ELITR Minuting Corpus

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

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- <laugh/>
(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. [...]

project meetings in English and Czech

120 meetings in English, 59 meetings in Czech

transcripts are segmented into utterances and diarized

manually corrected ASR transcripts, special vocal tags added

de-identified meeting transcripts and minutes

parts of minutes are aligned with corresponding parts of meeting transcripts (in ALIGNMEET tool, also presented at LREC 2022)

collected original minutes and specially created minutes, bullet points

Meeting minutes 1 *segment*

• [PROJECT9] remote presentations organization
- Discussion about the results: agreement on the pre-recorded pre-sentation 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

multiple minutes for a single meeting

Meeting minutes 2 *segment*

• 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.

Overall statistics of ELITR Minuting Corpus

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

Comparison of meeting summarization datasets

Dataset	lang.	minutes or summaries	real or acted	multiple summaries per meeting	# meetings	# hours	avg. words per transcript	avg. words per summary	avg. turns per transcript	avg. # of speakers
ELITR Minuting corpus	English	minutes	real	yes	120	117	7,066	373	727	5.9
ELITR Minuting corpus	Czech	minutes	real	yes	59	60	8,534	236	1,205	7.6
AMI	English	summaries	acted	no	137	100	6,970	179	335	4
ICSI	English	summaries	real	no	61	70	9,795	638	456	6.2

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, Ondřej Bojar
Charles University, Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

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.