

# Team Darbarer @ AutoMin2023: Transcription simplification for concise minute generation from multi-party conversations

Ismaël Rousseau, Loïc Fosse, Youness Dkhissi, Géraldine Damnati, Gwénolé Lecorvé

Orange Innovation, DATA&AI, Lannion, France

first.last@orange.com

## Abstract

This document reports the approach of our team Darbarer for the main task (Task A) of the AutoMin 2023 challenge. Our system is composed of four main modules. The first module relies on a text simplification model aiming at standardizing the utterances of the conversation and compressing the input in order to focus on informative content. The second module handles summarization by employing a straightforward segmentation strategy and a fine-tuned BART-based generative model. Then a titling module has been trained in order to propose a short description of each summarized block. Lastly, we apply a post-processing step aimed at enhancing readability through specific formatting rules. Our contributions lie in the first, third and last steps. Our system generates precise and concise minutes. We provide a detailed description of our modules, discuss the difficulty of evaluating their impact and propose an analysis of observed errors in our generated minutes.

## 1 Introduction

The COVID-19 pandemic has led to substantial changes in our way of communicating, interacting and collaborating. As the virus has required social distancing measures and the implementation of remote working across many industries, communication methods have shifted from traditional face-to-face interactions to virtual platforms. Consequently, the reliance on digital tools and technologies has grown exponentially, altering not only the nature of our conversations but also the means by which they are documented and managed. In this paper, we introduce a novel approach to automatic minuting tools tailored to address the unique challenges of online communication. We submitted this system for Task A of the AutoMin2023 challenge (Ghosal et al., 2023). The primary objective

of this task is to develop an automated system capable of generating minutes from multiparty meeting transcripts. The performance of the resulting summaries are to be assessed using a combination of automatic and manual evaluation metrics.

For this system, we only used the task training data as well as the additional data that was recommended. We did not use Large Language Models nor any additional training data, which positions our submission in the *constraint* category. Instead, we used “classical” language models derived from BART. While there is no strict parameter count that officially defines if a language model is “large”, at the time this paper was written, the consensus seems to be that any model exceeding 1 billion parameters with the capacity to be prompted qualifies as such in the work of Zhao et al. (2023). However, BART does not meet these criteria. We first describe in Section 2 the data provided for the AutoMin Shared Task, being the ELITR and the EuroParlMin Corpus. Then we provide related work in 3 before describing in Section 4 the different modules of our system. Finally, we provide in Section 5 insights on the results by analyzing the effect of each module on the metrics and detailing the different errors we’ve encountered in the generated minutes.

## 2 Presentation of the data

### 2.1 ELITR Minuting Corpus

The ELITR Minuting Corpus presented by Nedoluzhko et al. (2022) is a dataset containing de-identified transcripts of project meetings and their corresponding minutes, primarily focusing on the computer science domain. The Corpus contains meetings in English and meetings in Czech. The English part of the dataset predominantly includes discussions among computer science professionals, while the Czech portion encompasses deliberations

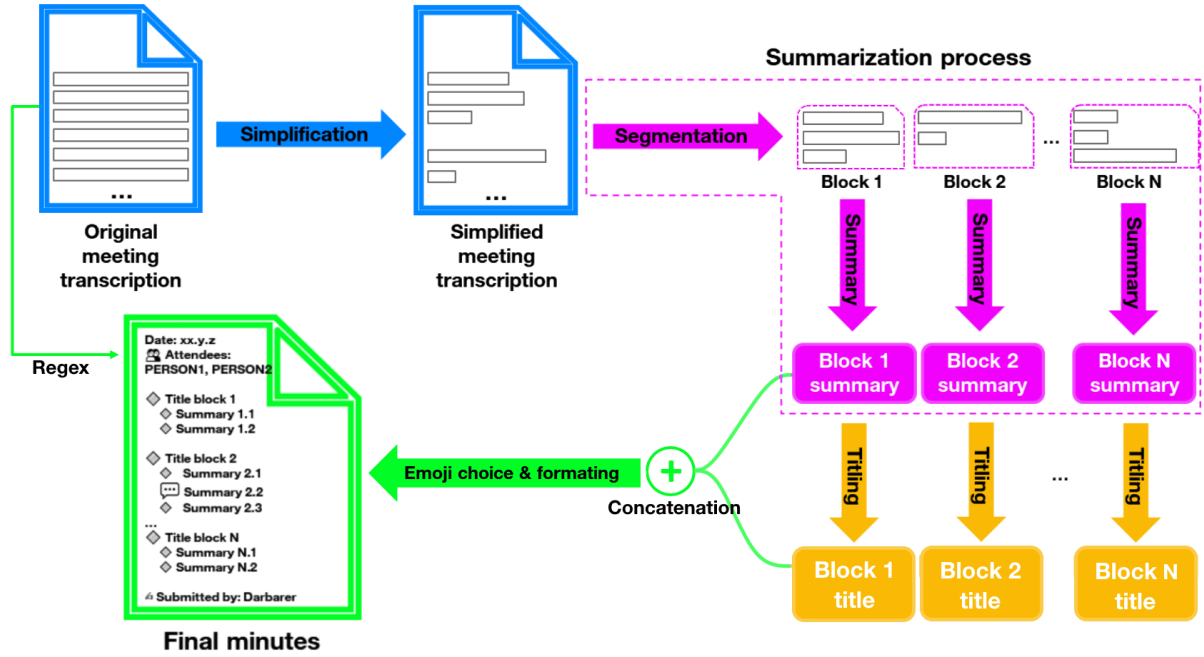


Figure 1: Processing chain (pipeline) for the automatic minute generation

from both the computer science and public administration fields.

The duration of the meetings captured in the dataset ranges from brief 10-minute exchanges to extensive discussions lasting over 2 hours.

One of the distinguishing features of the ELITR Minuting Corpus is the presence of multiple minutes files for a single conversation, thus offering a nuanced perspective on the variations in the interpretation and representation of meeting discussions. In addition, the Corpus includes, for some of the meetings, alignment files that facilitate the correlation between individual lines from the minutes files and the corresponding sections of the transcript files.

The minutes in the datasets are authored by various human annotators, each possessing distinct styles and perspectives on effective summarization. Consequently, the minutes exhibit substantial disparities in length, content, and organization. They may adopt flat or hierarchical structures, feature long sentences or keywords, and be arranged chronologically or by topic. These variations highlight the diverse approaches to summarization and offer a rich resource for studying the nuances of human-generated summaries. Table 1 shows statistics on the data and illustrates the disparities just mentioned earlier. The line compression ratio is the ratio between the number of lines in the transcripts and the number of lines in the annotated minutes.

Statistics	Mean	Std
nb. minutes per transcript	2.10	1.33
line compression ratio	12.15	21.06
nb. words per minute line	11.6	9.20

Table 1: Statistics of the ELITR Corpus

## 2.2 EuroParlMin Corpus

The EuroParlMin dataset is a subset of the broader EuroParl Corpus presented by Koehn (2005), focusing exclusively on English transcripts of European Parliament sessions from 2004 to 2011. Some sessions are split into chapters, and in that case there is one transcript file per chapter. The durations are not provided.

Unlike ELITR Minuting Corpus, each chapter of the sessions has only one associated minute. This reduces disparities in length between the minutes.

Statistics	Mean	Std
line compression ratio	6.62	13.05
nb. words per minute line	6.85	6.48

Table 2: Statistics of the EuroParlmin Corpus

## 3 Related work

Ghosal et al. (2021) give an overview of the systems submitted for the same task in the previous edition. Two systems stand out: that of the ABC

team (Shinde et al., 2021) and that of the Hitachi team (Yamaguchi et al., 2021). Both systems share a number of common features. Firstly, both teams have sought to partition the conversation in such a way that each part can fit the input of a transformer-based summarization module. Once each segment has been summarized, a concatenation is performed between the different summaries to obtain a global summary of the conversation.

In the case of the ABC team, conversation segmentation is carried out with a linear segmentation, cutting the conversation into blocks of tokens of uniform size. This segmentation is itself preceded by a rule-based block to remove redundant/repetitive elements.

In the case of the Hitachi team, segmentation is carried out automatically using a LongFormer (Beltagy et al., 2020) model, in order to select and group passages of interest in the conversation. The authors used manual annotation to train their segmentation method.

Our method is essentially based on these main steps (segmentation, summarization, concatenation), but we propose to add a text simplification module before segmentation in order to compress the text and increase the amount of information present in each segment.

## 4 Overview of the submitted system

This Section presents the Darbarer system<sup>1</sup> submitted for AutoMin2023 Task A, in the constraint category. The overall system is illustrated in Figure 1. Our system is composed of four main modules. The first module relies on a text simplification model aiming at standardizing the utterances of the conversation and compressing the input in order to focus on informative content. The second module handles summarization by employing a straightforward segmentation strategy and a fine-tuned BART-based generative model. Then a titling module has been trained in order to propose a short description of each summarized block. Lastly, we apply a post-processing step aimed at enhancing readability through specific formatting rules.

### 4.1 Transcription simplification

A conversation involves several people exchanging information about one or multiple topics. Each

person communicates in a manner that may vary significantly from one to another. This heterogeneity, notably put forward by Schiffrin (1990) can give rise to difficulties when trying to summarize spoken conversations. Additionally, disfluencies that are inherent to spontaneous speech, as well as discourse markers that help the intelligibility of speech in an interaction context, yield uninformative content in transcriptions that harms readability. Small talks can also be present and should not be transposed in the minutes. Even if the readability of the input transcript is not necessarily linked to the performance of a summarization model, it seems preferable to remove as much uninformative content as possible prior to performing the summarization task. Text simplification makes it possible to reduce the linguistic complexity of statements in a conversation and thus, in a way, allows to standardize each utterance. Text simplification has been studied initially as a way to increase accessibility of texts (for language learners for instance) and has also shown a real utility in many tasks in automatic language processing, including in summarization tasks. Silveira and Branco (2012) put forward that the simplification allows, in a summarization task, to more easily bring out the important information. In the same way, Hasler et al. (2017) shows that in translation tasks, text simplification has a positive effect. Simplification is also used for caption and subtitles generation (Buet and Yvon, 2021) as a way to compress information. It has been studied a lot in the biomedical domain as a way to increase accessibility of biomedical texts (Cardon and Grabar, 2020; Ondov et al., 2022).

In order to perform this text simplification, we first selected 86 utterances from a single conversation from the training set and manually annotated simplified version of each utterance, as can be seen in table 3. We then fine-tuned a mBART-based model named mBarthez (Kamal Eddine et al., 2021).<sup>2</sup> mBarthez, like mBART, is pretrained on denoising tasks on multilingual data. Preliminary experiments run on in-house data, that are not reported here, showed that mBarthez performed better than mBart on the simplification task. The model was trained with the following hyperparameters : a learning rate of 3e-6, a batch size of 3, 1200 for max-steps and 0 warmup steps.

Table 4 shows a few examples of the simplifi-

<sup>1</sup>The code and data specifically annotated to train some of the modules is available at <https://github.com/Orange-OpenSource/automin2023-darbarer>

<sup>2</sup>Model available on HuggingFace : <https://huggingface.co/moussaKam/mbarthez>

Original transcript	Human-written simplification
Well today (unintelligible) should be pretty simple, simple, simple meeting.	Today should be a simple meeting
And so we have collected all of them, and we are going to reveal presentation platform.	We are going to reveal the presentation platform
So you can you mute when you have some interview.	Can you mute?
Yeah, okay, yes, great.	Yes.

Table 3: Sample of the training data for text simplification

Original transcript	Automatic Simplification
Yeah, ye- ye-.	
But it's just to-	But it's just to-
To- I don't know, maybe writing things, we can think about uh alternatives or limitations.	We can think about alternatives or limitations.
And so it will be interesting to, to check this kind of thing so-	It will be interesting to check this kind of thing so-
Um so far I- I just organize the- the sections in bullet points.	I just organize the sections in bullet points.

Table 4: Examples of conversation lines before and after simplification

cation task, performed with our model. Note that the simplification model can yield empty outputs and thus can allow us to remove some lines of the transcriptions which are carrying little information.

It is noteworthy and surprising to observe that a relatively small number of training examples sufficed in achieving acceptable results for text simplification. We conducted an ablation study to further investigate the impact of the volume of the training data on the results. This analysis involved iteratively training the model with diminishing volumes of data and subsequently observing the variations in inference outcomes as well as System Output Against References and Input Sentence (SARI) scores (Xu et al., 2016). The test set consists of 20 utterances not present in the training set, as well as their corresponding human-written simplification. SARI is a metric specifically made to evaluate simplification results. It compares the system output not only against a reference (human simplified sentence), but also against the original complex sentence. This approach helps assess whether the system correctly keeps, adds, and deletes information. By looking at the inference results, we see that from [X] to [Y] examples, the model learns which sentences it should keep or not, but keeps the output sentence the exact same as the input. It is only after [Y] examples that the model starts to delete parts of the sentence that are not considered

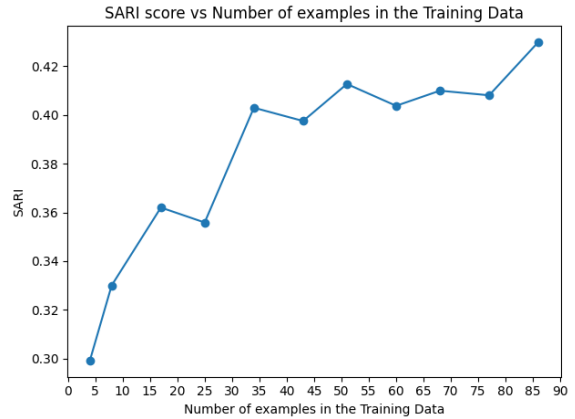


Figure 2: Evolution of the SARI score as we increase the number of simplification training examples

relevant. As seen in 2, we observe that the SARI score keeps rising as we increase the amount of examples. It does not seem to reach a plateau yet and giving the model more training examples might improve our results.

## 4.2 Summarization

The summarization module is the main component of our processing chain. For this task we decided to use the checkpoint of the BART model presented by Lewis et al. (2020). This model is trained on the XSum (Narayan et al., 2018) dataset which consists of short summaries of BBC articles and on the

SAMSum dataset (Gliwa et al., 2019) which is composed of conversations summaries.<sup>3</sup> This checkpoint showed interesting performances in the summarization task, especially on the ELITR dataset as shown by Nedoluzhko et al. (2022).<sup>4</sup> Unfortunately a major limitation of this model is the size of the text that it can take as input, which is currently limited to 1024 tokens. A naive way to deal with this, is to split the conversations into blocks of size 1024 and then summarize each block. This way of proceeding is sub-optimal since this segmentation can clearly cut the conversation in the middle of a topic and thus induce an important loss of information. We tested several more sophisticated methods, including clustering transcript lines in order to take into account the non linear nature of meeting topics. None of them improved the ROUGE score and the coherence of the resulting blocks was also perceptually degraded. We kept the fixed 1024 token segmentation for our system, but we believe that this should be further investigated.

Note that with the previous simplification step in place, blocks of 1024 tokens can now hold more information. Consequently, the average number of blocks per meeting has decreased.

To enhance the readability of the minutes, every summarized block undergoes post-processing steps which include titling and general formatting.

### 4.3 Summary block titling

Titles are a simple way to improve readability and overall comprehension in a document by providing a piece of context to the reader before the main content, as shown by Alba et al. (1981).

Thus, we once again fine-tuned mBarthez (Kamal Eddine et al., 2021), this time for the task of title generation, with the same hyper-parameters as for the text simplification. In order to achieve this, we specifically wrote relevant titles for 65 summarized blocks that were previously generated, as shown in Table 5 and used this annotated dataset to fine-tune mBarthez.<sup>5</sup>

Table 6 displays a few examples of the titles generated during inference for the ELITR dev-09 conversation using this model.

<sup>3</sup>*Ibid*: lidiya/bart-large-xsum-samsum

<sup>4</sup>particularly, see Table 6 in (Nedoluzhko et al., 2022)





<sup>5</sup>The spelling and capitalization errors of “PERSON” expressions occurred during generation and are explained in 5.2.2

### 4.4 Final formatting

Our objective is to produce meeting minutes that are neatly organized into blocks, where each block is defined by a specific title and comprises a list of indented bullet points. To achieve this, we rely on the prior stages of summarization and titling. In our setup, each sentence from a summary becomes a separate bullet point. This makes the information easier to break down and understand.

To improve readability even further, we have incorporated specific rules to generate emojis for each bullet point. This incorporation of visual cues is an additional step in our strategy to enhance minutes readability.

To generate the convenient emojis for each bullet point line, we defined a set of rules:

- If the line contains the word “date”, “deadline”, “afternoon”, “tomorrow”, “yesterday” or a day of the week, the emoji generated for this bullet point is the calendar emoji .
- If there is a discussion between some persons in the line (triggered by the verbs “discuss” or “talk”), the emoji associated to this bullet point will be the discussion balloon emoji .
- For the lines that evoke a deadline or the existence of some warnings in something (triggered by the words “deadline”, “warning” and “careful”), we add at the end of the bullet point the warning sign emoji .
- If the bullet point where there is a task still not complete or wait another task to be done (triggered by the words “still” and “wait”), we add at the end the hourglass emoji .

We also add a header containing the date and the attendees of the meeting using simple regular expressions on the transcript. Plus the signature at the end of the minute. Adding the header has an impact on the ROUGE scoring, as will be seen in Table 8 whereas the additional stylistic adjustments are not taken into account by the scoring methodology. We believe however that the latter may increase the *fluency* criterion during the human evaluation.

### 4.5 From English to Czech

All the models we’ve used thus far have been specifically fine-tuned on English corpora. The issue at hand now is the application of our method to the



Bullet points	Human-written title
PERSON7, PERSON8, PERSON9 and PERSON4 had a call last week. They will have to provide at least some prototype for the n-best list navigation and they will try to implement it into the final product.	Provision of navigation list
PERSON8 wants to have a single module that can have all the functionality of the browser translator.	Browser translator module

Table 5: Sample of the training data for title generation

Bullet points	Generated Title
PERSON6 is collecting data. He sends bad transcripts with bad quality to the annotator and asks him to correct them, then he sends it to Person6 via FileSender. Person6 sends him the pre-processed automatic speech reconstructed transcript.	Correction of transcripts
PERSON7 wants to know how the link works It is the same link as the one in the same meeting invite People can use it for all of their meetings The meeting is free for one hour, but they have to pay for it for the next month or so.	Working on the link

Table 6: Examples of section titles generated using the segment bullet points.

Czech transcriptions of the ELITR dataset. In order to re-use the same pipeline, we add two translation blocks. A first block that translates the transcriptions from Czech to English. We then generate our minutes (in English) with our processing chain, to finally translate back from English to Czech. For this purpose we use the (Tiedemann and Thottingal, 2020) models which offer the possibility to translate in both directions.<sup>6</sup> We did not perform any particular fine-tuning for this translation task.

## 5 Results

The AutoMin 2023 challenge provided three test sets: `elmiCS` and `elmiEN` for ELITR Meeting in Czech and English respectively and `europarl` in English. Full results and details about the evaluation process are provided in Ghosal et al. (2023). We obtain a ROUGE-1 score of 0.31 on `elmiCS`, 0.39 on `elmiEN` and 0.27 on `europarl`. Manual evaluation has been produced with the ALIGN-MEET tool (Polák et al., 2022), focusing on adequacy, grammaticality, fluency, relevance and at two different levels of granularity : at the document-level and the hunk-level (a hunk is defined as a set of dialog acts belonging to a summary point). Table 7 shows the results for of our system according to human annotators. Examples of generated minutes from the `test` partition are provided in Appendix. In this Section we provide additional objective evaluations and insights on observed er-

rors on the initially provided test datasets (`test` and `test2`).

### 5.1 Ablation studies

In order to evaluate the impact of each module, we use several metrics as can be seen in Table 8, with *Darbarer* being the final system we used to submit our minutes for the task. The baseline system applies the `bart-large-xsum-samsum` model on fixed blocks of 1024 tokens, without any pre-processing nor post-processing. For the second line, we applied Simplification prior to segmentation and summarization. The third line adds the titling step for each summarized block. And finally the formatting step is added to obtain the last line (*Darbarer*). We decided to evaluate the results with metrics usually used for the summarization task: ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019). These metrics are also used during the evaluation process of the task alongside a human evaluation. However, Ghosal et al. (2021) have shown that these metrics have poor correlations with human judgement. During our research, we found similar discrepancies with our results, which seemed perceptually better without an improvement of the scores. We thus decided to look at the number of words and blocks generated by our systems, with the assumption that shorter summaries will improve the overall readability and clarity of the minutes.

While the overall values of ROUGE and

<sup>6</sup>We used `Helsinki-NLP/opus-mt-cs-en` and `Helsinki-NLP/opus-mt-en-cs` checkpoints, available on HuggingFace.

<sup>7</sup><https://github.com/chakki-works/sumeval>

<sup>8</sup><https://pypi.org/project/bert-score/>

	elmiEN		europarl	
	Document-level	Hunk-level	Document-level	Hunk-level
<b>Adequacy</b>	$3.14 \pm 0.60$	$4.03 \pm 1.11$	$2.33 \pm 1.21$	$4.44 \pm 0.81$
<b>Grammaticality</b>	$4.92 \pm 0.18$	$4.93 \pm 0.41$	$5.00 \pm 0.00$	$5.00 \pm 0.00$
<b>Fluency</b>	$3.64 \pm 0.42$	$4.17 \pm 1.11$	$3.50 \pm 1.05$	$4.44 \pm 0.81$
<b>Relevance</b>	$4.67 \pm 0.67$	$4.46 \pm 0.71$	$4.83 \pm 0.41$	$4.94 \pm 0.25$

Table 7: Average human evaluation scores (1: worst, 5: best) for English meetings. The figures correspond to mean $\pm$ standard deviation.

		Summarization	Simplification	Titling	Formatting	R1	R2	RL	BERT Score (scaled)	Words	Blocks
test	baseline	✓				<b>0.32</b>	<b>0.08</b>	<b>0.18</b>	<b>0.44</b>	392	12
	+ simplification	✓	✓			0.29	0.06	0.16	0.42	<b>294</b>	<b>8,7</b>
	+ titling	✓	✓	✓		0.29	0.06	0.16	0.42	316	<b>8,7</b>
	Darbarer	✓	✓	✓	✓	0.30	0.06	<b>0.18</b>	<b>0.44</b>	330	<b>8,7</b>
test2	baseline	✓				<b>0.33</b>	<b>0.08</b>	0.19	0.41	417	14
	+ simplification	✓	✓			0.28	0.07	0.18	0.40	<b>310</b>	<b>9,8</b>
	+ titling	✓	✓	✓		0.29	0.07	0.18	0.40	339	<b>9,8</b>
	Darbarer	✓	✓	✓	✓	0.31	0.07	<b>0.20</b>	<b>0.43</b>	352	<b>9,8</b>

Table 8: Results of the ablation study. The ROUGE scores were computed using the Sumeval<sup>7</sup> library, removing stopwords from the provided list. The BERTScore was computed with the bert-score library<sup>8</sup> using the *rescale with baseline* option for a larger range and better human-readability of the score. The words and blocks column show the average number of words and blocks across minutes.

BERTScore do not strongly change for any of these steps, we observe some general patterns in the results:

- The simplification module seems to decrease both the ROUGE and BERTScore by a few points. However, it allows the system to produce shorter minutes (by about 33%) and of seemingly better quality when looking at the actual content of the minutes. Looking more closely to the results, we see that the precision component of the metrics increases, while the recall one falls by a few points.
- The titling module has little to no effect on the metrics, but allows for better readability.
- The formatting improves the ROUGE and BERTScore by a few points.

## 5.2 Error analysis

### 5.2.1 Simplification

The simplification process occasionally removes sections of the transcript that could be essential for creating an accurate summary. Additionally, it may inadvertently alter the meaning of certain sentences, potentially leading to misunderstandings or misinterpretations in the summarized content. The

following are examples of transcripts followed by their simplification (right side of the arrow). In the first case, the meaning is slightly altered and the information about “manual alignments” is removed. In the second example, the output is not simplified and removing the first part alters the general meaning. even though a thorough evaluation should be achieved. While these phenomena seem to be quite rare, a more thorough evaluation should be conducted in order to clearly quantify their frequency and impact.

- (Person6) Great, so we do alignment, fine the manual are done, but what is the final output?  $\rightarrow$  (Person6) *How is the final output?*
- There is nothing that I would know about that we need to discuss uh, like in in in a very big detail, ehm.  $\rightarrow$  *I would know about that we need to discuss uh, like in in in in a very big detail*

### 5.2.2 Summarization distortions

Some errors were produced during the summarization step. One of these recurrent errors was generating inaccurate tags (we refer to anonymized entities such as PERSON, ORGANIZATION, PROJECT and LOCATION as tags). For example, *Organizing6 / Organizer* instead of [ORGANIZATION6]

or *Person A* / *PERSO* / *PERSS* instead of [PERSON1].

We thus proceeded to a manual analysis on all the tags present on the `test` partition to extract the statistics shown on Table 9. As we can see, this type of error appears in only 3.6% of the generated tags, but they are particularly harmful for the general meaning of minutes.

# generated tags	358
# wrong tags	15
percentage of wrong tags	3.6%

Table 9: % Tag errors in generated minutes of `test`

### 5.2.3 Person tracking

Dialogue summarization models face a challenge when it comes to tracking the identity of speakers, addresses and people indirectly mentioned with third person pronouns, particularly when there are many parts in the conversation. Specifically, these models encounter difficulties in accurately determining the referent of a personal pronoun (e.g., “you”) when transitioning from direct speech in the conversation to an indirect speech format for the summary. This challenge appears because the model needs to infer the identity of the pronouns based on the conversation’s structure to appropriately assign them in the summarization process.

In order to overcome this challenge, researchers have explored various strategies to enhance the performance of dialogue summarization models. One promising approach that has been recently proposed by Fang et al. (2022) is to replace each pronoun with its specific noun. This technique helps the models to avoid misplacing the nouns during the summarization process, which can significantly improve the overall quality.

We also detected ambiguities regarding pronouns in the generated minutes. Some bullet points in the minutes were generated with pronouns such as *he*, *she* or *them*, while it is impossible to guess who they refer to without any context about the conversation. The following is an example :

- PERSON3 is not sure whether he will join

In this example, *he* seems to refer to PERSON3 while in the context it refers to PERSON11.

We checked the minutes generated for the first 9 transcriptions of the `test` partition to pull out the percentage of pronouns with unclear antecedent.

We observed 14 indefinite pronouns, among which 4 of them could not be resolved given the summarized context. Here again this type of error can be misleading for the general understanding of the minutes. An additional analysis on the first transcript from the `train` partition revealed that among the 89 occurrences of the pronoun *you*, 43 corresponded to the previous speaker and 33 corresponded to the last mentioned tag. This illustrates that resolving the *you* pronouns is not a trivial task. Further analysis should be achieved to better understand the impact of person tracking on the overall acceptability of the generated minutes.

### 5.2.4 Titling

The automated generation of titles in the dataset is not entirely error-free. Various issues can be observed, which may lead to misunderstandings while reading. These errors can be broken down into multiple types: grammatical mistakes (e.g. “*Meet today in person*”), semantic inaccuracies (e.g. “*Summarisation of the minutes annotation*”), or nonsensical phrases (e.g. “*Edit of eh*”).

We checked the minutes generated by our model on the first nine meetings on the `test` partition to see if the title for each block is coherent or if it contains grammatical mistakes or semantic inaccuracies. We observed that 54 titles out of 70 were fully coherent.

## 6 Conclusion and discussion

In this paper, we described our system for the AutoMin 2023 challenge Task A and detailed its four different modules: simplification, summarization, titling and formatting. Our submitted system produces meeting minutes that are concise, intelligible and that may already be usable without further modifications, in a multitude of use cases. However, it is not error-proof and still subject to improvement, regarding the way we could cleverly split the conversation into coherent segments, or how to ensure correctness in regards to grammar, semantics and person tracking. Moreover, this work highlights the need for better metrics for evaluating the results of abstractive summarization systems in order to make better informed decisions for the design of the whole pipeline. We believe this Shared Task to be very relevant, especially in times where automatic content summarization is becoming more and more common.



## References

- Joseph W Alba, Susan G Alexander, Lynn Hasher, and Karen Caniglia. 1981. The role of context in the encoding of information. *Journal of Experimental Psychology: Human Learning and Memory*, 7(4).
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- François Buet and François Yvon. 2021. Toward genre adapted closed captioning. In *Interspeech'21*.
- Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *The 28th International Conference on Computational Linguistics*.
- Yue Fang, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Bo Long, Yanyan Lan, and Yanquan Zhou. 2022. From spoken dialogue to formal summary: An utterance rewriting for dialogue summarization. In *Proc. NAACL*.
- Tirthankar Ghosal, Ondřej Bojar, Marie Hledíková, Tom Kocmi, and Anja Nedoluzhko. 2023. Overview of the second shared task on automatic minuting (automin) at inlg 2023. In *Proceedings of the 16th International Conference on Natural Language Generation: Generation Challenges*. Association for Computational Linguistics.
- Tirthankar Ghosal, Ondřej Bojar, Muskaan Singh, and Anja Nedoluzhko. 2021. Overview of the first shared task on automatic minuting (automin) at interspeech 2021. *Proceedings of the First Shared Task on Automatic Minuting at Interspeech*.
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsun corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.
- Eva Hasler, Adrià de Gispert, Felix Stahlberg, Aurelien Waite, and Bill Byrne. 2017. Source sentence simplification for statistical machine translation. *Computer Speech & Language*, 45:221–235.
- Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. **BARThez: a skilled pre-trained French sequence-to-sequence model**. In *Proc. EMNLP'21*, Dominican Republic.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. Machine Translation summit*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. **BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension**. In *Proc. ACL'20*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.
- Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. Elitr minuting corpus: A novel dataset for automatic minuting from multi-party meetings in english and czech. In *Proc. LREC'22*.
- Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11).
- Peter Polák, Muskaan Singh, Anna Nedoluzhko, and Ondřej Bojar. 2022. **Alignmeet: A comprehensive tool for meeting annotation, alignment, and evaluation**.
- Deborah Schiffrin. 1990. Conversation analysis. *Annual Review of Applied Linguistics*, 11:3–16.
- Kartik Shinde, Nidhir Bhavsar, Aakash Bhatnagar, and Tirthankar Ghosal. 2021. Team abc@ automin 2021: Generating readable minutes with a bart-based automatic minuting approach. *Proceedings of the First Shared Task on Automatic Minuting at Interspeech*.
- Sara Botelho Silveira and António Branco. 2012. Enhancing multi-document summaries with sentence simplification. In *Proc. ICAI'12*, page 1.
- Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proc. EAMT'20*, Lisbon, Portugal.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. volume 4, pages 401–415.
- Atsuki Yamaguchi, Gaku Morio, Hiroaki Ozaki, Kenichi Yokote, and Kenji Nagamatsu. 2021. Team hitachi@ automin 2021: Reference-free automatic minuting pipeline with argument structure construction over topic-based summarization. *arXiv preprint arXiv:2112.02741*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. **A survey of large language models**.

## Generated minute for the test\_04 meeting



Date: 11.9.2022



Attendees: PERSON13, PERSON3, PERSON10, PERSON14, PERSON7



### Making a presentation platform

- ◆ [PERSON10] and [PERSON3] are working on a presentation platform
- ◆ [PERSON10] wants to know if anyone is willing to play with it.



### Working on text client

- ◆ [PERSON10], [PERSON13] and [PERSON7] are working on a text client which should be able to send text to text worker
- ◆ It works with the Czech Czech machine translation, but there is a problem with the batch processing mode and the ASR.
- ◆ They need to perform more test next week.



### Quality of online ASR and machine translation

- ◆ [PERSON10], [PERSON13] and [PERSON14] are working on improving the quality of online ASR and machine translation.



### Batch mode and segmentation work

- ◆ [PERSON7] created a batch mode, but it didn't work
- ◆ The segmentation workers don't work with the CTM client
- ◆ The ASR worker starts producing text in the chat window



### Control of segmentation worker

- ◆ [PERSON10] wants to know more about the segmentation worker
- ◆ It only handles the text as it comes out of the segmentor
- ◆ There will be a session, session with and a test next week



### Process of the presentation platform

- ◆ [PERSON3] has prepared a presentation platform for June
- ◆ The presentation platform will have an operator monitoring the output of one of the re-speakers cabins and if the output from the floor is bad, the operator should kill the client and switch to the other provided translation



### Implementation of MT outputs

- ◆ There are 4000 people connected on the same WiFi network
- ◆ The current user is expected to be at June
- ◆ They want to know which of the MT outputs is the best at the moment
- ◆ They need to decide how to deliver the subtitles to the participants
- ◆ They have a year to find a better solution



Submitted by: Darbarer

## Generated minute for the test\_10 meeting



Date: 22.8.2022



Attendees: PERSON3, PERSON2, PERSON4, PERSON1



### Record the meeting

- ◆ [PERSON3], [PERSON2] and [PERSON3] are going to record the meeting
- ◆ [PERSON3] will send the poll for the next week as well
- ◆ Organizing Committee will divide the budget for the meeting among other parties, but each party will get their own funding.



### Preparation of work package

- ◆ [PERSON3] asks Organizing Committee to prepare a work package for the presentation application development for live meetings
- ◆ She also asks for a dry-run and a follow-up workshop



### Work plan for the project



[PERSON1], [PERSON2] and [PERSON3] are discussing the organization's work plan for the three-year-long project



### Preparation of speaker



[PERSON3] and [PERSON2] discuss how to prepare a speaker for a conference.



### recording and the adaptation of a voice

- ◆ [PERSON1] and [PERSON3] explain to each other what is required for the recording and the adaptation of a voice.



### Preparation of proposal

- ◆ [PERSON3] and [PERSON1] have 14 days to prepare a proposal
- ◆ They need the audio equipment for the re-speakers, and they need to check the availability of specific hardware
- ◆ They also need to work on the integration of ASR essential from multiple partners into the platform.



### Design of deliverables



[PERSON3], [PERSON4] and [PERSON1] discuss the design of deliverables for the project

- ◆ The deliverables should be in line with the timing of the work packages.



Submitted by: Darbarer