5	2	2	3	2.5	4	2.5	3	1.5	2
en_test_027	2.5	3	2.5	3	4.5	5	2.5	3	2.5	3	1.5	2	3	3	3.5	4	2	2	2	2
en_test_028	4	4	2.5	3	4.5	5	2.5	3	3.5	4	1	1	3	3	4	5	2.5	3	1	1

Table 10: *Adequacy scores of the participants (assessed against the transcripts only). † marks a late submission.*

Teams→	ABC		Auto Minuters		Hitachi		JU_PAD		M/F (baseline)		MTS		Symantlytical		The Turing TESTament		UEDIN		Zoom†	
Test Meetings↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	4.5	5	2.5	3	5	5	2.5	3	2.5	3	1.5	2	2	2	2	2	4	5	1.5	2
en_test_002	5	5	2.5	3	5	5	3	3	2.5	3	1	1	3	4	2.5	3	4	4	2	3
en_test_003	5	5	2	2	5	5	4	5	2	3	1.5	2	3	3	2.5	3	4.5	5	3	5
en_test_004	5	5	2.5	3	4.5	5	3.5	5	3	4	1	1	3	4	3	3	4	5	1	1
en_test_005	5	5	1.5	2	4	5	3	3	1	1	1	1	2	2	2	2	3.5	5	1	1
en_test_006	4.5	5	3	3	4	5	4	5	2	2	1.5	2	3	3	2.5	3	3.5	5	1	1
en_test_007	4.5	5	4	5	4	5	3.5	4	3	3	2.5	3	3	4	2.5	3	3	5	1	1
en_test_008	5	5	2.5	3	4	5	3.5	4	2	2	2	2	2.5	3	2.5	3	4	5	2.5	4
en_test_009	5	5	3	3	4	5	3	3	2	2	2	3	3	3	3	3	3	5	1	1
en_test_010	5	5	3	4	4.5	5	3.5	4	2	2	1.5	2	3	3	2.5	3	3.5	5	3	5
en_test_011	4	4	3	3	3.5	4	3.5	4	2	2	1.5	2	2	3	1.5	2	3.5	4	1	1
en_test_012	4	4	2.5	3	3.5	4	3	3	3	3	1.5	2	2.5	3	2	2	4	4	4	5
en_test_013	4	4	2.5	3	3	3	3	3	2.5	3	1.5	2	2	2	2	2	3	3	3.5	5
en_test_014	4	4	2.5	3	4	5	2.5	3	2	2	1.5	2	1.5	2	2	3	3.5	4	2.5	4
en_test_015	4	4	2.5	3	4	5	2.5	3	2	2	1.5	2	2.5	3	1.5	2	3.5	4	1	1
en_test_016	4	4	2	2	3.5	4	3	3	2	2	1	1	2	3	2	3	3.5	4	1	1
en_test_017	4	4	2	2	3.5	4	3	3	2	3	2	2	2	3	2.5	3	3.5	4	2.5	4
en_test_018	3.5	4	2	2	3.5	4	2.5	3	2	2	1	1	2.5	3	2.5	3	3	4	2.5	4
en_test_019	4	4	2	2	3	4	2.5	3	1.5	2	1	1	2.5	3	2	2	3	3	3	5
en_test_020	4	4	2.5	3	3	3	2	3	2	2	1	1	3	3	2	2	2.5	3	3	5
en_test_021	4.5	5	2	3	4.5	5	1.5	2	1	1	1	1	2.5	3	2	3	2	2	4	5
en_test_022	4.5	5	2.5	3	4	5	4	4	3.5	4	1.5	2	3.5	4	3.5	4	3.5	4	3	5
en_test_023	4	5	2	2	4	5	2	2	3.5	4	1	1	3	3	3.5	4	3.5	4	3	5
en_test_024	4	5	3	4	4.5	5	2.5	3	2.5	3	1	1	3	3	2.5	3	3.5	4	3.5	5
en_test_025	3.5	4	2.5	3	3.5	4	3	3	3	3	1.5	2	3	3	3.5	4	3	4	1.5	2
en_test_026	4.5	5	3	4	3.5	4	3	3	2	2	1.5	2	3	4	3	4	3	3	2	3
en_test_027	3	4	3	3	4	4	3	3	2	2	1.5	2	3	3	3	3	2.5	3	3.5	5
en_test_028	3.5	4	2.5	3	3.5	4	2.5	3	3	3	1	1	3	3	3	3	2.5	3	3.5	5

Table 11: *Fluency scores of the participants (assessed against the transcripts only). † marks a late submission.*

Teams →	ABC		Auto Minuters		Hitachi		JU_PAD		M/F (baseline)		MTS (P/S)		Symantlytical		Turing TESTament		UEDIN		Zoom†	
Test Meeting	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	4	4	2	2	4.5	5	2	2	3	3	2.5	3	2.5	3	3	3	4.5	5	5	5
en_test_002	4.5	5	2.5	3	4	5	3	3	3	3	2.5	3	3.5	4	2	3	4.5	5	5	5
en_test_003	4.5	5	2.5	3	4.5	5	3	3	2.5	3	2	3	3.5	4	2.5	3	4.5	5	5	5
en_test_004	4	4	1.5	2	4	5	2.5	3	3	3	1	1	1.5	2	2	3	4	4	1	1
en_test_005	4.5	5	2.5	3	4	4	2.5	3	2.5	4	2	3	3.5	4	3.5	4	4	4	1	1
en_test_006	3.5	4	3	4	4	4	3	3	2	2	2.5	3	3	3	2	3	3	3	1	1
en_test_007	4	4	2.5	3	4	4	2.5	3	3	3	2.5	3	2.5	3	2	3	2.5	4	1	1
en_test_008	4.5	5	2	2	3	3	2.5	3	2.5	3	2.5	3	2.5	3	2	3	4	4	5	5
en_test_009	4.5	5	2.5	3	3.5	4	2.5	3	3	3	2	3	3	3	2	3	3.5	4	1	1
en_test_010	4.5	5	3	3	4	5	2.5	4	3.5	4	2	3	2	3	2	3	3.5	4	5	5
en_test_011	5	5	3.5	4	4.5	5	3.5	4	4	4	2.5	3	2.5	3	3	3	4	5	1	1
en_test_012	5	5	3	3	4.5	5	3	3	3.5	4	2	2	2.5	3	3	3	4	5	5	5
en_test_013	4.5	5	3	3	4.5	5	2.5	3	3	3	2	2	2.5	3	2.5	3	3.5	4	5	5
en_test_014	4.5	5	3	3	4.5	5	3	3	3	3	2	3	2.5	3	3	3	3.5	4	5	5
en_test_015	5	5	2.5	3	4.5	5	2.5	3	3	3	2	2	2.5	3	3	3	3.5	4	1	1
en_test_016	5	5	2.5	3	4.5	5	2.5	3	2.5	3	2	2	3	3	3	3	4	4	1	1
en_test_017	5	5	3	3	5	5	3	3	2.5	3	2.5	3	2.5	3	3	3	3.5	4	3.5	4
en_test_018	4.5	5	3	4	4.5	5	3	3	2.5	3	2.5	3	3	3	3	3	4.5	5	4.5	5
en_test_019	4.5	5	3	3	4.5	5	2.5	3	2.5	3	2	2	3	3	3	3	4	4	4.5	5
en_test_020	4.5	5	3.5	4	4.5	5	2.5	3	2	2	2	2	2.5	3	3	3	3.5	4	5	5
en_test_021	4.5	5	1.5	2	4.5	5	2	2	3.5	4	1	1	4	4	3.5	4	3.5	4	4.5	5
en_test_022	4.5	5	2.5	3	4.5	5	4	5	3.5	4	1.5	2	3.5	4	3.5	4	4.5	5	3.5	5
en_test_023	4.5	5	2.5	3	4.5	5	2.5	3	2.5	3	1	1	3.5	4	4	5	3	3	3.5	5
en_test_024	4	5	2	2	5	5	3	3	3	3	1.5	2	4	4	3.5	4	4.5	5	3.5	5
en_test_025	4	4	3.5	4	4.5	5	3.5	4	3.5	4	2	2	4	4	4	4	5	5	4	5
en_test_026	4.5	5	3	3	4.5	5	4	4	3.5	4	1.5	2	4.5	5	4	4	3	3	4	5
en_test_027	4.5	5	2.5	3	4.5	5	3	3	2.5	3	2.5	3	2.5	3	3.5	4	5	5	5	5
en_test_028	4	5	2.5	3	4.5	5	3.5	4	3	3	1	1	3.5	4	3.5	4	3.5	4	5	5

Table 12: *Grammatical Correctness scores of the participants (assessed against the transcripts only). † marks a late submission.*

D. Detailed Automatic Scores of English Minutes

Teams→	ABC		Auto Minuters		Hitachi		JU_PAD		M/F (baseline)		MTS (P/S)		Symantlytical		Turing TESTament		UEDIN		Zoom†	
Test Meetings↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	0.25	0.35	0.21	0.29	0.17	0.24	0.22	0.33	0.14	0.19	0.19	0.30	0.21	0.29	0.15	0.21	0.18	0.24	0.05	0.05
en_test_002	0.29	0.40	0.20	0.26	0.26	0.40	0.21	0.29	0.16	0.23	0.17	0.23	0.19	0.26	0.20	0.35	0.14	0.18	0.05	0.07
en_test_003	0.24	0.29	0.15	0.20	0.18	0.27	0.19	0.29	0.13	0.21	0.15	0.19	0.18	0.27	0.13	0.23	0.18	0.23	0.08	0.11
en_test_004	0.14	0.18	0.10	0.12	0.05	0.06	0.13	0.15	0.06	0.08	0.09	0.09	0.10	0.10	0.04	0.05	0.16	0.17	0.00	0.00
en_test_005	0.28	0.33	0.17	0.19	0.14	0.17	0.18	0.20	0.08	0.09	0.09	0.10	0.12	0.13	0.08	0.09	0.13	0.14	0.01	0.01
en_test_006	0.27	0.28	0.24	0.25	0.24	0.30	0.22	0.26	0.19	0.24	0.14	0.16	0.29	0.33	0.14	0.17	0.18	0.21	0.00	0.00
en_test_007	0.25	0.31	0.19	0.27	0.17	0.25	0.22	0.29	0.15	0.22	0.19	0.24	0.22	0.29	0.09	0.15	0.08	0.13	0.00	0.01
en_test_008	0.29	0.46	0.18	0.27	0.18	0.27	0.17	0.28	0.10	0.15	0.17	0.25	0.17	0.29	0.14	0.21	0.14	0.22	0.05	0.05
en_test_009	0.36	0.42	0.22	0.25	0.23	0.24	0.31	0.34	0.12	0.15	0.29	0.33	0.29	0.34	0.14	0.16	0.19	0.24	0.00	0.01
en_test_010	0.28	0.33	0.21	0.26	0.23	0.31	0.27	0.35	0.15	0.19	0.05	0.08	0.24	0.28	0.16	0.23	0.15	0.24	0.07	0.09
en_test_011	0.24	0.31	0.21	0.28	0.18	0.24	0.19	0.26	0.19	0.24	0.20	0.28	0.18	0.24	0.10	0.15	0.15	0.19	0.00	0.00
en_test_012	0.29	0.31	0.26	0.27	0.31	0.32	0.31	0.38	0.22	0.24	0.22	0.25	0.25	0.27	0.32	0.32	0.14	0.19	0.07	0.08
en_test_013	0.19	0.33	0.12	0.24	0.10	0.23	0.14	0.23	0.13	0.25	0.06	0.09	0.12	0.24	0.09	0.20	0.16	0.19	0.08	0.12
en_test_014	0.21	0.27	0.15	0.19	0.18	0.21	0.14	0.17	0.11	0.16	0.14	0.19	0.12	0.16	0.09	0.14	0.18	0.20	0.02	0.05
en_test_015	0.21	0.21	0.13	0.13	0.15	0.16	0.16	0.19	0.15	0.15	0.10	0.11	0.17	0.18	0.09	0.10	0.20	0.31	0.00	0.00
en_test_016	0.35	0.44	0.26	0.33	0.25	0.41	0.22	0.29	0.22	0.35	0.23	0.28	0.24	0.33	0.21	0.35	0.17	0.20	0.01	0.02
en_test_017	0.26	0.34	0.21	0.33	0.19	0.33	0.22	0.32	0.14	0.24	0.04	0.08	0.23	0.38	0.14	0.26	0.18	0.26	0.04	0.07
en_test_018	0.31	0.33	0.21	0.28	0.25	0.33	0.23	0.25	0.20	0.26	0.14	0.15	0.24	0.29	0.18	0.25	0.17	0.23	0.05	0.06
en_test_019	0.41	0.41	0.35	0.35	0.26	0.26	0.34	0.34	0.21	0.21	0.18	0.18	0.33	0.33	0.18	0.18	0.27	0.27	0.02	0.02
en_test_020	0.29	0.29	0.23	0.23	0.27	0.27	0.15	0.15	0.20	0.20	0.08	0.08	0.24	0.24	0.17	0.17	0.15	0.15	0.08	0.08
en_test_021	0.22	0.22	0.15	0.15	0.12	0.12	0.15	0.15	0.10	0.10	0.00	0.00	0.14	0.14	0.08	0.08	0.17	0.17	0.03	0.03
en_test_022	0.21	0.21	0.15	0.15	0.09	0.09	0.12	0.12	0.07	0.07	0.11	0.11	0.13	0.13	0.06	0.06	0.22	0.22	0.09	0.09
en_test_023	0.24	0.24	0.28	0.28	0.25	0.25	0.27	0.27	0.20	0.20	0.04	0.04	0.25	0.25	0.16	0.16	0.22	0.22	0.07	0.07
en_test_024	0.30	0.30	0.25	0.25	0.26	0.26	0.30	0.30	0.24	0.24	0.12	0.12	0.25	0.25	0.19	0.19	0.18	0.18	0.05	0.05
en_test_025	0.42	0.42	0.30	0.30	0.30	0.30	0.31	0.31	0.29	0.29	0.23	0.23	0.31	0.31	0.21	0.21	0.25	0.25	0.06	0.06
en_test_026	0.40	0.40	0.32	0.32	0.39	0.39	0.37	0.37	0.31	0.31	0.07	0.07	0.36	0.36	0.33	0.33	0.19	0.19	0.04	0.04
en_test_027	0.32	0.37	0.25	0.26	0.35	0.38	0.26	0.33	0.22	0.23	0.13	0.14	0.25	0.29	0.26	0.27	0.15	0.16	0.06	0.07
en_test_028	0.37	0.46	0.24	0.28	0.32	0.34	0.20	0.26	0.28	0.32	0.03	0.04	0.23	0.26	0.23	0.26	0.18	0.18	0.03	0.03

Table 13: *ROUGE-1 scores of the participants against the test set reference minutes. Team MTS submitted three runs, we enlist the best run here. M/F→Matus_Francesco. † marks a late submission.*

Teams →	ABC		Auto Minuters		Hitachi		JU_PAD		M/F (baseline)		MTS (P/S)		Symantlytical		Turing TESTament		UEDIN		Zoom†	
Test Meetings↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	0.03	0.05	0.05	0.10	0.05	0.09	0.04	0.07	0.03	0.04	0.04	0.08	0.04	0.08	0.05	0.08	0.03	0.05	0.00	0.00
en_test_002	0.07	0.12	0.05	0.09	0.07	0.13	0.04	0.07	0.03	0.08	0.04	0.09	0.03	0.06	0.06	0.12	0.02	0.06	0.01	0.03
en_test_003	0.02	0.05	0.02	0.04	0.04	0.05	0.02	0.03	0.01	0.02	0.01	0.01	0.02	0.04	0.02	0.04	0.02	0.04	0.00	0.00
en_test_004	0.02	0.03	0.01	0.01	0.01	0.02	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00
en_test_005	0.06	0.09	0.02	0.02	0.04	0.05	0.03	0.03	0.01	0.01	0.00	0.00	0.01	0.01	0.02	0.02	0.03	0.03	0.00	0.00
en_test_006	0.03	0.05	0.02	0.03	0.05	0.05	0.03	0.04	0.03	0.04	0.01	0.02	0.06	0.06	0.05	0.06	0.02	0.04	0.00	0.00
en_test_007	0.05	0.08	0.03	0.04	0.05	0.09	0.04	0.07	0.04	0.06	0.03	0.03	0.05	0.09	0.03	0.04	0.00	0.00	0.00	0.00
en_test_008	0.07	0.12	0.02	0.05	0.05	0.10	0.02	0.04	0.02	0.04	0.04	0.09	0.04	0.08	0.04	0.07	0.02	0.04	0.00	0.00
en_test_009	0.11	0.15	0.06	0.07	0.05	0.06	0.11	0.12	0.02	0.02	0.05	0.06	0.06	0.09	0.04	0.05	0.02	0.03	0.00	0.00
en_test_010	0.05	0.07	0.03	0.05	0.04	0.06	0.04	0.06	0.03	0.04	0.01	0.03	0.05	0.08	0.04	0.06	0.04	0.08	0.01	0.01
en_test_011	0.06	0.08	0.04	0.07	0.04	0.06	0.03	0.06	0.04	0.05	0.03	0.04	0.03	0.05	0.02	0.04	0.02	0.02	0.00	0.00
en_test_012	0.07	0.07	0.09	0.12	0.10	0.11	0.08	0.11	0.05	0.06	0.06	0.08	0.06	0.08	0.12	0.13	0.04	0.07	0.00	0.00
en_test_013	0.04	0.07	0.02	0.05	0.02	0.05	0.02	0.06	0.01	0.03	0.00	0.00	0.01	0.04	0.01	0.03	0.03	0.05	0.00	0.00
en_test_014	0.04	0.04	0.01	0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.04	0.01	0.01	0.02	0.03	0.03	0.03	0.00	0.00
en_test_015	0.03	0.04	0.02	0.04	0.03	0.05	0.03	0.05	0.03	0.05	0.00	0.01	0.03	0.03	0.02	0.03	0.05	0.10	0.00	0.00
en_test_016	0.12	0.19	0.06	0.09	0.08	0.16	0.05	0.10	0.05	0.07	0.04	0.06	0.06	0.12	0.08	0.16	0.03	0.04	0.00	0.00
en_test_017	0.03	0.05	0.05	0.08	0.04	0.06	0.03	0.06	0.03	0.04	0.01	0.02	0.05	0.08	0.04	0.06	0.03	0.05	0.00	0.00
en_test_018	0.06	0.08	0.03	0.06	0.06	0.09	0.06	0.08	0.04	0.06	0.01	0.02	0.06	0.08	0.05	0.08	0.03	0.05	0.01	0.02
en_test_019	0.13	0.13	0.13	0.13	0.10	0.10	0.13	0.13	0.08	0.08	0.04	0.04	0.10	0.10	0.07	0.07	0.10	0.10	0.00	0.00
en_test_020	0.05	0.05	0.02	0.02	0.06	0.06	0.02	0.02	0.03	0.03	0.00	0.00	0.02	0.02	0.04	0.04	0.04	0.04	0.01	0.01
en_test_021	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.00	0.00	0.00	0.00	0.03	0.03	0.04	0.04	0.04	0.04	0.00	0.00
en_test_022	0.07	0.07	0.01	0.01	0.05	0.05	0.05	0.05	0.04	0.04	0.00	0.00	0.03	0.03	0.02	0.02	0.08	0.08	0.00	0.00
en_test_023	0.05	0.05	0.08	0.08	0.06	0.06	0.07	0.07	0.03	0.03	0.00	0.00	0.05	0.05	0.04	0.04	0.05	0.05	0.00	0.00
en_test_024	0.08	0.08	0.04	0.04	0.08	0.08	0.07	0.07	0.06	0.06	0.00	0.00	0.04	0.04	0.07	0.07	0.11	0.11	0.00	0.00
en_test_025	0.11	0.11	0.04	0.04	0.09	0.09	0.07	0.07	0.06	0.06	0.00	0.00	0.03	0.03	0.05	0.05	0.09	0.09	0.00	0.00
en_test_026	0.13	0.13	0.08	0.08	0.13	0.13	0.05	0.05	0.06	0.06	0.03	0.03	0.09	0.09	0.12	0.12	0.03	0.03	0.00	0.00
en_test_027	0.08	0.10	0.05	0.07	0.10	0.11	0.06	0.09	0.05	0.05	0.01	0.01	0.05	0.06	0.06	0.07	0.05	0.07	0.00	0.00
en_test_028	0.11	0.13	0.06	0.09	0.09	0.11	0.05	0.07	0.08	0.11	0.00	0.00	0.04	0.04	0.07	0.09	0.06	0.07	0.00	0.01

Table 14: *ROUGE-2 scores of the participants against the test set reference minutes. Team MTS submitted three runs, we enlist the best run here. M/F→Matus_Francesco. † marks a late submission.*

Teams →	ABC		Auto Minuters		Hitachi		JU_PAD		Matus_Francesco		MTS (P/S)		Symantlytical		Turing TESTament		UEDIN		Zoom†	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	0.13	0.19	0.13	0.19	0.10	0.16	0.12	0.20	0.08	0.11	0.10	0.16	0.09	0.14	0.09	0.12	0.11	0.17	0.03	0.03
en_test_002	0.18	0.24	0.13	0.19	0.16	0.26	0.13	0.17	0.09	0.16	0.11	0.17	0.12	0.18	0.14	0.27	0.10	0.13	0.04	0.07
en_test_003	0.13	0.14	0.09	0.11	0.10	0.14	0.09	0.12	0.07	0.10	0.08	0.10	0.10	0.13	0.07	0.11	0.10	0.13	0.06	0.08
en_test_004	0.08	0.09	0.05	0.05	0.03	0.04	0.07	0.08	0.04	0.05	0.03	0.04	0.05	0.05	0.02	0.03	0.10	0.11	0.00	0.00
en_test_005	0.13	0.18	0.08	0.09	0.07	0.08	0.09	0.10	0.03	0.04	0.04	0.04	0.05	0.05	0.04	0.04	0.08	0.10	0.01	0.01
en_test_006	0.14	0.14	0.13	0.13	0.10	0.12	0.11	0.12	0.08	0.10	0.07	0.08	0.16	0.17	0.07	0.07	0.12	0.13	0.00	0.00
en_test_007	0.12	0.15	0.09	0.12	0.10	0.13	0.11	0.14	0.08	0.10	0.09	0.11	0.11	0.15	0.05	0.08	0.07	0.13	0.00	0.01
en_test_008	0.16	0.25	0.09	0.14	0.10	0.15	0.10	0.15	0.06	0.08	0.12	0.17	0.11	0.18	0.09	0.13	0.11	0.14	0.04	0.05
en_test_009	0.21	0.30	0.11	0.14	0.11	0.13	0.17	0.23	0.06	0.06	0.15	0.19	0.13	0.16	0.08	0.10	0.10	0.16	0.00	0.01
en_test_010	0.15	0.19	0.12	0.15	0.12	0.15	0.13	0.17	0.08	0.09	0.04	0.07	0.12	0.16	0.10	0.13	0.10	0.16	0.05	0.07
en_test_011	0.15	0.18	0.10	0.15	0.09	0.13	0.10	0.14	0.11	0.12	0.09	0.12	0.09	0.12	0.07	0.09	0.09	0.10	0.00	0.00
en_test_012	0.18	0.21	0.20	0.21	0.19	0.23	0.19	0.25	0.14	0.14	0.14	0.18	0.15	0.15	0.22	0.25	0.10	0.15	0.05	0.06
en_test_013	0.12	0.20	0.07	0.14	0.06	0.12	0.08	0.14	0.06	0.11	0.03	0.04	0.07	0.14	0.05	0.10	0.10	0.13	0.06	0.08
en_test_014	0.11	0.13	0.06	0.07	0.09	0.10	0.08	0.08	0.05	0.07	0.08	0.10	0.06	0.06	0.05	0.07	0.10	0.11	0.01	0.02
en_test_015	0.12	0.13	0.07	0.07	0.08	0.09	0.10	0.13	0.08	0.09	0.04	0.05	0.10	0.10	0.05	0.06	0.14	0.23	0.00	0.00
en_test_016	0.22	0.29	0.13	0.17	0.14	0.21	0.14	0.19	0.11	0.16	0.11	0.14	0.12	0.19	0.12	0.20	0.10	0.12	0.01	0.02
en_test_017	0.12	0.15	0.11	0.16	0.09	0.13	0.11	0.14	0.06	0.09	0.03	0.08	0.11	0.17	0.08	0.13	0.13	0.18	0.03	0.04
en_test_018	0.18	0.21	0.11	0.15	0.13	0.18	0.14	0.15	0.10	0.15	0.07	0.08	0.13	0.17	0.11	0.17	0.10	0.13	0.04	0.06
en_test_019	0.27	0.27	0.20	0.20	0.16	0.16	0.22	0.22	0.15	0.15	0.11	0.11	0.18	0.18	0.14	0.14	0.21	0.21	0.02	0.02
en_test_020	0.15	0.15	0.13	0.13	0.12	0.12	0.09	0.09	0.10	0.10	0.05	0.05	0.12	0.12	0.08	0.08	0.13	0.13	0.05	0.05
en_test_021	0.12	0.12	0.12	0.12	0.08	0.08	0.09	0.09	0.10	0.10	0.00	0.00	0.09	0.09	0.05	0.05	0.13	0.13	0.03	0.03
en_test_022	0.16	0.16	0.11	0.11	0.08	0.08	0.11	0.11	0.07	0.07	0.05	0.05	0.07	0.07	0.05	0.05	0.18	0.18	0.04	0.04
en_test_023	0.12	0.12	0.12	0.12	0.10	0.10	0.13	0.13	0.09	0.09	0.03	0.03	0.11	0.11	0.06	0.06	0.13	0.13	0.05	0.05
en_test_024	0.17	0.17	0.13	0.13	0.11	0.11	0.16	0.16	0.09	0.09	0.05	0.05	0.10	0.10	0.12	0.12	0.16	0.16	0.02	0.02
en_test_025	0.24	0.24	0.14	0.14	0.17	0.17	0.18	0.18	0.15	0.15	0.10	0.10	0.14	0.14	0.12	0.12	0.17	0.17	0.04	0.04
en_test_026	0.21	0.21	0.18	0.18	0.23	0.23	0.19	0.19	0.14	0.14	0.06	0.06	0.21	0.21	0.21	0.21	0.11	0.11	0.04	0.04
en_test_027	0.19	0.25	0.11	0.14	0.17	0.21	0.15	0.21	0.11	0.13	0.04	0.04	0.12	0.15	0.13	0.14	0.10	0.14	0.04	0.05
en_test_028	0.22	0.27	0.14	0.18	0.17	0.20	0.13	0.18	0.16	0.21	0.02	0.02	0.11	0.13	0.13	0.17	0.12	0.12	0.02	0.02

Table 15: *ROUGE-L scores of the participants against the test set reference minutes. Team MTS submitted three runs, we enlist the best run here. M/F→Matus_Francesco. † marks a late submission.*

E. Late and Additional Submission Evaluation

Teams →	M/F (coref)†		M/F (final)†		MTS (T5)		MTS (Pegasus)		MTS (customized)	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	2.5	3	4	4	1	1	1	1	1	1
en_test_002	3	3	1.5	2	1	1	1	1	1	1
en_test_003	3	3	2	3	1	1	1	1	1.5	2
en_test_004	2.5	3	3.5	4	1	1	1	1	1.5	2
en_test_005	2	2	2.5	3	1	1	1	1	2	2
en_test_006	2.5	3	2.5	3	1	1	1	1	1	1
en_test_007	4	4	4	4	1	1	1	1	2.5	3
en_test_008	2.5	3	2.5	3	1	1	1	1	2	2
en_test_009	2.5	3	2.5	3	1.5	2	1	1	2	2
en_test_010	4	4	2.5	3	1	1	1	1	3	4
en_test_011	2	2	3.5	4	1	1	1	1	2	2
en_test_012	3.5	4	4	4	1	1	1.5	2	2	3
en_test_013	1.5	2	2.5	3	1.5	2	2	3	2	2
en_test_014	3	3	4.5	5	1.5	2	1.5	2	2.5	3
en_test_015	2.5	3	3.5	4	1	1	1.5	2	1.5	2
en_test_016	3	3	3	4	1	1	1.5	2	2.5	3
en_test_017	3	3	3.5	5	1	1	1.5	2	2	2
en_test_018	2.5	3	2	2	1.5	2	1.5	2	2	2
en_test_019	3.5	4	3	4	1.5	2	1.5	2	1.5	2
en_test_020	3	3	2.5	3	1	1	1.5	2	2	3
en_test_021	1	1	0	0	1	1	1.5	2	1.5	2
en_test_022	2.5	3	2.5	3	1	1	1.5	2	1.5	2
en_test_023	2.5	3	3.5	4	1	1	1	1	2	2
en_test_024	2	3	1.5	2	1	1	1	1	1.5	2
en_test_025	2	2	3.5	4	1	1	1	1	1.5	2
en_test_026	3	3	4	5	1	1	1	1	2	3
en_test_027	3	3	2	2	1	1	1	1	2	2
en_test_028	3	3	2.5	3	1.5	2	2	2	2.5	3

Table 16: *Adequacy scores of additional and late (†) submissions*

Teams →	M/F (coref)†		M/F (final)†		MTS (T5)		MTS (Pegasus)		MTS (customized)	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	3	4	3	3	2	2	2	2	1	1
en_test_002	3	3	1.5	2	2.5	3	2.5	3	1.5	2
en_test_003	3	3	2	3	1.5	2	4	5	1.5	2
en_test_004	2.5	3	3.5	4	3	5	2.5	4	1.5	2
en_test_005	2	2	2.5	3	1.5	2	2.5	3	2	2
en_test_006	2	2	2.5	3	2.5	4	1.5	2	2	2
en_test_007	3.5	4	3.5	4	2	3	1	1	2.5	3
en_test_008	2	2	2	2	1.5	2	1.5	2	1.5	2
en_test_009	2.5	3	2.5	3	1.5	2	1.5	2	2	2
en_test_010	4	4	3.5	4	1.5	2	2	3	3	4
en_test_011	2.5	3	4	4	1	1	1.5	2	2.5	3
en_test_012	3	3	4.5	5	1.5	2	1.5	2	2.5	3
en_test_013	2	2	2.5	3	2.5	3	2	2	2.5	3
en_test_014	3	3	4	4	3	4	1.5	2	1.5	2
en_test_015	2.5	3	3.5	4	1.5	2	1	1	1.5	2
en_test_016	2.5	3	4.5	5	1.5	2	2.5	3	2.5	3
en_test_017	2.5	3	4	5	2	2	2	2	2	2
en_test_018	2	2	3	4	1.5	2	2	2	1.5	2
en_test_019	3	3	3.5	5	1.5	2	2	2	1.5	2
en_test_020	3	4	2.5	3	2	2	1.5	2	1.5	2
en_test_021	3	3	0	0	1.5	2	2	2	1.5	2
en_test_022	3.5	4	3	4	1	1	1.5	2	1.5	2
en_test_023	2.5	3	4	4	2.5	4	1	1	2	2
en_test_024	2.5	3	2.5	3	1	1	1	1	2	2
en_test_025	2.5	3	4	5	1	1	1	1	2.5	3
en_test_026	3	3	4	5	1.5	2	2	3	2	2
en_test_027	3	3	3.5	5	1	1	1	1	2	2
en_test_028	3	3	3	4	1.5	2	2	2	2	2

Table 17: *Fluency scores of additional and late (†) submissions*

Teams →	M/F (coref)†		M/F (final)†		MTS (T5)		MTS (Pegasus)		MTS (customized)	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	3.5	4	4	4	2.5	3	2.5	3	1.5	2
en_test_002	3	3	3.5	4	3.5	4	2.5	3	2.5	3
en_test_003	3	3	3	4	2	2	3	3	2.5	3
en_test_004	3	3	3	3	2.5	3	2.5	3	2	2
en_test_005	3	4	3	4	2.5	3	3	3	3	3
en_test_006	3	3	3	3	4	4	3	3	3	3
en_test_007	3	4	3.5	4	2.5	3	2.5	4	3	3
en_test_008	3	3	3	3	2.5	4	2.5	3	2.5	3
en_test_009	3	3	3	3	3	3	3	4	2.5	3
en_test_010	3	4	4	4	3	3	3.5	4	3	3
en_test_011	3	3	3.5	4	2	3	2.5	3	2.5	3
en_test_012	3	4	3.5	4	2.5	3	2.5	3	3.5	4
en_test_013	3	3	3	3	3.5	4	3	3	3	3
en_test_014	3.5	4	3.5	5	3.5	4	2.5	3	1.5	2
en_test_015	3.5	4	3.5	4	1	1	1	1	1.5	2
en_test_016	3.5	4	5	5	2	2	3	3	2	2
en_test_017	3.5	4	4	5	2	2	2	2	2	2
en_test_018	3	3	4	4	3.5	4	3	3	2	2
en_test_019	3	3	4	5	3	4	2	2	1	1
en_test_020	3.5	4	3.5	4	3	3	3	3	2.5	3
en_test_021	4	4	0	0	3	3	2.5	3	2.5	3
en_test_022	3.5	4	4	5	2	2	2	2	2.5	3
en_test_023	3.5	4	3.5	4	3.5	4	2.5	3	2.5	3
en_test_024	2.5	3	3	3	1.5	2	1.5	2	1.5	2
en_test_025	3	3	4.5	5	2	3	2	2	2.5	3
en_test_026	3.5	4	4	5	2.5	3	3.5	4	2	2
en_test_027	3.5	4	4.5	5	1	1	3.5	4	2	2
en_test_028	2.5	3	4	4	2.5	3	3	3	2	2

Table 18: *Grammatical correctness scores of additional and late (†) submissions*

Teams →	M/F (coref)†		M/F (final)†		MTS (T5)		MTS (Pegasus)		MTS (customized)	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	0.19	0.25	0.15	0.20	0.01	0.02	0.01	0.03	0.13	0.18
en_test_002	0.22	0.33	0.12	0.17	0.06	0.12	0.10	0.23	0.07	0.13
en_test_003	0.19	0.27	0.17	0.24	0.05	0.07	0.04	0.08	0.12	0.15
en_test_004	0.08	0.10	0.10	0.13	0.05	0.07	0.07	0.11	0.13	0.15
en_test_005	0.14	0.17	0.13	0.16	0.01	0.01	0.05	0.08	0.15	0.16
en_test_006	0.22	0.27	0.24	0.28	0.11	0.15	0.11	0.13	0.15	0.18
en_test_007	0.21	0.30	0.16	0.26	0.08	0.11	0.06	0.08	0.19	0.24
en_test_008	0.17	0.26	0.13	0.20	0.07	0.11	0.09	0.13	0.13	0.19
en_test_009	0.22	0.22	0.17	0.19	0.13	0.16	0.13	0.14	0.14	0.16
en_test_010	0.23	0.29	0.22	0.26	0.05	0.07	0.11	0.12	0.24	0.27
en_test_011	0.17	0.18	0.26	0.28	0.02	0.02	0.04	0.05	0.22	0.23
en_test_012	0.34	0.37	0.22	0.22	0.04	0.04	0.09	0.11	0.20	0.21
en_test_013	0.13	0.17	0.15	0.27	0.04	0.04	0.06	0.09	0.09	0.16
en_test_014	0.17	0.24	0.12	0.17	0.07	0.08	0.07	0.09	0.15	0.17
en_test_015	0.12	0.13	0.24	0.28	0.08	0.10	0.06	0.07	0.17	0.19
en_test_016	0.23	0.33	0.26	0.30	0.05	0.09	0.02	0.03	0.22	0.24
en_test_017	0.18	0.29	0.20	0.25	0.02	0.04	0.08	0.10	0.17	0.19
en_test_018	0.25	0.29	0.16	0.21	0.06	0.08	0.12	0.18	0.21	0.23
en_test_019	0.33	0.33	0.24	0.24	0.05	0.05	0.04	0.04	0.25	0.25
en_test_020	0.29	0.29	0.15	0.15	0.01	0.01	0.07	0.07	0.21	0.21
en_test_021	0.07	0.07	0.00	0.00	0.05	0.05	0.08	0.08	0.23	0.23
en_test_022	0.17	0.17	0.19	0.19	0.03	0.03	0.04	0.04	0.21	0.21
en_test_023	0.19	0.19	0.27	0.27	0.02	0.02	0.01	0.01	0.20	0.20
en_test_024	0.28	0.28	0.16	0.16	0.01	0.01	0.04	0.04	0.23	0.23
en_test_025	0.32	0.32	0.21	0.21	0.04	0.04	0.04	0.04	0.21	0.21
en_test_026	0.30	0.30	0.24	0.24	0.04	0.04	0.06	0.06	0.21	0.21
en_test_027	0.30	0.32	0.10	0.10	0.00	0.00	0.02	0.02	0.17	0.17
en_test_028	0.30	0.36	0.16	0.20	0.03	0.06	0.07	0.10	0.21	0.26

Table 19: *ROUGE-1 scores of additional and late (†) submissions*

Teams →	M/F (coref)†		M/F (final)†		MTS (T5)		MTS (Pegasus)		MTS (customized)	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	0.03	0.04	0.03	0.03	0.00	0.00	0.00	0.00	0.01	0.03
en_test_002	0.05	0.11	0.04	0.08	0.00	0.02	0.02	0.06	0.01	0.04
en_test_003	0.02	0.03	0.03	0.05	0.00	0.00	0.00	0.00	0.01	0.03
en_test_004	0.01	0.02	0.01	0.02	0.00	0.00	0.00	0.00	0.03	0.05
en_test_005	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.02	0.03	0.04
en_test_006	0.03	0.05	0.04	0.06	0.03	0.06	0.01	0.02	0.03	0.04
en_test_007	0.05	0.08	0.04	0.05	0.01	0.03	0.01	0.02	0.03	0.05
en_test_008	0.05	0.10	0.05	0.10	0.00	0.00	0.00	0.00	0.01	0.02
en_test_009	0.05	0.05	0.05	0.06	0.03	0.04	0.02	0.04	0.02	0.03
en_test_010	0.04	0.06	0.05	0.09	0.00	0.00	0.01	0.02	0.03	0.06
en_test_011	0.02	0.03	0.05	0.07	0.00	0.00	0.00	0.00	0.04	0.06
en_test_012	0.10	0.12	0.05	0.05	0.00	0.00	0.01	0.03	0.06	0.06
en_test_013	0.02	0.03	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.01
en_test_014	0.01	0.02	0.02	0.03	0.02	0.03	0.02	0.02	0.02	0.03
en_test_015	0.01	0.02	0.06	0.11	0.00	0.00	0.00	0.00	0.02	0.03
en_test_016	0.05	0.09	0.04	0.06	0.00	0.00	0.00	0.01	0.03	0.03
en_test_017	0.04	0.05	0.04	0.08	0.01	0.02	0.02	0.02	0.03	0.04
en_test_018	0.05	0.08	0.04	0.07	0.00	0.00	0.00	0.01	0.04	0.07
en_test_019	0.12	0.12	0.07	0.07	0.01	0.01	0.00	0.00	0.08	0.08
en_test_020	0.04	0.04	0.01	0.01	0.00	0.00	0.02	0.02	0.03	0.03
en_test_021	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.06	0.06
en_test_022	0.05	0.05	0.09	0.09	0.00	0.00	0.00	0.00	0.08	0.08
en_test_023	0.01	0.01	0.08	0.08	0.00	0.00	0.00	0.00	0.02	0.02
en_test_024	0.05	0.05	0.02	0.02	0.00	0.00	0.00	0.00	0.10	0.10
en_test_025	0.07	0.07	0.02	0.02	0.00	0.00	0.00	0.00	0.03	0.03
en_test_026	0.05	0.05	0.04	0.04	0.01	0.01	0.02	0.02	0.04	0.04
en_test_027	0.05	0.06	0.03	0.05	0.00	0.00	0.00	0.00	0.02	0.03
en_test_028	0.09	0.11	0.02	0.03	0.00	0.00	0.03	0.05	0.06	0.09

Table 20: *ROUGE-2 scores of additional and late (†) submissions*

Teams →	M/F (coref)†		M/F (final)†		MTS (T5)		MTS (Pegasus)		MTS (customized)	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	0.10	0.13	0.10	0.13	0.01	0.02	0.01	0.02	0.06	0.09
en_test_002	0.14	0.22	0.08	0.13	0.05	0.09	0.07	0.15	0.05	0.08
en_test_003	0.10	0.12	0.11	0.13	0.03	0.03	0.02	0.03	0.07	0.09
en_test_004	0.05	0.07	0.06	0.07	0.04	0.05	0.05	0.07	0.08	0.08
en_test_005	0.07	0.09	0.07	0.08	0.01	0.01	0.05	0.08	0.08	0.09
en_test_006	0.11	0.14	0.12	0.13	0.09	0.13	0.09	0.13	0.09	0.09
en_test_007	0.11	0.14	0.08	0.10	0.06	0.09	0.06	0.08	0.08	0.09
en_test_008	0.11	0.18	0.10	0.17	0.05	0.08	0.07	0.10	0.06	0.09
en_test_009	0.13	0.16	0.11	0.13	0.08	0.08	0.07	0.08	0.08	0.10
en_test_010	0.13	0.15	0.11	0.15	0.03	0.05	0.07	0.10	0.10	0.11
en_test_011	0.11	0.11	0.14	0.16	0.02	0.02	0.03	0.04	0.10	0.13
en_test_012	0.21	0.26	0.14	0.17	0.03	0.03	0.05	0.06	0.12	0.12
en_test_013	0.10	0.13	0.07	0.13	0.03	0.04	0.04	0.07	0.05	0.08
en_test_014	0.08	0.10	0.06	0.07	0.07	0.08	0.05	0.07	0.08	0.09
en_test_015	0.07	0.08	0.14	0.18	0.07	0.10	0.05	0.07	0.10	0.12
en_test_016	0.11	0.17	0.10	0.13	0.04	0.07	0.02	0.03	0.09	0.11
en_test_017	0.09	0.12	0.11	0.16	0.02	0.04	0.07	0.10	0.11	0.16
en_test_018	0.14	0.19	0.10	0.12	0.04	0.04	0.07	0.11	0.11	0.15
en_test_019	0.20	0.20	0.14	0.14	0.04	0.04	0.04	0.04	0.16	0.16
en_test_020	0.12	0.12	0.07	0.07	0.01	0.01	0.06	0.06	0.10	0.10
en_test_021	0.07	0.07	0.00	0.00	0.05	0.05	0.06	0.06	0.15	0.15
en_test_022	0.12	0.12	0.17	0.17	0.03	0.03	0.04	0.04	0.13	0.13
en_test_023	0.09	0.09	0.15	0.15	0.01	0.01	0.01	0.01	0.09	0.09
en_test_024	0.12	0.12	0.07	0.07	0.01	0.01	0.02	0.02	0.18	0.18
en_test_025	0.17	0.17	0.07	0.07	0.03	0.03	0.03	0.03	0.12	0.12
en_test_026	0.16	0.16	0.12	0.12	0.02	0.02	0.04	0.04	0.10	0.10
en_test_027	0.14	0.18	0.07	0.08	0.00	0.00	0.02	0.02	0.09	0.10
en_test_028	0.15	0.19	0.07	0.08	0.01	0.02	0.06	0.07	0.11	0.15

Table 21: *ROUGE-L scores of additional and late (†) submissions*

G The University of Edinburgh's Submission to the First Shared Task on Automatic Minuting (to be published soon)

Team UEDIN @ AutoMin 2021: Creating Minutes by Learning to Filter an Extracted Summary

Philip Williams, Barry Haddow

School of Informatics, University of Edinburgh, Scotland
philip.williams.edin@icloud.com , bhaddow@ed.ac.uk

Abstract

We describe the University of Edinburgh's submission to the First Shared Task on Automatic Minuting. We developed an English-language minuting system for Task A that combines BERT-based extractive summarization with logistic regression-based filtering and rule-based pre- and post-processing steps. In the human evaluation, our system averaged scores of 2.1 on adequacy, 3.9 on grammatical correctness, and 3.3 on fluency.
Index Terms: automatic minuting, extractive summarization, meeting summarization

1. Introduction

The University of Edinburgh participated in the main task of the First Shared Task on Automatic Minuting [1]. We developed a pipelined system that employs (more-or-less) off-the-shelf extractive summarization together with rule-based and learned components. The output of our system is a short list of bullet points (roughly 3% the length of the original transcript) together with a list of participants. The bullet points are sentences derived from the original transcript, but cleaned up to remove speech disfluencies and transcription artifacts. Figure 1 shows a sample of the resulting minutes.

While automatic minuting is a form of meeting summarization, minutes typically emphasize certain aspects of the meeting, such as decisions that have been reached or actions that are to be taken. We found that extractive summarization alone produced mixed results in terms of selecting sentences that are appropriate for use in minutes. We therefore employed a post-summarization step that filtered the summarizer output. To do this we first hand-labelled a sample of summarizer output for the training data then trained a logistic regression model to score extracted sentences according to their 'minute-worthiness.'

Ultimately, we did not make use of the minutes provided in the training data except as a guide for making system design choices. This was due in part to the wide variety of minuting styles used by the annotators, with wide variations in minute length as well as structural choices, such as grouping bullet points by topic or by speaker. This diversity of styles made the data challenging to utilize effectively for machine learning and is an aspect of minuting that sets it apart from other summarization tasks.

Submissions to this task were evaluated automatically using ROUGE [2] and manually using human judgment of adequacy, fluency, grammatical correctness. In the human evaluation, our system averaged scores of 2.1 on adequacy, 3.9 on grammatical correctness, and 3.3 on fluency. During system development we evaluated minuting quality on the dev set using ROUGE-1 and ROUGE-2 scores computed using sacreROUGE [3] and we report those scores in this paper. In practice, we found that ROUGE scores calculated against the supplied references were not sufficiently reliable to differentiate systems and we relied on manual sample checking for making most model design choices.

```
Attendees: PERSON1, PERSON2, PERSON3,
PERSON4, PERSON5

 * [PERSON1]: I plan to go there, but like,
   we need a back-up person.
 * For the [PROJECT2] event.
 * We need someone to take care of the
   recording, so the archiver person.
 * [PERSON3]: I think we need to improve
   our segmenter, the worlds are getting
   revised fine.
 * [PERSON3]: I'll first ask him to correct
   the current [PROJECT6]L for the correct
   type we have.
 * Maybe it will be better for us to attend
   the call with the [PERSON7].
 * We will separately need to ship the
   audio to the English [PROJECT5]
   separately.
 * [PERSON1]: If you have good data for the
   language pair, then yes, it is better to
   go directly.
```

Figure 1: *Sample system output (test meeting 27).*

2. The Automatic Minuting Pipeline

Figure 2 outlines the pipeline that transforms a raw transcript into minutes. This section describes the individual steps in detail.

2.1. Preprocessing

The preprocessing step of the pipeline performs three main tasks: it normalizes speaker attributions, records the list of participants, and removes speech artifacts.

2.1.1. Speaker Attribution

The raw transcripts contain a speaker attribution, such as `[PERSON5]`, at the start of each turn. Since summarization will be performed at the sentence-level, we copy speaker attributions to the start of each sentence in order that attributions will persist through the summarization and filtering steps (although we may later choose to discard some of them). The training data uses a mix of square and round brackets, which we standardize as square brackets.

2.1.2. Participant List

The list of participants is recorded at this stage since later steps may remove the contributions of some speakers.

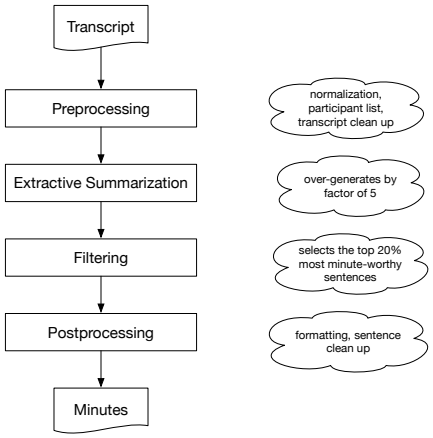


Figure 2: *The automatic minuting pipeline.*

2.1.3. *Removal of Speech and Transcription Artifacts*

The raw meeting transcripts faithfully reproduce speech disfluencies as well as adding annotations in the form of tags, such as `<laugh>` or `<other_language>`. We use hand-written rules to repair sentences where possible and to remove sentences that are incomplete. Specifically, we:

- Remove filler words, such as `um`, `er`, and `ehm`;
- Repair restarts such as, `we sh-, we should`;
- Remove sentences containing the `<unintelligible>` tag
- Remove incomplete words (ending with the – character);
- Remove incomplete sentences (not ending `.` or `?`); and
- Remove any remaining annotation tags.

For example, after preprocessing, the sentence,

```
[PERSON3]: will send uh, SLT findings uh,
findings PDF. <parallel_talk>
```

becomes,

```
[PERSON3]: will send SLT findings PDF.
```

The removal of incomplete and unintelligible (or partially unintelligible) sentences constitutes a significant level of filtering prior to the main summarization and filtering stages, reducing the total number of sentences by approximately 28%

2.2. Extractive Summarization

For summarization, we used `lecture-summarizer`[1] [4] with minor modifications. `lecture-summarizer` is an extractive summarizer based on the BERT [5] pre-trained language model. It is designed to summarize transcripts of university lectures. In brief, it works by using BERT to encode sentences, clustering the sentence embeddings, and then finding the nearest sentence to the centroid of each cluster. The number of clusters is configurable and can be specified either as a fixed number of sentences or as a ratio. We chose a ratio of 0.035, which is toward the lower-end of transcript / minute ratios in the training data. We chose the exact value based on personal preference, since the ROUGE metric was an unreliable guide, almost always favouring longer minutes.

[1] `https://github.com/dmmiller612/lecture-summarizer`

2.2.1. *Modifications to* `lecture-summarizer`

`lecture-summarizer` uses spaCy [6] to split the input into sentences. We found that the spaCy sentencizer would separate speaker attribution tags from sentences, and since our transcripts had already been split into sentences, we modified the code to use line breaks as sentence delimiters instead.

In order that we could over-generate sentences (ahead of the subsequent filtering stage), we modified the code to produce the k closest sentences to each centroid (k was set to 5 in our final system).

2.3. Filtering

In preliminary systems, we noticed that the output of the summarizer would typically contain a small number of sentences that were perfect to include in the minutes as-is, among a larger number of sentences that were unsuitable, because they were irrelevant, vague, or lacked context. While the suitability of many sentences is borderline or subject to opinion, it was clear that there were some features that could differentiate the most suitable sentences from the least. We therefore tried hand-labelling a set of sentences produced by the summarizer from the training and dev meetings and training a regression model to score candidate sentences.

In total, we labelled 1,107 sentences from the training meetings, assigning 266 (24.0%) to the positive class (minute-worthy) and 841 (76.0%) to the negative class. We labelled 451 sentences from the dev meetings, assigning 176 (41.9%) to the positive class and 244 (58.1%) to the negative class.

For model development, we created a balanced training set containing all 266 positive examples and 266 randomly sampled negative examples from the training meetings. Similarly, we created a test set containing all 176 positive examples and 176 randomly sampled negative examples from the dev meetings.

We used scikit-learn [7] to train a logistic regression model with unigram and bigram TF-IDF features. On our test set, this achieved a precision of 66.7%, recall of 65.9% and F1 score of 66.3%. Figure 3 gives a sample of low- and high-scoring sentences from our test set.

In the pipeline, we used this model to score candidate sentences produced by the summarization step, taking the top-scoring 20%

2.4. Postprocessing

Postprocessing performs three final tasks that are primarily stylistic: it removes conjunctions and exclamations from the starts of sentences, selectively drops speaker attributions, and formats the participant list and summary.

2.4.1. *Conjunction and Exclamation Removal*

Spoken sentences frequently begin with a conjunction, such as `so` or `because` or with an exclamation such as `yeah` or `oh`. We remove these when they occur at the start of a sentence, based on a list of words observed in the output of the training data.

```
0.03 So does it work, if you are not searching
     for any words?
0.06 I, maybe I just didn't compile it
     properly.
0.07 So probably it's not that serious.
0.08 but I think, you saying, you've got an
     email from a project officer?
0.10 I 'm not, you know, wanting is like,
     yes?

0.90 So I agree with what [PERSON6] suggested
     that [PERSON4] and [PERSON12] should
     focus on the selection of the input.
0.92 and we have the reviewer chosen.
0.92 This week I work on do the collection is
     business for [OTHER1] and English.
0.95 I still have to look at it and then we
     have to prepare for the posters.
0.98 We will buy some extra time for from
     them.
```

Figure 3: *A sample of low-scoring and high-scoring sentences, as scored by our logistic-regression model.*

2.4.2. Speaker Attribution Removal

In order that the minutes appear less like direct speech, we remove the speaker attributions for any sentence that does not include a reference to the first or second person (`I`, `me`, `your`, etc.).

2.4.3. Formatting

Finally, we format and output the list of participants that was saved during preprocessing and we add bullet points to the summary.

3. Experiments

Here we give results for our submitted system in contrast to some baseline and variant systems that were created during system development. We used ROUGE for evaluation, since that is the task's primary automatic metric, although, as already mentioned, we found it to have limited use during system development.

The systems are as follows:

`baseline-random` Randomly selects sentences from the transcript. No pre- or postprocessing except for speaker attribution normalization, bullet points, and participant list generation.

`baseline-lecsum` As `baseline-random` but uses `lecture-summarizer` to select sentences.

`submitted` Submitted system

`no-filter` As `submitted`, but does not include filtering step and does not over-generate during extractive summarization.

Table 1 gives average ROUGE-1 and ROUGE-2 scores on the dev set for systems tuned to produce output of approximately the same length. In the case that multiple references were available for a meeting, we computed scores against all references and took the maximum. Note that this differs from the official evaluation method, which takes the average.

While the submitted system is the highest-scoring (on ROUGE-2), the differences in score are small and we found that substantial differences in quality between system were not reflected in the scores.

System	ROUGE-1	ROUGE-2
`baseline-random`	27.8	5.0
`baseline-lecsum`	29.6	5.8
`submitted`	29.4	6.5
`no-filter`	26.0	5.4

Table 1: *ROUGE-1 and ROUGE-2 scores on the dev set.*

4. Discussion

The minutes produced by our system give a sense of what a meeting was the about and tend to include at least some of the actions and outcomes. However, the minutes are unsatisfactory in a number of important ways, largely resulting from the use of extractive summarization:

- Many sentences lack context and are unable to stand alone;
- The minutes contain direct speech where reported speech would be more natural;
- There is no means for the system to encapsulate portions of the meeting in a single sentence (e.g. '[PERSON4] and [PERSON7] discussed arrangements for the upcoming conference');
- The minutes are unstructured.

In an attempt to address the problem of sentences lacking context, we experimented with coreference resolution, using the neuralcoref[2] package to replace corefering mentions with main mentions. However, we found that harmful substitutions (where correct terms were replaced with incorrect ones) were more common than beneficial substitutions.

Before developing our current system, we briefly experimented with abstractive summarization (specifically with the Pegasus model [8] in Huggingface [9]). Abstractive summarization is appealing for this task and would potentially solve at least some of the problems listed above. It has been successfully applied to meeting summarization for the AMI [10] and ICSI [11] datasets [12]. However, it was unclear to us how to address some significant challenges posed by this dataset, most notably, the diversity in minuting styles and the length of the transcripts (even the Longformer[13] model available in Huggingface 'only' supports an input of 4,096 tokens, which is far short of the meeting transcript lengths).

In a first attempt to make the data more amenable to learning with a transformer-based model, we began chunking the transcript and manually aligning bullet points, with the goal of creating smaller training examples, but found this was difficult in practice for many of the minutes due to the extreme summarization and restructuring of material in the minutes.

We also attempted to segment the transcripts into parts that could be tackled separately during inference. We experimented with the NLTK implementation of TextTiling [14] but on inspecting results, it didn't appear to pick up meaningful boundaries.

[2] `https://github.com/huggingface/neuralcoref`

5. Conclusion

We have described the University of Edinburgh's submission to the First Shared Task on Automatic Minuting. Our minuting system was based on extractive summarization with logistic regression-based filtering and rule-based pre- and post-processing steps. While our system performed satisfactorily in terms of grammatical correctness and fluency, it performed less well in terms of adequacy, which we attribute to the use of extractive summarization.

6. Acknowledgments

This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460 (ELITR).

7. References

[1] T. Ghosal, O. Bojar, M. Singh, and A. Nedoluzhko, "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, pp. 1–25. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-1

[2] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[3] D. Deutsch and D. 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)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 120–125. [Online]. Available: https://www.aclweb.org/anthology/2020.nlposs-1.17

[4] D. Miller, "Leveraging BERT for extractive text summarization on lectures," *CoRR*, vol. abs/1906.04165, 2019. [Online]. Available: http://arxiv.org/abs/1906.04165

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[6] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.1212303

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," 2019.

[9] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[10] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*, S. Renals and S. Bengio, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 28–39.

[11] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The icsi meeting corpus," 2003, pp. 364–367.

[12] X. Feng, X. Feng, and B. Qin, "A survey on dialogue summarization: Recent advances and new frontiers," *CoRR*, vol. abs/2107.03175, 2021. [Online]. Available: https://arxiv.org/abs/2107.03175

[13] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020.

[14] M. A. Hearst, "Texttiling: A quantitative approach to discourse segmentation," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, mar 1997.