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

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, Ondřej Bojar
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague
(nedoluzhko,singh,hledikova,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 relatively unstructured writing act and is affected by who takes the minutes and for whom the minutes are intended. With the rise of online meetings, automatic minuting would be an important use-case for the meeting participants and those who might have missed the meeting. However, automatically generating meeting minutes is a challenging problem due to various factors, including the quality of automatic speech recognition (ASR), public availability of 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, consisting of transcripts generated from ASRs, manually corrected, and minuted by several annotators. Our dataset, *ELITR Minuting Corpus*, consists of 120 English and 59 Czech meetings, covering about 180 hours of meeting content. The corpus is publicly available at `http://hdl.handle.net/11234/1-4692` 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, ELITR Minuting Corpus is probably the first resource on minuting in English and also in a language other than English (Czech).

Keywords: automatic minuting, meeting summarization, multi-party dialogues

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 the necessary context switching give 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 essential issues, action points, decisions, or proposed activities discussed during the meeting. Manually writing minutes takes time and distracts attention from the discussion. Hence we believe that an automatic minuting solution will be a practical 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 (Ghosal et al., 2022).

Although both may seem similar, for this paper, we make a distinction between the task of *meeting summarization* and *minuting*. Whereas meeting summarization intends to sum up the central concepts of the meeting (and 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 take the form of a structured bulleted list of important 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 because it includes meetings in Czech and not just English as all similar datasets we are aware of in the literature.

The two existing benchmark meeting datasets in English, the AMI (Mccowan et al., 2005) and the ICSI (Janin et al., 2003) corpus are aimed at meeting summarization. They contain meeting transcripts, extractive summaries (selected relevant transcript lines), and abstractive summaries in the form of coherent paragraphs. Our *ELITR Minuting Corpus*[1] is comparable in size to AMI and ICSI. However, we differ in three significant aspects: (i) we focus on minuting, so our summaries are organized as bulleted lists, typical for project meeting minutes we have more commonly seen; (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. Minuting is a subjective activity. Different people may have different perspectives and objectives while writing a meeting minute. Hence in our dataset, we include multiple minutes written independently by different persons to observe the variance of outputs when humans are carrying out the task. We used ELITR Minuting Corpus to conduct the AutoMin shared task at Interspeech 2021 (Ghosal et al., 2021). AutoMin aspires to be a community-driven initiative to attempt this complex yet a timely task.

2. Related Work

Given the lack of proper minuting datasets, we survey a few existing datasets on meeting and dialogue summarization,

[1] our dataset derives its name from the H2020 European Live

(A) Meeting transcript segment:

(PERSON10) 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?
Because [PERSON16] already asks uh, asked a while ago.
Uh, are you making any steps in this, or decisions?
(PERSON14) No, I haven't done any steps and decisions,
I thought sort of you'd ask with doing it-
(PERSON10) Yeah.
(PERSON14) And the, coordinating.
(PERSON10) Yeah.
(PERSON14) So what what's your propose?
I mean, what we have proposed in the a in the offline track seems quite a reasonable. [...]
(PERSON10) Uh, uh, so let's start with the um, um, with the uh, uh, the the postponed review.
So it's seems, that people have not uh-
So [PERSON11], uh, please, let let us know what this doodle is.
This is that we need to figure out, the date.
(PERSON22) Okay, so it's because the the review will postpone till September.
We should give uh, our project officer the new ah, a new date.
And I see more people finally voted it, so- [...]
(PERSON10) Whether we want get little time extension, uh,
uh, little or longer time extension uh, of the project.
So I don't know if [PERSON22] is aware any date until we should make our uh, mind. [...]
(PERSON19) Um, if we um, ask for an extension, I will be <unintelligible/> automatically.
(PERSON10) Okay. [...]

(B) Meeting minutes by annotator 1:

- [PROJECT9] remote presentations organization
  - Discussion about the results: agreement on the pre-recorded presentation for the [PROJECT5] 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

(C) Meeting minutes by annotator 2:

- Organization of the [PROJECT9] presentations
  - There is organized a panel and there will always have 2 time slots for the presentation of the papers.
  - The papers, also the presentations, can be pre-recorded.
- Postponed review
  - It is needed to figure out the date.
  - Because the review will postpone till September and is needed to get to project officer a new date.
- Time extension
  - Agree, that by mid next week everybody should fill the table by how many months would you like the project to be extended.

Figure 1: An example from ELITR Minuting Corpus showing a segment from a meeting transcript (A) along with two independently created corresponding minutes (B, C). As the data has been anonymized, "PERSON*number*" and "PROJECT*number*" denote persons' and projects' placeholders respectively. The underlining has been added only in this figure for an improved presentation.

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).

Translator (ELITR) project of which it originated: `https://elitr.eu`

The AMI and ICSI are the most commonly used for meeting summarization experiments among the meeting datasets. The AMI Meeting corpus (Mccowan et al., 2005) contains 100 hours of meeting discussions, two-thirds of which are, however, meetings enacted artificially according to a script (not real meetings). 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 from 17 to 103 minutes and involve 3 to 10 participants. Interestingly, the corpus contains many non-native English speakers, varying in fluency from nearly native to challenging-to-transcribe. Other meeting collections are substantially smaller, such as 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 councils).

MEDIASUM is another conversational dataset with 463.6K transcripts and short abstractive summaries of Public Radio (NPR) and CNN television interviews from multiple domains (Zhu et al., 2021). 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-patient conversations (Krishna et al., 2020) and some others.

Table 1 compares our dataset with relevant others, distinguishing meeting collections (top) and other dialogue corpora (bottom of the table). Except for our data, we reuse the statistics reported by (Zhu et al., 2021). Among the meeting collections, only *ELITR Minuting Corpus* 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 meeting transcripts in English and Czech and their corresponding minutes. The English part includes project meetings from the computer science domain, with

Dataset	A	B	C	D	E	F	G	H	I	J
ELITR Minuting Corpus (English)	MM	project meetings	✓	✓	✓	120	7,066	373	727	5.9
ELITR Minuting Corpus (Czech)	MM	project meetings	✓	✓	✓	59	8,534	236	1,205	7.6
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 (for ELITR Minuting Corpus, we average across multiple summaries regardless of the meeting they belong to), I – avg. turns per transcript, J - avg. number of speakers.

prevailing non-native speakers of English. The discussions in the Czech part are from computer science and public administration domains; participants in Czech meetings are primarily Czech native speakers, but some Slovak native speakers also appeared (speaking Slovak because Czech and Slovak are mutually intelligible). 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 exceptions, whereas meetings longer than two hours are topic-oriented mini-workshops, also rather occasional.

In ELITR Minuting Corpus, a meeting usually contains one manually corrected transcript, one original minute[2] (created by a meeting participant; in some cases, these minutes are a detailed agenda that got further updated during or after the meeting), and one or more minutes are generated by our annotators. Original minutes are missing for some meeting sessions, but each meeting must contain at least one generated minute. To conform to GDPR and the consent 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

Our minuting corpus consists primarily of 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 has to be used to record the audio. There are also a 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 (Nguyen et al., 2020; Kratochvíl et al., 2020). 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 high-varying recording conditions and domain-specific terminology, the ASR outputs are often of low quality. We also note that the difference between ASR performance in the lab and real-world settings is striking, see, e.g., Macháček et al. (2019). Along with the recordings, we also collected original minutes prepared by one of the meeting participants. These minutes are stored together with the specially created minutes (described in Section 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 utterances Here, we divide the ASR outputs into utterances (lines in the transcript). Our transcript segmentation is based on syntactic and prosodic features of utterances.[3]

The main segmentation criterion is the **syntactic** one. The annotators were instructed to distinguish speech segments that roughly correspond to sentences in the written discourse. Thus, for example, the sequence "*So the – ... We should be muting ourselves when we are not talking*" should be annotated as two utterances because it consists of two sentence-like segments, the unfinished "*So the –*" and "*We should be muting ourselves when we are not talking*". On the other hand, "*Lets – okay, let's get started anyway*" is annotated as one utterance. Although it includes correction, it is further continued with the same syntactic structure and corresponds to one sentence-like segment.

However, in spontaneous speech, sentence boundaries are not always easy to distinguish, so another segmentation hint is the **prosodic** one: It is always preferable to break transcripts at such locations where the speaker pauses. So longer statements of one speaker, which could be hardly syntactically segmented, were recommended to break wherever the most noticeable pauses occur.

[2]We use the non-standard singular "minute" to highlight that we are talking about a single instance of meeting minutes.

[3]Our decision for this issue is different from the one applied in AMI and ICSI meeting corpora, where the segmentation into dialogue acts (DAs) had been primarily based on their functions. The reason for our different decision is that segmentation based on the functions of utterances turned out to be unreliable and too subjective in the case of our transcripts. We believe that it is more adequate to define communicative functions of utterances after the text segmentation, choosing more formal criteria for segmentation itself.

As a general rule, no segment should be longer than a minute, but most of them are much shorter. For an example of our transcript segmentation, see part A of Figure 1.

Diarize the transcript The change of speakers is always marked as a new utterance. The speaker's code is inserted in round brackets at the beginning of the first utterance of the speaker.

Correct the transcript The low-quality ASR outputs have been manually corrected from the linguistic point of view.[4] The annotators recognize and correct the sequences of words used in the meeting. A transcription should be an accurate record of what was actually said; no changes are made to make transcripts more grammatically correct. For example, if the participant says "*I dunno*", this should be transcribed as it is heard, not as "*I don't know*". The transcript should include all repeated words, grammatical and semantic errors (e.g., incorrect word choice in non-native speech).

This step also includes adding correct letter casing and adequate punctuation. The intuitive guidelines for using punctuation marks and dealing with abbreviations and numbers are given to annotators similar to those used in the AMI and ICSI meeting corpora. Word fragments, unfinished words or sentences are graphically marked (for example, "*let-*" for the unfinished word "*letter*" or "*I decided to –*" for e.g. "*I decided to resign*"). Parentheses within the transcribed texts are exclusively used for cases where the annotator is unsure if they recognized the word or phrase correctly.

Special vocal tags describe sounds that are made using the mouth or nose but that do not have standard lexical representations, such as <laugh/>, <cough/>, or <other_noise/>. Unique tags are used for unrecognized speech segments (<unintelligible/>), utterances or words in a language that is different from the meeting language (<another_language/>), speech recorded within the meeting but not part of it (for example, if one of the meeting members has a telephone call, <parallel_talk/>), etc.
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).

The data pre-processing stage is very time-consuming. Cleaning and preparing a meeting transcript (breaking it into utterances, diarizing it, and correcting it) for subsequent computational processing takes approximately four times as long as the meeting's duration.

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, some general recommendations for creating minutes include being concise and concrete, avoiding overuse of people's 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 50 words). Longer minutes contain hundreds (and occasionally even thousands) of words.

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 for 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 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 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 assume that the names of these entities cannot lead to personal identification.

Person, organisation, project and location (in specific cases) names were replaced with the lexical substitute strings [PERSON*number*], [ORGANIZATION*number*], [PROJECT*number*] and [LOCATION*number*] respectively. Additionally, we replaced the names of annotators mentioned in minutes with [ANNOTATOR*number*]. 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.[5] 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. Apart from this, some information like phone numbers and passwords also must be removed. These were replaced with the strings [URL], [PASSWORD], [EMAIL], [PHONE], [NUMBER] and [PATH].

The de-identification was completed within our web-based tool (Polák et al., 2022) and finalized in a series of scans using Unix text processing tools across the whole corpus.

[4] In some particular cases, where the sound quality of the recording was especially poor, or the accents of speakers were especially strong, it was easier to transcribe the recording manually from the very beginning rather than correcting the ASR output. However, such cases were rare.

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

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 2: Overall statistics of ELITR Minuting Corpus

Lang	Set	Total # Meetings	Total # Words	Per Meeting: # 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 3: Transcript size statistics of ELITR Minuting Corpus. We report averages ± standard deviations.

3.5. Alignment between Transcripts and Minutes

Having prepared the transcripts and minutes, the next step was equipping a subset of the data with manual alignments between transcript utterances and minutes lines. This was done by our annotators using the annotation tool ALIGNMEET[6] (Polák et al., 2022), which was specifically designed for this purpose. We hope that this type of annotation will be helpful in more precise evaluation methods for the meeting minuting task in the future, with segment-level evaluation and transcript coverage calculation. For details on this idea and a pilot study, see Polák et al. (2022).

The alignment maps each utterance to either the line of the minutes in which it is summarized, a problem label, both, or neither. The alignments are done in such a way that the whole long piece of conversation is aligned to the same minutes line, which summarizes it. Most of our minutes cover the transcripts completely and mention all critical 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 five problem types, which our annotators were assigning to sections in transcripts as needed. These types are *Organizational* (organizational talk not directly related to the subject of the meeting, e.g., discussing technical issues with the video call), *Speech incomprehensible* (it is not clear what the speaker is saying), *Small talk* (casual conversation unrelated to the subject of the meeting, e.g., discussing the weather), *Censored* (a section of the transcript removed for privacy reasons) and *Other issue*. A single utterance can be assigned both a summarizing minutes line and a problem label, as well as for it to remain completely unaligned.

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

Lang	Set	Total # Minutes	Per Average Minute: # 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 4: Minuting size statistics of ELITR Minuting Corpus.

3.6. Final ELITR Minuting Corpus

The final layout of the corpus is captured in Table 2. 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 that have even more manual minutes: up to 11 and 3 on average.

The last column in Table 2 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 3 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).

	English		Czech	
Meeting Minuted	#meetings	#hours	#meetings	#hours
Once	30	28	8	9
Twice	65	67	20	20
More than twice	25	22	31	31
Total	120	117	59	60

Table 5: Duration of the recordings and number of minutes per meeting for ELITR Minuting Corpus.

Table 4 summarizes the statistics of the minutes in the corpus. For instance, the 18 English test set meetings have in total of 55 minutes. We report the average of averages in the subsequent columns: an average minute (across the multiple minutes created for a given meeting) has about 279 words and 36 lines. About 5.6 persons are mentioned in a minute on average.

Table 5 shows how many meetings have been minuted once, twice, or more than twice and informs about the total duration of the collected meetings in hours.

In Table 1, we position our ELITR Minuting Corpus with respect to the meeting summarization datasets: AMI (Mccowan et al., 2005) and 61 sessions of ICSI (Janin et al., 2003) and dialogue summarization datasets, namely MEDIASUM (Zhu et al., 2021), SAMSUM (Gliwa et al., 2019), CRD3 (Rameshkumar and Bailey, 2020), DiDi (Liu et al., 2019a) and MultiWoz (Budzianowski et al., 2020). We report average words per transcript and summary, turns per transcript and number of speakers. We can clearly notice only ELITR Minuting Corpus has multiple summaries for single transcripts.

3.7. Manual Evaluation for Human Annotated Minutes

To better understand the quality of minutes in our dataset, we manually evaluated three meetings[7] which had been independently minuted by 8, 8, and 11 people, respectively. In five experts, we scored the minutes on a scale of 1 (worst) – 5 (best) according to several generally accepted summary estimation criteria, such as adequacy, topicality, readability, relevance, grammaticality, fluency, coverage, informativeness, and coherence (Kryściński et al., 2019; Zhu et al., 2020; Lee et al., 2020).

We ranged the minutes according to these criteria and thoroughly discussed the details to understand (i) what we based our judgments on and (ii) which criteria appeared to be the most useful and comprehensive. The results show very similar assessments by different experts: for example, all experts selected the same minutes as the best ones. As for evaluation criteria, we found *adequacy* (the judgment if summary sentences represent conclusions clearly visible in the transcripts of the discussions), *relevance* (how well the summary sums up the main idea of the meeting) and *topicality* (whether summary sentences cover topics that are discussed in the transcript) most helpful. Our typical objections were, for example, missing relevant information, unclear extractive segments revealing no content value, misunderstanding the content, including non-relevant information, or chaotic structure. However, most (24 out of 26) minutes were evaluated as acceptable in the experiment. Surprisingly, “winners" for each meeting were minutes specially created by our annotators. Original minutes (created by meeting participants) included too much unnecessary information or, on the contrary, were too short.

However, the criteria used for manual evaluation are still relatively informal, and their rigorous definition and assessment of inter-annotator agreement are part of our future work.

3.8. 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.

3.9. Handling Ethical Issues

All meeting participants gave their consent to make the data publicly available. We provided participants with a preview of the full texts for the meetings they participated in to check the de-identified transcripts and minutes by themselves and ensure that no unwanted sensitive information would be 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 sensitive 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, the participants cannot effectively give informed consent without actually browsing the data planned to be released. For that reason, we consider it obligatory to give all participants the possibility to preview and check the final version of the data before the release.

In the case of our dataset, although we had prior consent from all the participants, we performed this additional check of the de-identified transcripts and minutes. 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. Experiments

We evaluate the performance of state-of-art text summarization models on the English part of ELITR Minuting Corpus. We assess both extractive and abstractive methods of summarization. Extractive methods, given a transcript, select a subset of the words or sentences which best represent the discussion of the meeting. Abstractive methods generate a concise minute that captures the salient notions of the meet-

[7]They are *meeting_en_train_039*, *meeting_en_test_007* and *meeting_en_dev_010* in ELITR Minuting Corpus.

	Method	BLEU	BERTScore_F1	ROUGE_1	ROUGE_2	ROUGE_WE	ROUGE_L
dev	(A) BART-xsum-samsum	6.17±3.88	49.02±24.99	42.41±22.03	11.13±5.84	10.88±5.63	23.25±12.71
dev	(A) BART	≀ 6.26±2.89	46.40±23.94	37.48±18.37	8.93±4.18	8.71±4.54	18.82±9.57
dev	(A) T5	5.16±3.12	≀ 49.07±24.81	36.41±18.26	7.30±3.40	8.07±3.67	≀ 19.17±9.96
dev	(A) Pegasus	3.26±2.13	41.30±20.39	33.80±16.94	6.73±3.49	7.50±3.90	17.30±9.19
dev	(A) BERT2BERT	≀ 5.50±3.04	27.33±13.94	32.72±16.45	6.25±3.19	6.59±3.36	15.38±8.26
dev	(E) LSA	5.48±3.15	≀ 30.02±12.84	31.03±16.05	≀ 7.08±4.15	≀ 7.99±4.32	≀ 15.54±8.68
dev	(E) Luhn	5.31±2.87	23.09±10.50	30.37±15.63	≀ 7.23±4.01	≀ 8.72±4.62	15.18±8.44
dev	(E) LexRank	5.25±2.44	≀ 35.52±15.47	29.70±16.33	4.85±3.16	6.22±3.37	13.84±8.31
dev	(E) TextRank	≀ 5.46±3.02	25.22±11.27	29.27±15.35	≀ 6.35±3.64	≀ 8.68±4.39	≀ 13.99±8.01
dev	(A) BERTSUM	4.47±2.43	≀ 47.64±22.63	28.55±15.85	4.20±2.32	4.43±2.02	≀ 14.33±8.25
dev	(A) Roberta2Roberta	≀ 4.96±3.18	44.11±22.92	25.67±12.42	≀ 4.32±2.12	3.48±1.55	13.29±7.14
dev	(E) TF-IDF	4.22±2.15	10.84±5.62	22.38±12.01	≀ 5.91±3.32	≀ 7.18±3.69	12.16±6.89
dev	(A) LED	3.92±2.92	≀ 48.83±29.15	14.85±8.29	1.73±0.80	0.94±0.32	9.79±5.85
test	(A) BART-xsum-samsum	5.64±2.56	50.27±30.96	37.92±22.78	7.80±4.74	6.37±4.17	21.09±12.47
test	(A) BART	≀ 5.98±2.97	≀ 50.31±31.00	35.20±20.98	7.27±3.92	≀ 7.43±4.64	19.48±11.18
test	(A) Pegasus	3.09±2.03	49.10±29.93	34.16±20.48	6.33±3.75	≀ 9.08±5.95	19.04±11.48
test	(A) T5	≀ 5.62±2.84	≀ 49.94±30.83	34.02±20.21	≀ 7.38±4.23	≀ 11.38±6.79	18.78±11.03
test	(E) LSA	≀ 5.72±2.30	25.70±15.16	32.57±19.16	≀ 7.72±4.63	9.82±6.21	18.45±10.94
test	(A) BERTSUM	5.31±2.07	≀ 51.16±29.77	31.50±19.15	5.38±3.48	6.39±4.17	17.58±10.61
test	(A) BERT2BERT	≀ 5.32±2.72	32.62±19.90	31.06±18.16	≀ 5.41±3.21	5.73±4.19	15.33±9.14
test	(E) Luhn	≀ 5.85±2.31	20.11±11.94	30.60±18.22	≀ 7.06±4.15	≀ 9.56±6.16	≀ 17.19±10.22
test	(E) TextRank	5.82±2.24	≀ 21.28±12.66	30.35±17.96	6.69±3.83	9.13±5.94	16.72±9.90
test	(A) Roberta2Roberta	4.55±2.54	≀ 50.30±30.90	24.89±14.48	4.53±2.53	5.40±2.93	14.39±8.27
test	(E) TF-IDF	≀ 4.70±1.95	12.97±7.84	24.11±13.99	≀ 6.61±3.71	≀ 8.62±5.63	≀ 14.53±8.45
test	(A) LED	3.70±3.41	≀ 46.18±24.96	15.36±7.99	1.29±0.61	0.86±0.69	10.75±5.70
test2	(A) BART-xsum-samsum	7.92±1.53	30.45±19.58	22.18±12.20	6.10±3.07	4.54±2.80	12.39±6.57
test2	(A) BART	7.72±1.54	≀ 35.96±18.38	20.37±11.26	5.61±3.31	≀ 5.24±2.84	10.85±6.30
test2	(A) T5	7.30±1.59	33.37±18.36	19.71±10.37	5.59±2.79	≀ 6.25±2.90	10.34±5.59
test2	(A) BERT2BERT	7.10±1.54	19.87±9.86	19.29±9.61	4.40±1.99	5.26±2.34	9.14±4.78
test2	(A) Pegasus	3.65±2.67	≀ 32.53±16.39	19.15±10.07	≀ 5.21±2.64	≀ 5.37±2.94	≀ 10.35±5.56
test2	(E) LSA	≀ 7.62±2.61	23.73±11.97	18.47±10.58	5.01±2.62	≀ 5.39±3.05	10.04±5.43
test2	(E) Luhn	6.33±1.92	18.20±9.72	17.99±9.34	4.31±2.34	≀ 5.43±2.96	9.28±4.96
test2	(E) TextRank	≀ 6.85±2.19	≀ 19.76±10.72	17.59±9.04	4.01±2.19	5.08±2.97	8.74±4.77
test2	(E) TF-IDF	≀ 6.92±1.88	8.11±3.96	15.28±7.46	≀ 4.17±2.06	≀ 5.91±3.04	8.52±4.29
test2	(A) Roberta2Roberta	6.16±1.52	≀ 32.94±16.41	14.47±7.76	2.84±1.61	2.41±1.21	7.46±4.25
test2	(A) BERTSUM	≀ 6.80±2.18	30.13±15.25	14.40±7.35	2.37±1.23	≀ 2.65±1.67	7.28±3.97
test2	(E) LexRank	6.42±2.37	20.08±13.02	12.81±8.05	≀ 2.56±1.64	≀ 4.14±2.85	≀ 7.62±4.49
test2	(A) LED	3.62±2.53	12.70±8.84	6.14±4.08	0.74±0.64	1.79±1.20	3.87±2.65

Table 6: Automatic evaluation of abstractive (A) and extractive (E) summarization methods on the English part of ELITR Minuting Corpus. Within each test set, the methods are sorted by decreasing Rouge_1. The symbol "≀" used in other columns highlights places the order of the other metric does not correspond to that of ROUGE_1.

ing. The generated abstractive minute potentially contains new phrases and sentences that do not appear in the meeting transcript. We briefly describe the abstractive and extractive models included in this small experimental study in the following two paragraphs.

BART (Lewis et al., 2019) uses denoising autoencoder using pretraining sequence to sequence tasks for generation and understandability. BERTSUM (Liu and Lapata, 2019) uses a document-level encoder on top of BERT, which generates extractive and abstractive language learning models. BERT2BERT (Rothe et al., 2020) uses BERT checkpoints to initialize the encoder-decoder to provide a better understanding of input, mapping of input to context, and generation from context while the attention variables initialize randomly. LED (Beltagy et al., 2020) is a variant of Longformer, an adapted Transformer (Vaswani et al., 2017) model which supports sequence-to-sequence transformation for long documents. 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) is a pretrained model with a novel objective function designed for summarization by which important sentences are removed from an input document and then generated from the remaining sentences. Roberta2Roberta (Liu et al., 2019b) is an encoder-decoder model, meaning that both the encoder and the decoder are RoBERTa models. T5 (Raffel et al., 2019) is 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. BART_XSum_Samsum[8] is a BART denoising autoencoder, pretrained on XSum and further fine-tuned on Samsum dataset (Gliwa et al., 2019).

TextRank (Mihalcea and Tarau, 2004) is based on graph modeling techniques and takes each sentence of a given transcript as vertices and similarity score as edges. The top-ranked sentences are extracted to generate the minutes. LexRank (Erkan and Radev, 2004) 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. It sets 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 (Luhn, 1958) is based on the frequency of words. It is a naive approach based on TF-IDF (Christian et al., 2016) and focuses 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 (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.

[8] https://huggingface.co/Salesforce/bart-large-xsum-samsum

In Table 6, we assess the quality of the generated summaries with automatic summarization metrics like ROUGE (1, 2, L, WE) (Lin, 2004; Ng and Abrecht, 2015), BERTScore (Zhang et al., 2019) and BLEU (Papineni et al., 2002). The scores are averaged across the meetings in the given test set.
In the abstractive methods, we see that 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.

5. Conclusions and Future Work

In this paper, we present the first version of ELITR Minuting Corpus, a dataset designed to develop methods that 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 down by different human scribes. We extensively describe and analyze the annotations (minute creation) both quantitatively and qualitatively and compare them with other meeting datasets. Finally, we provide an extensive collection of summarization baseline results on the English part of our dataset. The corpus is publicly available in the Lindat repository at `http://hdl.handle.net/11234/1-4692`.

Automatic Minuting is a time-critical application of speech and language processing, and we claim that *ELITR Minuting Corpus* is a first-of-its-kind dataset to address this use case. Also, ELITR Minuting Corpus is the first meeting dataset to have instances of meetings and minutes in a 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.

6. Acknowledgment

This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460 (ELITR), the grant 19-26934X (NEUREM3) of the Czech Science Foundation, and has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2018101 LINDAT/CLARIAH-CZ.

7. Language Resource References

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

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.

Ghosal, T., Bojar, O., Singh, M., and Nedoluzhko, A. (2021). 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.

Ghosal, T., Singh, M., Nedoluzhko, A., and Bojar, O. (2022). 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.

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.

Kratochvíl, J., Polák, P., and Bojar, O. (2020). Large corpus of czech parliament plenary hearings. In Nicoletta Calzolari, et al., editors, *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 6363–6367, Marseille, France. European Language Resources Association.

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*.

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.

Macháček, D., Kratochvíl, J., Vojtěchová, T., and Bojar, O. (2019). A speech test set of practice business presentations with additional relevant texts. In Carlos Martín-Vide, et al., editors, *Statistical Language and Speech Processing*, pages 151–161, Cham. Springer International Publishing.

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.

Ng, J. and Abrecht, V. (2015). Better summarization evaluation with word embeddings for ROUGE. In Lluís Màrquez, et al., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1925–1930. The Association for Computational Linguistics.

Nguyen, T.-S., Pham, N.-Q., Stüker, S., and Waibel, A. (2020). High performance sequence-to-sequence model for streaming speech recognition. In *INTERSPEECH*.

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.

Polák, P., Singh, M., Nedoluzhko, A., and Bojar, O. (2022). ALIGNMEET: A comprehensive tool for meeting annotation, alignment, and evaluation. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC-2022)*, Marseille, France, June. European Language Resources Association (ELRA).

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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

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*.