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

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

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

last-name@ufal.mff.cuni.cz

Abstract

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

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

1. Introduction

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

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

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

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

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

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

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

*equal contribution

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

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

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

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

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

Some unique features of AutoMin 2021 were:

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

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

2. Earlier Efforts and Related Literature

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

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

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

Our AutoMin challenge is motivated along similar lines. We present our baseline experiments using *off-the-shelf* text summarization models in [28]. We envisage AutoMin to evolve as a platform for community investigation into tasks pertaining to automatically generating minutes from multi-party meetings.

3. Task Descriptions and Evaluation Procedure

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

The provided dataset is detailed in Section 4 below.

3.1. Task A

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

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

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

to usual paragraph-like text summaries.

3.2. Task B

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

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

3.3. Task C

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

3.4. Evaluation Procedure

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

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

3.4.1. Human Evaluation of Task A

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

1. **Adequacy** assesses if the minute adequately captures the major topics discussed in the meeting, also considering coverage (all such topics reflected).
2. **Fluency** reflects if the minute consists of fluent, coherent texts and is readable to the evaluator.
3. **Grammatical Correctness** checks the level to which the minute is grammatically consistent.

In each of these criteria, the evaluators rated the minutes on a Likert Scale [29] of 1 to 5 where 1 signifies the worst and 5 signifies the best output. Furthermore, we asked the evaluators to try to assess each of these qualities as independently of the other ones as possible.

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

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

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

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

3.4.2. Automatic Evaluation of Task A

For our automatic evaluation of Task A, we relied on the widely popular text summarization metric ROUGE [30] in its three variants: ROUGE-1, ROUGE-2, ROUGE-L.

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

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

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

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

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

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

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

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

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

3.4.3. Task B and C Evaluation

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

Our F1 score is calculated as:

$$F1 = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

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

$$\text{Precision} = \frac{\text{\# Correct YES Predictions}}{\text{Total \# of YES Predictions}} \tag{4}$$

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

Our Accuracy is:

$$\text{Accuracy} = \frac{\text{\# Correctly Answered Items}}{\text{\# Total Items}} \tag{6}$$

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

4. Dataset Description

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

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

The full processing sequence was this:

1. A preliminary consent to use meeting data was negotiated with meeting organizes and meeting participants. The consent was to record and process the full meeting data at Charles University with the foreseen publication of de-identified transcripts and minutes. We had a second round of explicit consents with publication from the meeting participants which we discuss later in this section.

2. The sound recordings from the online meetings were obtained.

3. The recordings were automatically transcribed using our own Automatic Speech Recognition (ASR) systems for English [31] and Czech [32].

4. We provided our annotators with the audio recording and the raw transcripts from the ASR. It demanded some manual effort to correct the ASR-generated transcript. We asked our annotators to carry out the following tasks:

   (a) Clean and correct the raw ASR-generated transcript (e.g., spelling corrections, mispronunciations, typos, etc.).

   (b) Break the transcript into smaller segments at natural linguistic points in the speech such as sentence or phrase boundaries, speech vs. silence/pauses, or speaker change. We warned that segmentation of spontaneous speech into sentences is often difficult and we expect that different transcribes could arrive at different segmentations. We nevertheless hope that with a fixed sequence of uttered words, the different segmentations should not affect minuting very often.

   (c) Diarize the transcripts, i.e., add speakers' codes ("PERSON1" etc.) at the beginning of each speaker's utterance in round brackets.

   (d) Format the transcript according to the agreed guidelines (in short: one sentence per line, focus on recognizing the sequence of words, preserve colloquial speech and speech errors, including errors in grammar, add punctuation and letter casing).

   Sometimes it required multiple rounds of communication between meeting participants and the annotators to resolve ambiguities. Since the annotators were not part of the actual meetings, hence sometimes they missed prior context in the meetings.

5. **Creating Minutes** After the transcript was manually corrected, the next step was to create the minutes. In most cases, it was external annotators who provided minutes for the meetings. For some meetings, we also had the minutes created by some of the meeting participants; these variants of minutes were more precise in the points they mentioned but they seemed less reliable in terms of coverage and structural match with the actual meeting.

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

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

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

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

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

   Examples of the minutes from our dataset are provided in Appendix A.

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

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

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

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

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

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

6. **De-identification** To avoid any personally-identifiable information or sensitive project-specific information getting leaked out, we took care to de-identify the entire dataset. We replaced person, project, and organization mentions with identifier strings: [PERSON*number*], [ORGANIZATION*number*] and [PROJECT*number*], respectively. We kept the identifiers stable throughout our dataset, so whenever the annotators were able to establish the identity of a given person, the same identifier was used. We note that in practice, this was complicated by unclear speech, unknown spelling of various foreign names, and lack of knowledge of people's voices.

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

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

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

4.1. Training, Development and Test Sets for AutoMin

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

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

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

5. Shared Task Timeline and Procedure Overview

AutoMin followed this timeline:

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

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

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

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

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

6. Participating Teams and Approaches

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

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

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

- **Team ABC [33]** participated in all tasks (A, B, and C) for EN data. They employed a BART-based [34] minuting architecture trained on the SAMSum corpus [14] with certain pre-processing (e.g., a simple rule-based transcript segmentation) to generate the bullet-point minutes. For Task B and Task C, the authors employed a feature-based approach with Support Vector Machines and Random Forest classifiers. They used the same set of features for both tasks: cosine similarity between the vectors, ROUGE-1, ROUGE-L, Jaccard similarity, Sequence Matcher, named entity match, ratio of the most common words to the total number of unique words, etc.
- **Team Auto Minuters [35]** participated in all tasks again only for EN meetings. The authors use a pre-trained T5-base model for summarization and fine-tuned the model on the shared task dataset for the minuting Task A. For Tasks B and C, they use several similarity scores (Jaccard and cosine similarity in particular) which they use as input to a K-Nearest Neighbour (KNN) classifier.
- **Team Hitachi [36]** participated in all the tasks for both EN and CS data. They did not use the provided reference minutes for training of Task A. Instead, they topically segmented the transcript and used a BART-based summarizer trained on SAMSum dialogue summarization dataset. In addition, they applied argumentation mining techniques on the generated minutes to improve their coherence and internal structured. The authors resolve Task A in Czech cross-lingually: they use mBART [37] to translate the Czech transcripts to English, process English and then translate the generated minutes back from English to Czech.

  For Tasks B and C, team Hitachi used multiple relevance scores and trained several machine learning models such as SVM, Logistic Regression, Random Forests, and Multi-Layer Perceptron for subsequent classification. They used Optuna[6] for hyperparameter optimization to select the best model for Task B and C.
- **Team JU_PAD [38]** participated in Task A (EN). They stacked pre-trained models for extractive (TextRank [39]) and abstractive summarization (BART trained on CNN/Daily Mail [40]) to generate the minutes. They followed these steps to generate their minutes: Pre-processing (speaker identification, speaker-dialogue separation), Part-of-Speech tagging (each word is attached with its POS), Sentence/Dialogue Processing (dialog act tagging), Extractive Summarization, Abstractive Summarization, and Minute generation (generating the bullet-point minutes from flat paragraph-like summaries). They used a pre-trained a CRF-based model on Switchboard Dialog Act Corpus [41] for dialogue act tagging.
- **Team Matus_Francesco [42]** participated in Task A (EN). They base their minuting system on the PEGASUS [43] summarization model (Pre-training with Extracted Gap-sentences for Abstractive Summarization). They perform certain pre-processing steps including removal of filler words and small talk, co-reference resolution (replacing pronouns like *you*, *I* with the corresponding named-entities) and dialogue partitioning (segment the longer transcripts into shorter chunks) prior to the summarization model. The in-time submission was M/F (baseline). The authors also made two late submissions M/F (coref) and M/F (final) where they further fine-tune their model on the AutoMin dataset followed by decoder optimization and add a certain post-processing (removal of non-important and irrelevant information via TF-IDF scoring with a user-tunable threshold).
- **Team MTS [44]** made four different submissions with different approaches for Task A (EN). They used a pipeline system for speech recognition (on AMI and ICSI corpus where the audio is available) and summarization. Their four submissions made use of PreSumm (MTS (P/S)) [45], Google Text-to-Text Transfer Transformers [46] (MTS (T5)), Pegasus [43] (MTS (Pegasus)) and a customized clustering and vectorization approach (MTS (customized)), resp., to generate the minutes. The authors use off-the-shelf pre-trained transformer-based models for the summarization part. The customized approach included steps like syntactic phrase extraction, deletion of redundant words, a vectorization step in combination with TF-IDF scores and Universal Sentence Encoder [47] followed by the final clustering (Affinity Propagation clustering [48]) step.
- **Team Symantlytical [49]** participated in all tasks for EN meetings. For Task A, they made use of Generative Pretrained Transformer (GPT-2) [50] model to generate the meeting minutes. For Tasks B and C, they used sentence vector representations: BERT [51] trained on SNLI [52] and Paraphrase RoBERTa [53]) with cosine similarity. Finally they used a thresholding scheme on the similarity values to determine the classes for the two tasks (the final threshold value for both the tasks was 0.65).
- **Team Turing TESTament [54]** participated in all tasks for EN data. For Task A, the team employed a feature-based approach (sentence length, unigram frequency, presence of numerical entities, topics from LDA, proper nouns, number of affirmative utterances) with the ranker method TOPSIS[7] to extract the most significant statements from the transcripts with a rule-based heuristic, and finally simply concatenating them as minute items in the end.

  For Tasks B and C, they used sentence representations from BERT trained on the SNLI dataset [52] with cosine similarity to find the similarity of the transcripts and the minutes. Finally, they used a similarity threshold (0.75) for the classification.
- **Team UEDIN [55]** participated in Task A (EN). They developed a minuting system that combines BERT-based extractive summarization with logistic regression-based filtering and certain rule-based pre- and post-processing steps. They leveraged *lecture summarizer*[8] which was originally designed to summarize transcripts of university lectures.
- **Team Zoom [56]** participated in AutoMin, making a late submission to Task A (EN). They used the MediaSum [13] corpus to train their transformer-based summarization model SEAL [57], and fine-tuned on AutoMin, AMI, and ICSI datasets.

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

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

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

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

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

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

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

7. Evaluation

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

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

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

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

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

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

7.1. Basic Statistics

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

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

Some participant minutes were not in the form of bulleted list like we intended to have. Instead, they produced flat summaries; some even generated one long single-line summary which was difficult to interpret in manual evaluation, see Zoom or some of MTS models. Zoom and T5 by MTS have also produced by far the shortest outputs, suggesting that their Transformer-based models may suffer from the length overfitting issue [58].

7.2. Task A Manual Evaluation Results

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

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

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

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

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

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

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

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

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

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

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

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

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

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

7.3. Task A Automatic Evaluation Results

For automatic evaluation, we took the usual text summarization metric ROUGE [30] in its three variants (ROUGE-1, ROUGE-2, and ROUGE-L). As described, each meeting has multiple reference minutes to allow for at least partial reflection of the fact that minuting styles across people differ.

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

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

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

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

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

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

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

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

7.4. Correlation between Automatic and Human Evaluation

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

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

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

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

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

7.5. Additional Submissions to EN Task A

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

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

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

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

7.6. Task B Evaluation Results

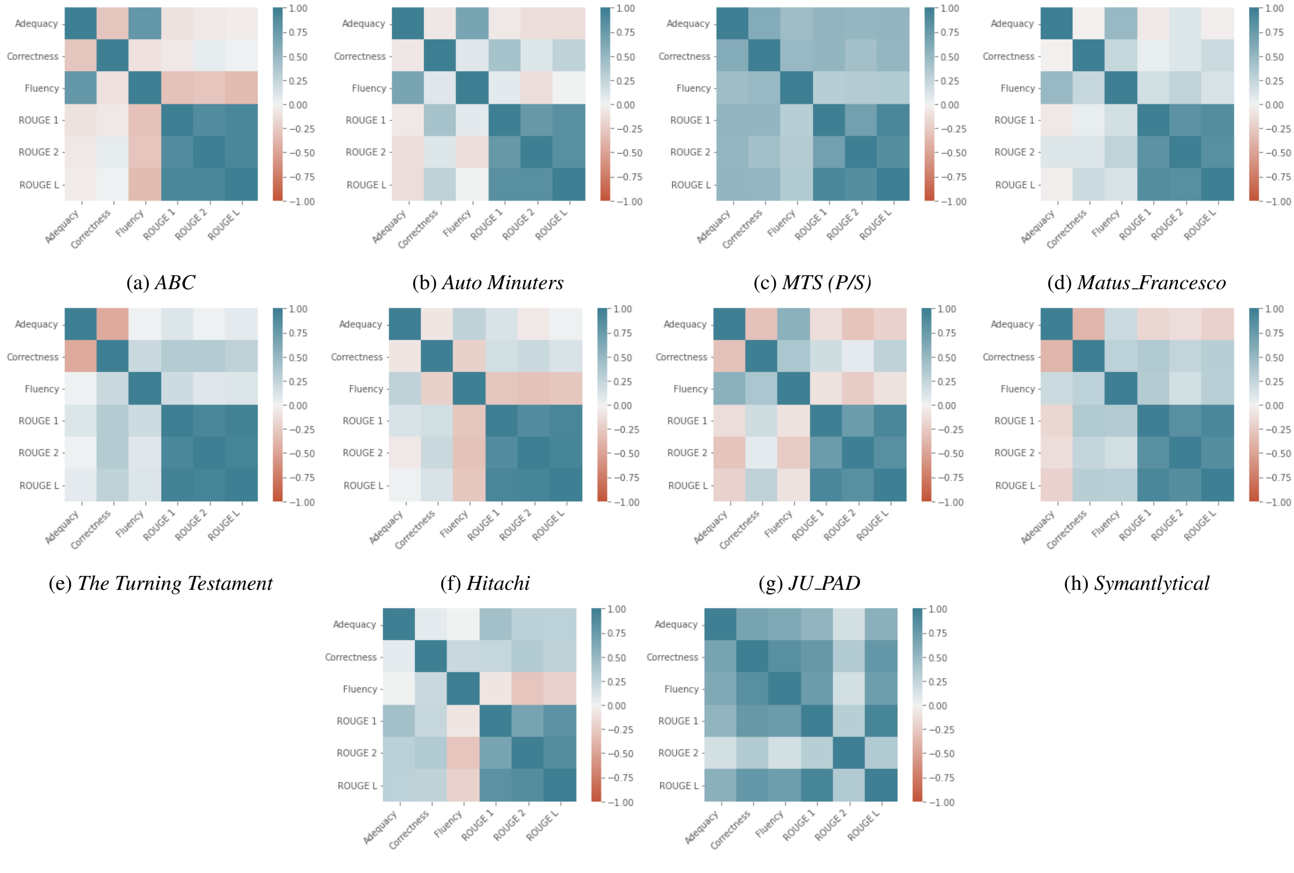
For Task B, five teams participated with their methods for the EN meetings while only Team Hitachi participated for the CS meetings.

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

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

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

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



(a) *ABC*

(b) *Auto Minuters*

(c) *MTS (P/S)*

(d) *Matus_Francesco*

(e) *The Turning Testament*

(f) *Hitachi*

(g) *JU_PAD*

(h) *Symantlytical*

(i) *UEDIN*

(j) *Zoom*†

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

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

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

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

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

Based on system description papers, we see that Team Hitachi used multiple similarity and relevance features (tf-idf, cosine similarity, named entity overlap ratio, date consistency, and BERTScore [59]) with adequate hyperparameter optimization. Team ABC, too, used several features (as mentioned in Section 6) but apparently, they missed to properly weigh the contribution of their features in the task. Other teams like Symantlytical and The Turing TESTament used manual threshold-based schemes on their features which may have resulted in their poorer performance. Team AutoMinuters used simple features like cosine and Jaccard similarity between the transcript and minute pairs and fed those to a kNN classifier. They performed second to Hitachi both in terms of Accuracy and F_1 score.

7.7. Task C Evaluation Results

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

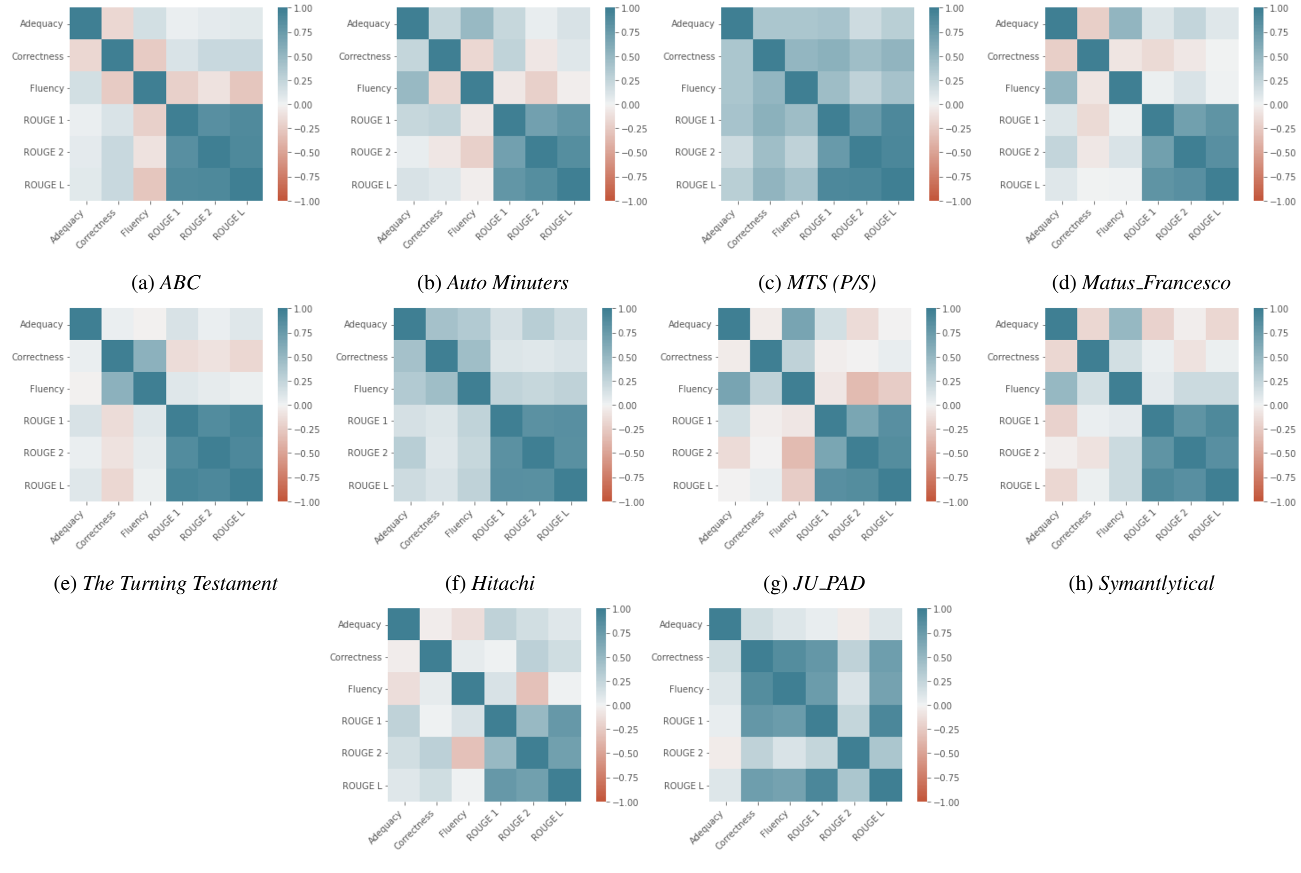
Table 9 shows the performance of the participating teams in Task C.

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

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

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

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



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

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

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

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

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

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

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

8. Findings of AutoMin

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

1. The best-performing systems in AutoMin suggest: BART-based deep neural models perform comparatively better than other transformer models to generate readable minutes.
2. As is evident from the several submissions (ABC, Hitachi, JU_PAD, Matus_Francesco, etc.), segmentation of the long meeting transcripts (either topically or via simple segmentation schemes) is crucial to the performance of the subsequent summarization modules in the proposed systems. Existing summarization models apparently have certain limitations in the number of tokens they can process to produce a good output. It would be interesting to know where the limitation effectively comes from: the inability to capture all the necessary information from a long input, the inability to produce longer output [58], or both.
3. Considering the current non-availability of large-scale domain datasets on multi-party meeting summarization (even AutoMin dataset is small-scale), the best recipe that evolved out for Task A looks like: train a deep neural model on available dialogue summarization datasets (SAMSum [14], DialSum [19], etc.) and further fine-tune it on the minuting or meeting summarization datasets (AMI [8], ICSI [9], AutoMin).
4. Simply using off-the-shelf text summarization models trained on text summarization datasets from other do-

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

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

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

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

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

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

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

9. As discussed in the previous point, minuting is a subjective activity. Different note-takers/participants would have different perspectives/expectations on what are the best possible minutes. Hence, effort towards personalized minutes generation is a worthy research direction. Generating a query-focused summary from meetings [60] is a nice example of this kind.

10. The community acknowledges a dire need for better evaluation metrics for text summarization including meeting summarization [4]. As we documented for ROUGE in our correlation results, the current automatic metrics (ROUGE, BERTScore [59], etc.) are not a good estimator of the quality of the summaries. We see here a large room for improvement from further research.

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

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

9. Conclusions and Future Plans

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

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

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

10. Acknowledgement

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

We organized AutoMin as part of our involvement in the

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

European Live Translator (ELITR) project, which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460. Ondřej Bojar would also like to acknowledge the support from the grant 19-26934X (NEUREM3) of the Czech Science Foundation.

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

11. References

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

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

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

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

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

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

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

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

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

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

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

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

[13] C. Zhu, Y. Liu, J. Mei, and M. Zeng, "MediaSum: A large-scale media interview dataset for dialogue summarization," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 5927–5934. [Online]. Available: https://aclanthology.org/2021.naacl-main.474

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[28] M. Singh, T. Ghosal, and O. Bojar, "An empirical performance analysis of state-of-the-art summarization models for automatic minuting," in *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. Shanghai, China: Association for Computational Lingustics, 11 2021, pp. 50–60. [Online]. Available: https://aclanthology.org/2021.paclic-1.6

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

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

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

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

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

[34] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

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

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

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

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

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

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

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

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

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

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

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

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

[47] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil, "Universal sentence encoder for English,"

in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 169–174. [Online]. Available: https://aclanthology.org/D18-2029

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

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

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

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

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

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

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

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

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

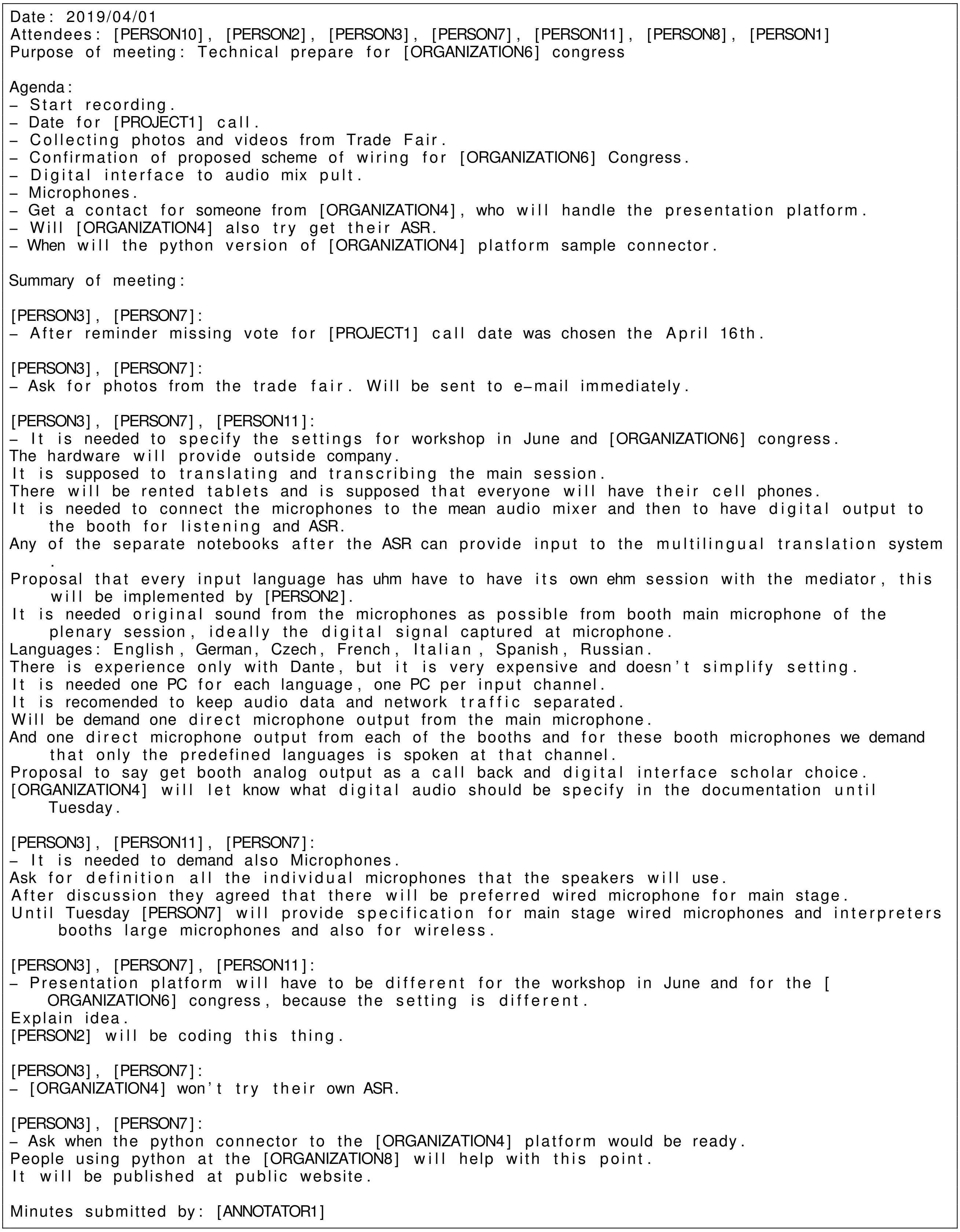
[57] Y. Zhao, M. Saleh, and P. J. Liu, "Seal: Segment-wise extractive-abstractive long-form text summarization," *ArXiv*, vol. abs/2006.10213, 2020.

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

[59] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[60] M. Zhong, D. Yin, T. Yu, A. Zaidi, M. Mutuma, R. Jha, A. H. Awadallah, A. Celikyilmaz, Y. Liu, X. Qiu, and D. Radev, "QMSum: A new benchmark for query-based multi-domain meeting summarization," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 5905–5921. [Online]. Available: https://aclanthology.org/2021.naacl-main.472

A. Sample Reference Minutes created by our Annotators



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

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

Summary of meeting:

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

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

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

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

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

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

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

Minutes submitted by: [ANNOTATOR1]
```

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

B. Sample Minutes from AutoMin participants

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

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


SUMMARY-

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


Minuted by: Team ABC
```

Figure 4: *A sample minute from Team ABC*

```
Summary:

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

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

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

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

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

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

Figure 7: *Sample Minute from Team Zoom*

C. Detailed Manual Scores of English Minutes

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

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

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

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

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

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

D. Detailed Automatic Scores of English Minutes

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

Table 13: *ROUGE-1 scores of the participants against the test set reference minutes. Team MTS submitted three runs, we enlist the best run here. M/F→Matus_Francesco. † marks a late submission.*

Teams →	ABC		Auto Minuters		Hitachi		JU_PAD		M/F (baseline)		MTS (P/S)		Symantlytical		Turing TESTament		UEDIN		Zoom†	
Test Meetings↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	0.03	0.05	0.05	0.10	0.05	0.09	0.04	0.07	0.03	0.04	0.04	0.08	0.04	0.08	0.05	0.08	0.03	0.05	0.00	0.00
en_test_002	0.07	0.12	0.05	0.09	0.07	0.13	0.04	0.07	0.03	0.08	0.04	0.09	0.03	0.06	0.06	0.12	0.02	0.06	0.01	0.03
en_test_003	0.02	0.05	0.02	0.04	0.04	0.05	0.02	0.03	0.01	0.02	0.01	0.01	0.02	0.04	0.02	0.04	0.02	0.04	0.00	0.00
en_test_004	0.02	0.03	0.01	0.01	0.01	0.02	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00
en_test_005	0.06	0.09	0.02	0.02	0.04	0.05	0.03	0.03	0.01	0.01	0.00	0.00	0.01	0.01	0.02	0.02	0.03	0.03	0.00	0.00
en_test_006	0.03	0.05	0.02	0.03	0.05	0.05	0.03	0.04	0.03	0.04	0.01	0.02	0.06	0.06	0.05	0.06	0.02	0.04	0.00	0.00
en_test_007	0.05	0.08	0.03	0.04	0.05	0.09	0.04	0.07	0.04	0.06	0.03	0.03	0.05	0.09	0.03	0.04	0.00	0.00	0.00	0.00
en_test_008	0.07	0.12	0.02	0.05	0.05	0.10	0.02	0.04	0.02	0.04	0.04	0.09	0.04	0.08	0.04	0.07	0.02	0.04	0.00	0.00
en_test_009	0.11	0.15	0.06	0.07	0.05	0.06	0.11	0.12	0.02	0.02	0.05	0.06	0.06	0.09	0.04	0.05	0.02	0.03	0.00	0.00
en_test_010	0.05	0.07	0.03	0.05	0.04	0.06	0.04	0.06	0.03	0.04	0.01	0.03	0.05	0.08	0.04	0.06	0.04	0.08	0.01	0.01
en_test_011	0.06	0.08	0.04	0.07	0.04	0.06	0.03	0.06	0.04	0.05	0.03	0.04	0.03	0.05	0.02	0.04	0.02	0.02	0.00	0.00
en_test_012	0.07	0.07	0.09	0.12	0.10	0.11	0.08	0.11	0.05	0.06	0.06	0.08	0.06	0.08	0.12	0.13	0.04	0.07	0.00	0.00
en_test_013	0.04	0.07	0.02	0.05	0.02	0.05	0.02	0.06	0.01	0.03	0.00	0.00	0.01	0.04	0.01	0.03	0.03	0.05	0.00	0.00
en_test_014	0.04	0.04	0.01	0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.04	0.01	0.01	0.02	0.03	0.03	0.03	0.00	0.00
en_test_015	0.03	0.04	0.02	0.04	0.03	0.05	0.03	0.05	0.03	0.05	0.00	0.01	0.03	0.03	0.02	0.03	0.05	0.10	0.00	0.00
en_test_016	0.12	0.19	0.06	0.09	0.08	0.16	0.05	0.10	0.05	0.07	0.04	0.06	0.06	0.12	0.08	0.16	0.03	0.04	0.00	0.00
en_test_017	0.03	0.05	0.05	0.08	0.04	0.06	0.03	0.06	0.03	0.04	0.01	0.02	0.05	0.08	0.04	0.06	0.03	0.05	0.00	0.00
en_test_018	0.06	0.08	0.03	0.06	0.06	0.09	0.06	0.08	0.04	0.06	0.01	0.02	0.06	0.08	0.05	0.08	0.03	0.05	0.01	0.02
en_test_019	0.13	0.13	0.13	0.13	0.10	0.10	0.13	0.13	0.08	0.08	0.04	0.04	0.10	0.10	0.07	0.07	0.10	0.10	0.00	0.00
en_test_020	0.05	0.05	0.02	0.02	0.06	0.06	0.02	0.02	0.03	0.03	0.00	0.00	0.02	0.02	0.04	0.04	0.04	0.04	0.01	0.01
en_test_021	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.00	0.00	0.00	0.00	0.03	0.03	0.04	0.04	0.04	0.04	0.00	0.00
en_test_022	0.07	0.07	0.01	0.01	0.05	0.05	0.05	0.05	0.04	0.04	0.00	0.00	0.03	0.03	0.02	0.02	0.08	0.08	0.00	0.00
en_test_023	0.05	0.05	0.08	0.08	0.06	0.06	0.07	0.07	0.03	0.03	0.00	0.00	0.05	0.05	0.04	0.04	0.05	0.05	0.00	0.00
en_test_024	0.08	0.08	0.04	0.04	0.08	0.08	0.07	0.07	0.06	0.06	0.00	0.00	0.04	0.04	0.07	0.07	0.11	0.11	0.00	0.00
en_test_025	0.11	0.11	0.04	0.04	0.09	0.09	0.07	0.07	0.06	0.06	0.00	0.00	0.03	0.03	0.05	0.05	0.09	0.09	0.00	0.00
en_test_026	0.13	0.13	0.08	0.08	0.13	0.13	0.05	0.05	0.06	0.06	0.03	0.03	0.09	0.09	0.12	0.12	0.03	0.03	0.00	0.00
en_test_027	0.08	0.10	0.05	0.07	0.10	0.11	0.06	0.09	0.05	0.05	0.01	0.01	0.05	0.06	0.06	0.07	0.05	0.07	0.00	0.00
en_test_028	0.11	0.13	0.06	0.09	0.09	0.11	0.05	0.07	0.08	0.11	0.00	0.00	0.04	0.04	0.07	0.09	0.06	0.07	0.00	0.01

Table 14: *ROUGE-2 scores of the participants against the test set reference minutes. Team MTS submitted three runs, we enlist the best run here. M/F→Matus_Francesco. † marks a late submission.*

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

Table 15: *ROUGE-L scores of the participants against the test set reference minutes. Team MTS submitted three runs, we enlist the best run here. M/F→Matus_Francesco. † marks a late submission.*

E. Late and Additional Submission Evaluation

Teams →	M/F (coref)†		M/F (final)†		MTS (T5)		MTS (Pegasus)		MTS (customized)	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	2.5	3	4	4	1	1	1	1	1	1
en_test_002	3	3	1.5	2	1	1	1	1	1	1
en_test_003	3	3	2	3	1	1	1	1	1.5	2
en_test_004	2.5	3	3.5	4	1	1	1	1	1.5	2
en_test_005	2	2	2.5	3	1	1	1	1	2	2
en_test_006	2.5	3	2.5	3	1	1	1	1	1	1
en_test_007	4	4	4	4	1	1	1	1	2.5	3
en_test_008	2.5	3	2.5	3	1	1	1	1	2	2
en_test_009	2.5	3	2.5	3	1.5	2	1	1	2	2
en_test_010	4	4	2.5	3	1	1	1	1	3	4
en_test_011	2	2	3.5	4	1	1	1	1	2	2
en_test_012	3.5	4	4	4	1	1	1.5	2	2	3
en_test_013	1.5	2	2.5	3	1.5	2	2	3	2	2
en_test_014	3	3	4.5	5	1.5	2	1.5	2	2.5	3
en_test_015	2.5	3	3.5	4	1	1	1.5	2	1.5	2
en_test_016	3	3	3	4	1	1	1.5	2	2.5	3
en_test_017	3	3	3.5	5	1	1	1.5	2	2	2
en_test_018	2.5	3	2	2	1.5	2	1.5	2	2	2
en_test_019	3.5	4	3	4	1.5	2	1.5	2	1.5	2
en_test_020	3	3	2.5	3	1	1	1.5	2	2	3
en_test_021	1	1	0	0	1	1	1.5	2	1.5	2
en_test_022	2.5	3	2.5	3	1	1	1.5	2	1.5	2
en_test_023	2.5	3	3.5	4	1	1	1	1	2	2
en_test_024	2	3	1.5	2	1	1	1	1	1.5	2
en_test_025	2	2	3.5	4	1	1	1	1	1.5	2
en_test_026	3	3	4	5	1	1	1	1	2	3
en_test_027	3	3	2	2	1	1	1	1	2	2
en_test_028	3	3	2.5	3	1.5	2	2	2	2.5	3

Table 16: *Adequacy scores of additional and late (†) submissions*

Teams →	M/F (coref)†		M/F (final)†		MTS (T5)		MTS (Pegasus)		MTS (customized)	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	3	4	3	3	2	2	2	2	1	1
en_test_002	3	3	1.5	2	2.5	3	2.5	3	1.5	2
en_test_003	3	3	2	3	1.5	2	4	5	1.5	2
en_test_004	2.5	3	3.5	4	3	5	2.5	4	1.5	2
en_test_005	2	2	2.5	3	1.5	2	2.5	3	2	2
en_test_006	2	2	2.5	3	2.5	4	1.5	2	2	2
en_test_007	3.5	4	3.5	4	2	3	1	1	2.5	3
en_test_008	2	2	2	2	1.5	2	1.5	2	1.5	2
en_test_009	2.5	3	2.5	3	1.5	2	1.5	2	2	2
en_test_010	4	4	3.5	4	1.5	2	2	3	3	4
en_test_011	2.5	3	4	4	1	1	1.5	2	2.5	3
en_test_012	3	3	4.5	5	1.5	2	1.5	2	2.5	3
en_test_013	2	2	2.5	3	2.5	3	2	2	2.5	3
en_test_014	3	3	4	4	3	4	1.5	2	1.5	2
en_test_015	2.5	3	3.5	4	1.5	2	1	1	1.5	2
en_test_016	2.5	3	4.5	5	1.5	2	2.5	3	2.5	3
en_test_017	2.5	3	4	5	2	2	2	2	2	2
en_test_018	2	2	3	4	1.5	2	2	2	1.5	2
en_test_019	3	3	3.5	5	1.5	2	2	2	1.5	2
en_test_020	3	4	2.5	3	2	2	1.5	2	1.5	2
en_test_021	3	3	0	0	1.5	2	2	2	1.5	2
en_test_022	3.5	4	3	4	1	1	1.5	2	1.5	2
en_test_023	2.5	3	4	4	2.5	4	1	1	2	2
en_test_024	2.5	3	2.5	3	1	1	1	1	2	2
en_test_025	2.5	3	4	5	1	1	1	1	2.5	3
en_test_026	3	3	4	5	1.5	2	2	3	2	2
en_test_027	3	3	3.5	5	1	1	1	1	2	2
en_test_028	3	3	3	4	1.5	2	2	2	2	2

Table 17: *Fluency scores of additional and late (†) submissions*

Teams →	M/F (coref)†		M/F (final)†		MTS (T5)		MTS (Pegasus)		MTS (customized)	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	3.5	4	4	4	2.5	3	2.5	3	1.5	2
en_test_002	3	3	3.5	4	3.5	4	2.5	3	2.5	3
en_test_003	3	3	3	4	2	2	3	3	2.5	3
en_test_004	3	3	3	3	2.5	3	2.5	3	2	2
en_test_005	3	4	3	4	2.5	3	3	3	3	3
en_test_006	3	3	3	3	4	4	3	3	3	3
en_test_007	3	4	3.5	4	2.5	3	2.5	4	3	3
en_test_008	3	3	3	3	2.5	4	2.5	3	2.5	3
en_test_009	3	3	3	3	3	3	3	4	2.5	3
en_test_010	3	4	4	4	3	3	3.5	4	3	3
en_test_011	3	3	3.5	4	2	3	2.5	3	2.5	3
en_test_012	3	4	3.5	4	2.5	3	2.5	3	3.5	4
en_test_013	3	3	3	3	3.5	4	3	3	3	3
en_test_014	3.5	4	3.5	5	3.5	4	2.5	3	1.5	2
en_test_015	3.5	4	3.5	4	1	1	1	1	1.5	2
en_test_016	3.5	4	5	5	2	2	3	3	2	2
en_test_017	3.5	4	4	5	2	2	2	2	2	2
en_test_018	3	3	4	4	3.5	4	3	3	2	2
en_test_019	3	3	4	5	3	4	2	2	1	1
en_test_020	3.5	4	3.5	4	3	3	3	3	2.5	3
en_test_021	4	4	0	0	3	3	2.5	3	2.5	3
en_test_022	3.5	4	4	5	2	2	2	2	2.5	3
en_test_023	3.5	4	3.5	4	3.5	4	2.5	3	2.5	3
en_test_024	2.5	3	3	3	1.5	2	1.5	2	1.5	2
en_test_025	3	3	4.5	5	2	3	2	2	2.5	3
en_test_026	3.5	4	4	5	2.5	3	3.5	4	2	2
en_test_027	3.5	4	4.5	5	1	1	3.5	4	2	2
en_test_028	2.5	3	4	4	2.5	3	3	3	2	2

Table 18: *Grammatical correctness scores of additional and late (†) submissions*

Teams →	M/F (coref)†		M/F (final)†		MTS (T5)		MTS (Pegasus)		MTS (customized)	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	0.19	0.25	0.15	0.20	0.01	0.02	0.01	0.03	0.13	0.18
en_test_002	0.22	0.33	0.12	0.17	0.06	0.12	0.10	0.23	0.07	0.13
en_test_003	0.19	0.27	0.17	0.24	0.05	0.07	0.04	0.08	0.12	0.15
en_test_004	0.08	0.10	0.10	0.13	0.05	0.07	0.07	0.11	0.13	0.15
en_test_005	0.14	0.17	0.13	0.16	0.01	0.01	0.05	0.08	0.15	0.16
en_test_006	0.22	0.27	0.24	0.28	0.11	0.15	0.11	0.13	0.15	0.18
en_test_007	0.21	0.30	0.16	0.26	0.08	0.11	0.06	0.08	0.19	0.24
en_test_008	0.17	0.26	0.13	0.20	0.07	0.11	0.09	0.13	0.13	0.19
en_test_009	0.22	0.22	0.17	0.19	0.13	0.16	0.13	0.14	0.14	0.16
en_test_010	0.23	0.29	0.22	0.26	0.05	0.07	0.11	0.12	0.24	0.27
en_test_011	0.17	0.18	0.26	0.28	0.02	0.02	0.04	0.05	0.22	0.23
en_test_012	0.34	0.37	0.22	0.22	0.04	0.04	0.09	0.11	0.20	0.21
en_test_013	0.13	0.17	0.15	0.27	0.04	0.04	0.06	0.09	0.09	0.16
en_test_014	0.17	0.24	0.12	0.17	0.07	0.08	0.07	0.09	0.15	0.17
en_test_015	0.12	0.13	0.24	0.28	0.08	0.10	0.06	0.07	0.17	0.19
en_test_016	0.23	0.33	0.26	0.30	0.05	0.09	0.02	0.03	0.22	0.24
en_test_017	0.18	0.29	0.20	0.25	0.02	0.04	0.08	0.10	0.17	0.19
en_test_018	0.25	0.29	0.16	0.21	0.06	0.08	0.12	0.18	0.21	0.23
en_test_019	0.33	0.33	0.24	0.24	0.05	0.05	0.04	0.04	0.25	0.25
en_test_020	0.29	0.29	0.15	0.15	0.01	0.01	0.07	0.07	0.21	0.21
en_test_021	0.07	0.07	0.00	0.00	0.05	0.05	0.08	0.08	0.23	0.23
en_test_022	0.17	0.17	0.19	0.19	0.03	0.03	0.04	0.04	0.21	0.21
en_test_023	0.19	0.19	0.27	0.27	0.02	0.02	0.01	0.01	0.20	0.20
en_test_024	0.28	0.28	0.16	0.16	0.01	0.01	0.04	0.04	0.23	0.23
en_test_025	0.32	0.32	0.21	0.21	0.04	0.04	0.04	0.04	0.21	0.21
en_test_026	0.30	0.30	0.24	0.24	0.04	0.04	0.06	0.06	0.21	0.21
en_test_027	0.30	0.32	0.10	0.10	0.00	0.00	0.02	0.02	0.17	0.17
en_test_028	0.30	0.36	0.16	0.20	0.03	0.06	0.07	0.10	0.21	0.26

Table 19: *ROUGE-1 scores of additional and late (†) submissions*

Teams →	M/F (coref)†		M/F (final)†		MTS (T5)		MTS (Pegasus)		MTS (customized)	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	0.03	0.04	0.03	0.03	0.00	0.00	0.00	0.00	0.01	0.03
en_test_002	0.05	0.11	0.04	0.08	0.00	0.02	0.02	0.06	0.01	0.04
en_test_003	0.02	0.03	0.03	0.05	0.00	0.00	0.00	0.00	0.01	0.03
en_test_004	0.01	0.02	0.01	0.02	0.00	0.00	0.00	0.00	0.03	0.05
en_test_005	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.02	0.03	0.04
en_test_006	0.03	0.05	0.04	0.06	0.03	0.06	0.01	0.02	0.03	0.04
en_test_007	0.05	0.08	0.04	0.05	0.01	0.03	0.01	0.02	0.03	0.05
en_test_008	0.05	0.10	0.05	0.10	0.00	0.00	0.00	0.00	0.01	0.02
en_test_009	0.05	0.05	0.05	0.06	0.03	0.04	0.02	0.04	0.02	0.03
en_test_010	0.04	0.06	0.05	0.09	0.00	0.00	0.01	0.02	0.03	0.06
en_test_011	0.02	0.03	0.05	0.07	0.00	0.00	0.00	0.00	0.04	0.06
en_test_012	0.10	0.12	0.05	0.05	0.00	0.00	0.01	0.03	0.06	0.06
en_test_013	0.02	0.03	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.01
en_test_014	0.01	0.02	0.02	0.03	0.02	0.03	0.02	0.02	0.02	0.03
en_test_015	0.01	0.02	0.06	0.11	0.00	0.00	0.00	0.00	0.02	0.03
en_test_016	0.05	0.09	0.04	0.06	0.00	0.00	0.00	0.01	0.03	0.03
en_test_017	0.04	0.05	0.04	0.08	0.01	0.02	0.02	0.02	0.03	0.04
en_test_018	0.05	0.08	0.04	0.07	0.00	0.00	0.00	0.01	0.04	0.07
en_test_019	0.12	0.12	0.07	0.07	0.01	0.01	0.00	0.00	0.08	0.08
en_test_020	0.04	0.04	0.01	0.01	0.00	0.00	0.02	0.02	0.03	0.03
en_test_021	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.06	0.06
en_test_022	0.05	0.05	0.09	0.09	0.00	0.00	0.00	0.00	0.08	0.08
en_test_023	0.01	0.01	0.08	0.08	0.00	0.00	0.00	0.00	0.02	0.02
en_test_024	0.05	0.05	0.02	0.02	0.00	0.00	0.00	0.00	0.10	0.10
en_test_025	0.07	0.07	0.02	0.02	0.00	0.00	0.00	0.00	0.03	0.03
en_test_026	0.05	0.05	0.04	0.04	0.01	0.01	0.02	0.02	0.04	0.04
en_test_027	0.05	0.06	0.03	0.05	0.00	0.00	0.00	0.00	0.02	0.03
en_test_028	0.09	0.11	0.02	0.03	0.00	0.00	0.03	0.05	0.06	0.09

Table 20: *ROUGE-2 scores of additional and late (†) submissions*

Teams →	M/F (coref)†		M/F (final)†		MTS (T5)		MTS (Pegasus)		MTS (customized)	
Test Meetings ↓	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
en_test_001	0.10	0.13	0.10	0.13	0.01	0.02	0.01	0.02	0.06	0.09
en_test_002	0.14	0.22	0.08	0.13	0.05	0.09	0.07	0.15	0.05	0.08
en_test_003	0.10	0.12	0.11	0.13	0.03	0.03	0.02	0.03	0.07	0.09
en_test_004	0.05	0.07	0.06	0.07	0.04	0.05	0.05	0.07	0.08	0.08
en_test_005	0.07	0.09	0.07	0.08	0.01	0.01	0.05	0.08	0.08	0.09
en_test_006	0.11	0.14	0.12	0.13	0.09	0.13	0.09	0.13	0.09	0.09
en_test_007	0.11	0.14	0.08	0.10	0.06	0.09	0.06	0.08	0.08	0.09
en_test_008	0.11	0.18	0.10	0.17	0.05	0.08	0.07	0.10	0.06	0.09
en_test_009	0.13	0.16	0.11	0.13	0.08	0.08	0.07	0.08	0.08	0.10
en_test_010	0.13	0.15	0.11	0.15	0.03	0.05	0.07	0.10	0.10	0.11
en_test_011	0.11	0.11	0.14	0.16	0.02	0.02	0.03	0.04	0.10	0.13
en_test_012	0.21	0.26	0.14	0.17	0.03	0.03	0.05	0.06	0.12	0.12
en_test_013	0.10	0.13	0.07	0.13	0.03	0.04	0.04	0.07	0.05	0.08
en_test_014	0.08	0.10	0.06	0.07	0.07	0.08	0.05	0.07	0.08	0.09
en_test_015	0.07	0.08	0.14	0.18	0.07	0.10	0.05	0.07	0.10	0.12
en_test_016	0.11	0.17	0.10	0.13	0.04	0.07	0.02	0.03	0.09	0.11
en_test_017	0.09	0.12	0.11	0.16	0.02	0.04	0.07	0.10	0.11	0.16
en_test_018	0.14	0.19	0.10	0.12	0.04	0.04	0.07	0.11	0.11	0.15
en_test_019	0.20	0.20	0.14	0.14	0.04	0.04	0.04	0.04	0.16	0.16
en_test_020	0.12	0.12	0.07	0.07	0.01	0.01	0.06	0.06	0.10	0.10
en_test_021	0.07	0.07	0.00	0.00	0.05	0.05	0.06	0.06	0.15	0.15
en_test_022	0.12	0.12	0.17	0.17	0.03	0.03	0.04	0.04	0.13	0.13
en_test_023	0.09	0.09	0.15	0.15	0.01	0.01	0.01	0.01	0.09	0.09
en_test_024	0.12	0.12	0.07	0.07	0.01	0.01	0.02	0.02	0.18	0.18
en_test_025	0.17	0.17	0.07	0.07	0.03	0.03	0.03	0.03	0.12	0.12
en_test_026	0.16	0.16	0.12	0.12	0.02	0.02	0.04	0.04	0.10	0.10
en_test_027	0.14	0.18	0.07	0.08	0.00	0.00	0.02	0.02	0.09	0.10
en_test_028	0.15	0.19	0.07	0.08	0.01	0.02	0.06	0.07	0.11	0.15

Table 21: *ROUGE-L scores of additional and late (†) submissions*