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

Anonymous ACL ARR submission

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).

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

(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.

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 bridge that gap. Also, our dataset offers *automatic minuting investigations on a low-resource language* (Czech) for this problem (we are not aware of any dataset on automatic minuting or meeting summarization on languages other than English). 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 transcripts, extractive summaries (selected relevant transcript lines), and abstractive summaries in the form of coherent paragraphs. AutoMin is comparable in size to the 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

As we mention, there is a lack of a proper minuting dataset, we survey few existing datasets on meeting and dialogue summarization which seems 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 dialogue and 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 for meeting summarization experiments. The AMI Meeting corpus (Mccowan et al., 2005) contains 100 hours of meeting discussions, two-thirds of which are, however meeting 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) includes 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)) or unprocessed (Parliament and other available political meetings in the official domain).

Some recently released conversational datasets are comparatively much larger. For example, the MEDIASUM (Zhu et al., 2021) dataset includes 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 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, the collection of doctor–patients 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). 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 (our project meetings and the project meetings of our colleagues), with prevailing non-native speakers of English. The discussions in the Czech part are from computer science and public transport domains; all meeting participants are native speakers of Czech. The **length** 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 after the meeting), and one or more generated minutes (by annotators). Original minutes are missing for some meeting sessions, but each meeting must contain 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-COVID), 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 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 more than once 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

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.

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 same 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.

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 planning 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,

	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.

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 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. The de-identification was completed using our web-based tool (see Appendix A Figure 6, 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

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

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 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 discussions concerning their ongoing projects being publicly released. Second, without actually browsing the data released, the participants cannot effectively give informed consent. That's why we consider it obligatory to give all participants the possibility to check the final version of data before the release.

In the case of our dataset, although we had prior consent of all the participants, we performed the final 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 sections, we discuss the quality of minutes (Section 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 (Section 4.1), topic diversity (Section 4.2), dialogue act diversity (Section 4.3) and speaker diversity (Section 4.4).

4.1 Level of Abstractiveness

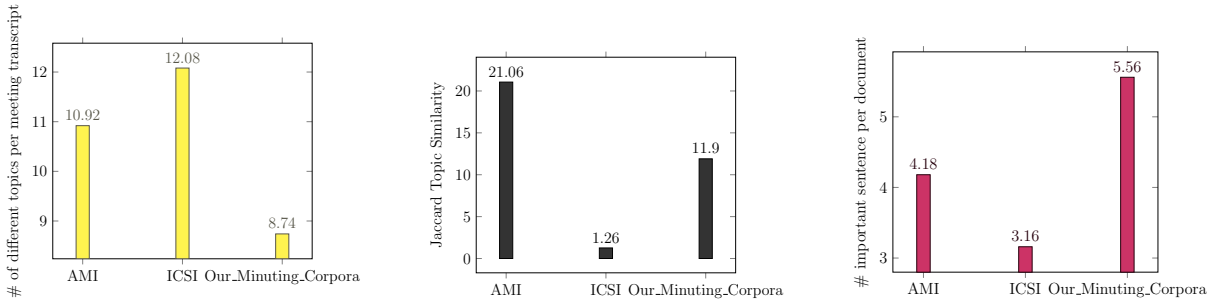
Abstractive summaries involve paraphrasing and are likely to contain words not seen in the transcript. We can thus estimate the *level of abstractiveness* simply by checking what portion of the vocabulary extracted from the minutes is covered by the wording of the transcript. For this analysis, we lemmatize words and exclude stopwords. Figure 3d 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. Figure 4 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.

Figure 2b 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 min-



(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

utes using Jaccard similarity (Niwattanakul et al., 2013). Figure 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 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

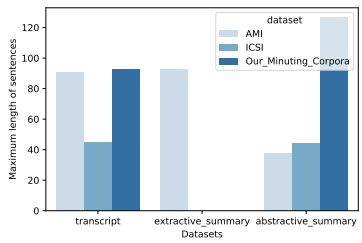
To observe the biasness in speaker diversity, we calculated the Perplexity in Figure 3e and Entropy 3f of our minuting dataset. We modeled a different number of speakers and their corresponding count of words. Further we averaged across the dataset. Next we visualize the data distribution by mapping the frequency of parameters across the entire meeting corpora. We plot the number of turns in all meeting corpora in Figure 4a and report the presence of multi-party dialogues in Figure 4b and summary tokens in Figure 4c. We also investigate whether a similar positional bias is present in multi-party dialogues in Figure 5. 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.

	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

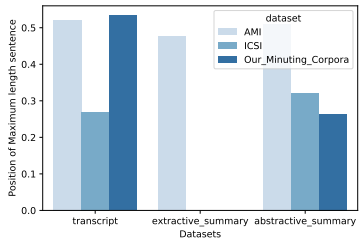
Table 3: Automatic Evaluation of Human Annotated Minutes

4.5 Data Quality

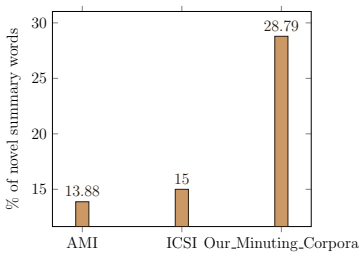
Estimating the quality of meeting minutes is a very subjective task. People differ in selecting topics which are essential and worthy to be included in the minutes, how much detailed one should be, or how to use different language expressions to describe a meeting action. For some minutes from a series 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.



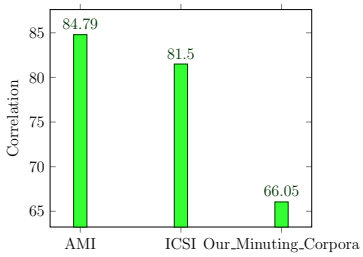
(a) Maximum sentence length over the transcript and minutes



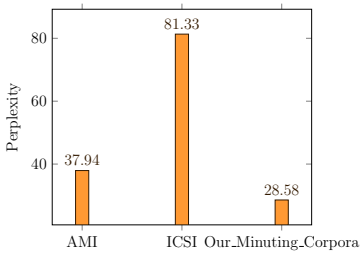
(b) Maximum sentence length position over the transcript and minute



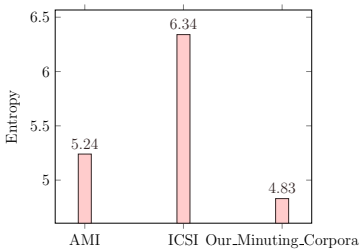
(c) Novel Summary Words



(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 (e,f) Speakers diversity mapped to observe the bias nature of minuting corpora with the help of perplexity and entropy

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.

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.

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

	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.

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.

5 Evaluation

We evaluate our minuting dataset on three possible use-cases described briefly in Appendix A.

	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.

Essentially, we consider evaluating our minuting corpora with the existing summarization models. We assess both extractive and abstractive methods of summarization (refer Appendix A.3). 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 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 Appendix A.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 Appendix C. 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 Table 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 Table 5 we see BART pretrained on XSum and fine-tuned on Samsum achieves most readable human evaluation scores.

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

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

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

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.

References

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

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2020. Multiwoz – a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling.

Susanne Burger, Victoria MacLaren, and Hua Yu. 2002. The isl meeting corpus: The impact of meeting type on speech style.

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.

Ann Clifton, Aasish Pappu, Sravana Reddy, Yongze Yu, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. The spotify podcasts dataset. *arXiv preprint arXiv:2004.04270*.

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.

Raquel Fernández, Matthew Frampton, John Dowding, Anish Adukuzhiyil, Patrick Ehlen, and Stanley Peters. 2008. Identifying relevant phrases to summarize decisions in spoken meetings. In *Ninth Annual Conference of the International Speech Communication Association*.

Matthew Frampton, Jia Huang, Trung Bui, and Stanley Peters. 2009. Real-time decision detection in multi-party dialogue. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1133–1141.

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*.

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.

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

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2020. Generating soap notes from doctor-patient conversations. *arXiv preprint arXiv:2005.01795*.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Dongyub Lee, Myeongcheol Shin, Taesun Whang, Seungwoo Cho, Byeongil Ko, Daniel Lee, Eunggyun Kim, and Jaechoon Jo. 2020. Reference and document aware semantic evaluation methods for korean language summarization. *arXiv preprint arXiv:2005.03510*.

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*.

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.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 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.

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. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

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

M. Michel, J. Ajot, and J.G. Fiscus. 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.

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.

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*.

Anna Nedoluzhko and Ondrej Bojar. 2019. Towards automatic minuting of the meetings. In *ITAT*, pages 112–119.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384.

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*.

Revanth Rameshkumar and Peter Bailey. 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.

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.

Lu Wang and Claire Cardie. 2016. Summarizing decisions in spoken meetings. *arXiv preprint arXiv:1606.07965*.

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.

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

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

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*.

A Appendices

A.1 Manual Evaluation of Human Annoated Minutes

We then thoroughly discussed the details to understand the basis of our judgment. The results show very similar assessments (e.g., all experts selected the same meetings as the best ones). We found adequacy (the judgment if summary sentences represent conclusions 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, e.g., missing relevant information, unclear extractive segments revealing no content value, misunderstanding the content, including non-relevant information, or chaotic structure. However, evaluated most (24 out of 26) minutes in the experiment as acceptable. Surprisingly, all three “winners" were minutes created by our annotators. Original minutes included too much unnecessary information or were too short.

A.2 Use cases

The primary usage of the data set consists of automatically creating minutes from multiparty meeting transcripts. Additionally, the dataset can also identify similarity between (i) given a pair of meeting transcript and minute; the task is to identify whether the minute belongs to the transcript. We found this use case challenging during our data preparation from meetings on similar topics given the similarity in various named entities. (ii) Given a pair of minutes, the task is to identify whether the two minutes belong to the same or different meetings. These use cases are essential as we want to uncover how minutes created by two different persons for the same meeting may differ in content and coverage.

Table 7: Use-case description

A:**(Generation)**	Transcript → Minute
B:**(Verification)**	Transcript + Minute → True/False (true corresponding to a pair of matching transcript and minute, and vice versa)
C:**(Comparison)**	Minute + Minute → True/False (true corresponding to a pair minutes that belong to the same transcript)

A.3 Methods

The evaluation of our novel dataset id performed on different existing baseline extractive and abstractive summarization methods. We present a brief overview of these methods in Table 8.

B Hyperparameter

The hyperparameter setting with a learning rate of 1e-5, weight decay of 0.001, max. Grad. Norm of 1.0 warmup steps of 1300 and batch size of 24 with max epochs as 4. In run-time, 1 GPU with GeForce RTX is 2080 Ti, used 11 GB GPU RAM and 248.8 machine RAM to execute examined models.

C Generated Samples

Given below is an example of minutes generated by our best model of our minuting corpora of use-case 1

DATE : 2021-07-21
ATTENDEES : PERSON4, PERSON5, PERSON8, PERSON10, PERSON13

SUMMARY-

- The deadline for the project is next Monday, June 15th.
  Someone from the project needs to be registered there.
  PERSON8 will try to register today.

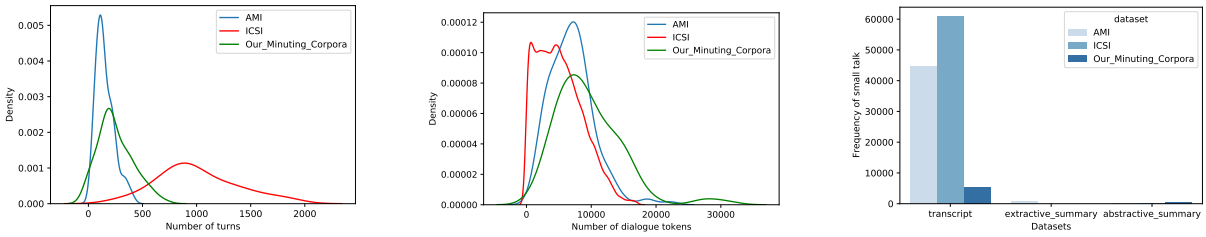
- PERSON13 is going with PERSON4 to LOCATION5.
  They have a meeting before lunch on Monday.
  They have one more paper, she wants to submit it to Archive and PROJECT8 so that someone can read it.

- PERSON10 is on holiday for next two days.
  They have written one and half paragraph of the book yesterday, and will work on the book from now on.

- PERSON4 will write half of the chapters.

- PERSON8 will organize the chapters.
  They added some information from papers.
  They will write a preface to the book.
  He needs to generate, to get the similar metrics from the PROJECT3 and the rest.

- PERSON5 is going to write his survey.
  They will work with PERSON8.



(a) Frequency of Turns (b) Frequency of dialogue tokens (c) Frequency of summary tokens

Figure 4: Speakers diverse nature with frequency of turns, dialogue and summary tokens across entire minuting dataset with their position bias

Table 8: Description of evaluation of our minuting corpus across different extractive and abstractive summarization techniques

Model	Description
	Supervised
SVM(Wang and Cardie, 2016; Fernández et al., 2008; Frampton et al., 2009)	In this paper, we use the supervised approach to determine the similarity use-case described in section 5.1. After the pre-processing, we apply a total of 6 similarity scores: cosine similarity, Rouge-1, Rouge-L, Sequence Matching, and some defined methods for cross-verification of noteworthy mentions (such as ORGANIZATION, LOCATION, PROJECT, etc.) in both - the minute and transcript, as well as checking the presence of some rarely used set of words, in both of them. A combination of these scores is used in classification to determine the best possible outcome.
	Deep Learning
BART(Lewis et al., 2019)	uses the basic seq2seq architecture with bidirectional encoder as in BERT with additional left-to-right denoising autoencoder. 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 significant performance gains when finetuned 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 to distinguish multiple sentences. These embeddings are summed and input to several bidirectional transformer layers, generating contextual vectors and further decoding. Additionally, a new finetuning schedule adopts different optimizers for the encoder and decoder to alleviate the mismatch(as the encoder is pre-trained while the 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[7] to match the pretraining 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 long former 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., 2019b)	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 pretraining 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.
BART_XSum_Samsum(Lewis et al., 2019)	introduces a denoising autoencoder for pretraining seq-2-seq tasks, which applies to both natural language understanding and generation tasks. In this work, we use pre-trained BART on XSUM and further finetune it on the SAMSUM dataset.
	Graph Modelling
TextRank(Mihalcea and Tarau, 2004)	is a text summarization technique based on a graph algorithm. The input transcript has individual sentences, each represented by 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.
	Ranking
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 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.
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. 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.
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.

- ALL are working on the papers.
  The deadline for feedback is at the end of June.
  The reviewers for PROJECT5 need to be at least a professor, but don't have to be from the university.
  The grant will be 5000 for it.
  The deadline for PROJECT7 should be in November.
  The conference will be virtualised and take

Similarity	Classifier	Accuracy	Precision	Recall	F1
Transcript-Minute	**Random Forest**	**0.91**	**0.71**	**0.62**	**0.66**
	SVM	0.88	0.65	0.40	0.49
Minute-Minute	**Random Forest**	**0.85**	**0.42**	**0.61**	**0.5**
	SVM	0.77	0.26	0.53	0.35

Table 9: Similarity use case results

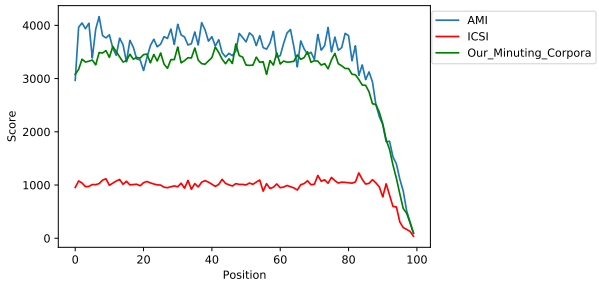


Figure 5: Position Bias

place in 2021.

- PERSON8, PERSON13, PERSON5 and PERSON10 discussed the details of the conference.
  The abstract submission is on Monday, June 15th.
  PERSON5 and PERSON8 are going to write a survey for the project. They want to introduce new people to it.

- ALL discussed about the amount of money they are getting from the university.
  The money for this year cannot be used for bonuses.
  PERSON7 bought the computer that he is now using for some grant.

- PERSON8 got a mail from PR person saying that they can come to the official event.

And, on the following page there is a true positive instance predicted by our model, for meeting similarity use-case:

Minute:A)

PROJECT3 31. 08. 2020
Attendees: PERSON1, PERSON9, PERSON2
Purpose of meeting: Preparing for the demo, choosing the right people and language combination Summary

- PERSON9 sent email to PERSON11
- PERSON1 checked PROJECT5 emails
- Discussed about the attendees during the demo
- Discussed input language
- Discussed language translation combination
- PERSON9 offered help with finding Romanian speaker
- Discussed person involved in the testing
- Discussed about date of the demo
- Discussed about a ORGANIZATION8 ASR
- Discussed about risk of Italian source
- Discussed a Session closing day date

Milestones

- PERSON8 will be person from ORGANIZATION2
- PERSON8 will be person from ORGANIZATION5
- German will be OK as input language
- PERSON1 does not have access to Romanian speaker
- PERSON1 will fill the Doodle

Minute:B)

Organizational stuff

- Monthly call will be on Thursday, 5 PM LOCATION1 time
  - At least PERSON14 and PERSON10 should take part
  - PERSON14 will care about including PERSON6 into the mailing list
- PERSON6's coming to LOCATION1

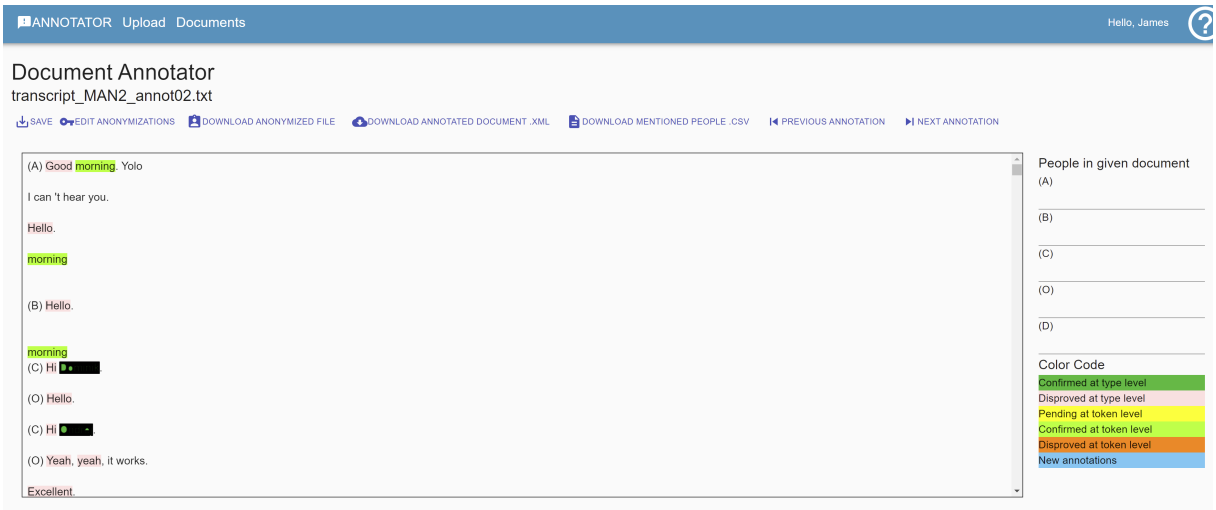


Figure 6: De-identification Toolkit

  - It is very desirable that PERSON6 comes to LOCATION1 in person
  - Visa issues due to Covid situations

PROJECT2

- PERSON10 is trying to contact ORGANIZATION5 colleagues, the communication is not completely perfect
- PERSON4 is preparing the leaflets, LOCATION1 is waiting

Progress on PROJECT6

- PERSON10 is trying the back-translation
  - It's low priority, is running on server, but may be stopped if needed.
  - No interesting results to discuss yet. Should be discussed with PERSON15 first, what to do next
  - PERSON10 may try the translations on CPUs

PROJECT4

- No special updates for now
- a related paper on BLEU that might be useful for evaluation
- Discussing metrics, using semantic metrics, different kinds of metrics
- Why do we need special metrics for MT