Team AutoMinuters @ AutoMin 2021: Leveraging state-of-art Text Summarization model to Generate Minutes using Transfer Learning

Parth Mahajan[1]*,Muskaan Singh*[2]*, Harpreet Singh*[1]

[1]Thapar Institute of Engineering and Technology, Patiala, India
[2] IDIAP Research Institute, Martigny, Switzerland

pmahajan_be18@thapar.edu ,msingh@idiap.ch, akalharpreet@gmail.com

Abstract

This paper presents our submission for the first shared task of automatic minuting (AutoMin@Interspeech 2021). The shared task consists of one main task generate minutes from the given meeting transcript. For this challenge, we leveraged state-of-art text summarization models to generate minutes using the transfer learning approach. We also provide an empirical analysis of our proposed method with other text summarization approaches. We evaluate our system submission quantitatively with 33% BERTscore and 11.6 % ROUGE_L, which is relatively higher than the average submission in the shared task. Along with the qualitative evaluation, we also vouch for quantitative assessment, where we achieve (2.32, 2.64, 2.52) scores out of five for adequacy, grammatical correctness, and fluency. For the other two tasks, we use Jaccard and cosine text similarity metrics for a given transcript-minute pair corresponding to the same meeting (Task B) and if a given pair of meeting minutes belong to the same meeting (Task C). However, our simple approach yielded 94.8 % (task B) and 92.3% (task C), clearly outperforming most submissions in the challenge.

We make our codebase release here https://github.com/mahajanparth19/Automin_Submission.

Index Terms: speech recognition, human-computer interaction, computational para-linguistics

1. Introduction

Due to the onset of the covid pandemic, most of our interactions shifted to the virtual environment. Due to this, the need for automatic summarization of meetings became imminent. The demand for summarizing generation systems increased significantly with a large transcripts size. Meeting summarization is currently considered a crucial topic in natural language processing. Survey has shown that the task of summarizing meetings by generating summaries in the form of structured minutes from speech can potentially save up to 80% of time. The task of automatic minuting seems closely related to meeting summarization; however, both tasks have some differences. Automatic minuting is a challenging task because there is no universal framework for generating minutes, and it varies across different types of meetings, subjects, and objectives. Hence, one annotator might give importance to some topic while others may have a completely different minute for the same meeting. Also, the content in the minutes might depend on its intended audience as different people might require additional information from the same meeting. There is no fixed format of representation, so different persons might represent the information with a different perspectives regarding the use of novel words, briefness in discussing topics, ignorance of information, and use of phrases. Apart from this, a minute is also judged based on its readability, adequacy, grammar, clarity, coverage, etc.

Automatic minuting is closely related to and often considered the same as meeting summarization. Recently, there have been many attempts to solve this problem by using abstractive summarization techniques and extractive summarization techniques. Here, we have mentioned some of the work in this field. Abstractive Meeting Summarization with Entailment and Fusion[1] proposes an end to end abstractive summarization framework in which the input senteces ared clustered into communities and those communities are used to build an entailment graph. It selects the most relevant sentences and then uses ranking strategies to find the best path in the graph to form abstract sentences. Abstractive Multi-Modal Meeting Summarization[2] proposes a abstractive meeting summarizer that uses multi-modal information to generate summaries. This multi-modal hierarchical attention is used on segment, utterance, and word levels, and also topic segmentation is utilized to get topic-relevant segments. In this paper we address this problem with our submission to the AutoMin shared task[3].The shared task presented by AutoMin has 1 main task and 2 sub tasks. The main task being generation of minutes from the transcript, and the 2 sub tasks were of validation of transcript and minute pair, and comparison of 2 minutes to identify whether they belong to same meeting or not. We experiment on their dataset and try to generate a system which can generate minutes, validate if a minute belongs to meeting and also compare if two minutes belong to same meeting.

2. Task A: Generating Automatic Minutes

This shared task's main task was to generate minutes from a given multiparty transcript automatically. Both automatic and manual methods would then evaluate the generated transcript. The generated summaries are compared with original summaries using the ROUGE [4] metric, *Adequacy*, *Relevance*, *Coverage*, *Readability*, and *Grammaticality* (i.e Human Evaluation) will be considered as the most critical aspects evaluating generated minutes

Task-A's dataset has 85, 10, and 28 instances for train, dev, and test sets. Each instance comprises (i) a meeting transcript; (ii) one or more than one annotated meeting minutes corresponding to the given transcript. Table 1 show relevant statistics of the datasets that we have used in experimentation. Each train and dev set instance contained one transcript and multiple minutes. Some were generated by people present in the meetings and others by those who had just read the transcript.

The transcript in the dataset contained turn-wise separated dialogues, and the speaker of the dialogue is denoted by parenthesis and is present at the start of the line. The transcript also included unique entities mentioned in square brackets (Used for referencing a person or some organization). The data is first pre-processed. The steps involved in pre-processing were

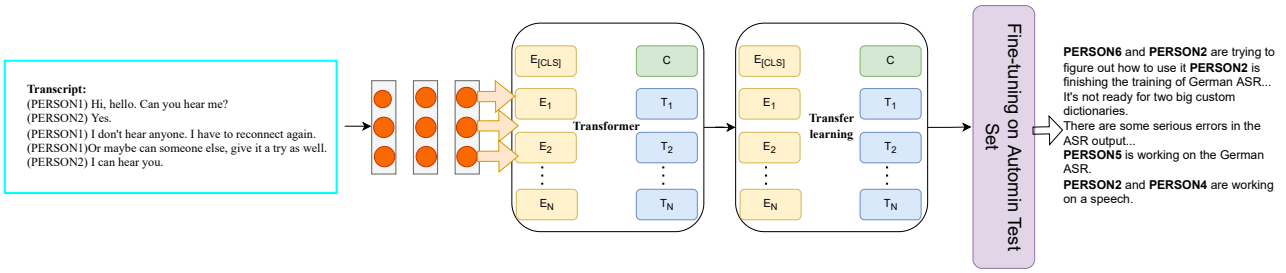

Figure 1: *Proposed System*

Datasets	#Meetings	avg words per Transcript	avg words per summary	avg turns per transcript	avg # speakers
AMI	137	6,970	179	335	4
ICSI	61	9,795	638	456	6.2
Automin 120	7,066	373	727	5.9	

Table 1: *Statistics of the datasets being used in our experiments*

changing the case and converting all the words in the transcript to lowercase; after this step, stop words removal was done using the NLTK library[5] available in python. It is relevant from all the words in the transcript to find the most commonly used words that did not provide any significant information, along with the stop words; these were also removed from the transcript. Also, we removed words between angular brackets as they did not provide significant information. The words ended with a hyphen as they were incomplete and, hence, no use for experimentation. We also did convert the dataset into a pandas data frame to experiment. So, for our task, we utilized transfer learning to leverage the pre-trained T5 model[6] and fine-tune it according to our requirement. The model we have used is T5-base by google provided by hugging face transformer available here. The model is fine-tuned on AutoMin dataset and presented in Fig.1.

It is an encoder-decoder transformer model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format. T5 was pre-trained by fill-in-the-blank-style denoising objectives. The text-to-text framework of the T5 model allows it to use the same hyper-parameters and loss functions on any NLP task. As the architecture of the T5 model supports multiple tasks, we have to append a particular word to the start of the text. It helps the model in recognizing the task to be performed. To achieve this, we had to append the word ”summarize” to the start of all texts. After that, we tokenized the text using the T5 tokenizer.
Then, the data was fed to the model using the trainer function available in the transformer library. Using this, we were able to fine-tune our dataset. The hyperparameters for this experiment were: learning rate 2e-8, and it ran for ten epochs. Rouge metric was used to evaluate the performance of the model.
The fine-tuned model was then used to generate summaries for the test set. As the model cannot handle extensive transcripts, we had to break the transcripts into segments of 500 words that were then fed to the model, and the outputs of all segments were then combined to generate the complete summary of the transcript. We evaluated the quantitative results of AutoMin (Refer table 2), and out of all these, T5 [6] gave better results in terms of sentence formation and readability. Table 3 shows the comparison of average scores for humans and automatic evaluation for our system with the average of all other submissions. We also experiment with Pegasus[7], BART[8], BERT[9], LED[10], Roberta2Roberta[11] and some extractive techniques such as word frequency[12], Luhn[13], LSA[14], TextRank[15], LexRank[16] and unsupervised[17]. Pegasus[18] transformer-based encoder decoder model pre-trained on massive text corpora with a new self-supervised objective. Pegasus is trained by removing/masking important sentences from input document and these are generated together as an output sequence from remaining sentences, similar to extractive summary. Another denoising auto-encoder BART[19] for pretraining sequence-to-sequence models. It is trained by corrupting text with an arbitrary noising function, and learning a model to reconstruct the original text. Bidirectional Encoder Representations from Transformers or BERT[**?**] for short is a machine learning technique. It is based on transformers for natural language processing. BERT reads the entire text at once rather than from left to right or right to left, and hence called bidirectional. This helps BERT to understand the entire context of the word, from left as well as right, providing a much more accurate representation. As part of the training process, it masks roughly 15% of the words, and attempts to predict them from the surrounding context. Moreover, it provides pairs of sentences as inputs and aims to check whether the two sentences are immediately consecutive in the provided text.

3. Task B & C: Semantic Similarity

The second subtask, Task-B, aims to identify whether or not the minute in a given minute-transcript pair matches with the transcript paired with it. The gist of this task lies in understanding the differences between two documents and determining how similar their context is. While the third subtask, Task-C, aims to identify whether or not a given pair of meeting minutes corresponds to the same meeting transcript. As discussed earlier, a given transcript minuted separately by multiple annotators, can have significantly different minutes in terms of the information within them, semantics used by the annotators, their briefness and the structure. Task-B's dataset comprises transcript-minute pairs and corresponding labels: i.e., TRUE and FALSE. Here, the Truth indicates that the minute matches with the paired transcript, and viva versa. As annotations done by different anno-

Table 2: *Quantitative results for the automin*

Model	R_1	R_2	R_We	MoverScore	BERTScore	BLEU
Bart	0.2488	0.063695	0.062282	0.002088	0.320863	0.152472
Bert	0.207337	0.036772	0.049543	-0.014163	0.289421	0.228029
Bert2Bert	0.235137	0.051975	0.06228	-0.014168	0.194232	0.15545
LED	0.092421	0.012805	0.0052	-0.040083	0.358017	0.262168
Pegasus	0.227211	0.045568	0.046668	-0.040189	0.291273	0.166803
Roberta2Roberta	0.166744	0.03124	0.031326	-0.024291	0.280946	0.28908
Word Frequency	0.190632	0.032974	0.036378	-0.016208	0.25305	0.129849
Luhn	0.237608	0.064312	0.080381	0.005292	0.166229	0.190505
LSA	0.235285	0.077394	0.089034	0.00823	0.146165	0.2243
TextRank	0.229602	0.054555	0.071968	0.007953	0.179273	0.183297
Unsupervised	0.234509	0.050407	0.026865	0.003819	0.299319	0.226086
LexRank	0.225526	0.041421	0.051358	-0.010625	0.249453	0.160923
Our Proposed System	**0.270173**	**0.06711**	**0.07597**	**-0.007101**	**0.3331**	**0.167922**

Teams	**Adequacy**	**Grammatical Correctness**	**Fluency**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Average of all teams	2.81	3.25	2.92	0.203	0.0458	0.114
Ours	**2.32**	**2.64**	**3.25**	**0.212**	**0.0440**	**0.114**

Table 3: *Results for subtask A on test data. The adequacy, grammatical correctness, and fluency are evaluated manually by two annotators and assessed on a Likert Scale of 1 to 5. These scores are based on official results provided by the organisers and are averaged across all test set samples.*

Dataset	True tag	false tag	**Total**
Task B	115	731	**846**
Task C	74	660	**734**

Table 4: *Class-wise distribution of train and dev data.*

tators can vary in terms of writing style, grammar, tone, vocabulary, etc., the minutes curated by each annotator are different and have a unique format. On the other hand, Task-C's dataset comprises meeting minute pairs and their corresponding labels, again TRUE/FALSE. Here, the Truth indicates that the paired minutes correspond to the same meeting transcript, and viva versa. Both these datasets have a vast imbalance between the instances belonging to the two classes (as represented by Table 4). In both cases, the Truth, the minority class, has merely 15% representation in the entire data. The task in this sub-task was to validate if the minute belonged to the given transcript or not. For this task, we used different similarity scores to find the relation between the transcript and the minute file, To achieve this, we use Jaccard score[20] in Eq.1 and Cosine similarity [21] in Eq.2 value for a minute and transcript pair, and with these, we also stored the ground truth label. After this, we fed this dataset to a KNN[22] model, which then uses this dataset to classify whether a new minute belongs to the corresponding transcript or not. Using the validation set, we tried to find the optimal value of k and found that even though values were close, keeping k as 7 worked well. We present our results.

$$Similarity(A, B) = \frac{\|A \cap B\|}{\|A \cup B\|} \tag{1}$$

$$Similarity(A, B) = \frac{A.B}{\|A\|.\|B\|} \tag{2}$$

We tried multiple values K 5 and found the one that gave the best results.

We can observe from the table that when we selected k=7 and after seven, the accuracy and other metrics started to decrease.

Table 5: *Task- B (Dev-Set)*

K	Accuracy	Precision	Recall	F1
1	0.818	0.83	0.82	0.82
3	0.857	0.84	0.86	0.85
5	0.875	0.86	0.88	0.87
7	**0.886**	**0.86**	**0.89**	**0.87**
9	0.882	0.85	0.88	0.86

Table 6: *Task- B (Test-Set (k=7))*

System	Accuracy	Precision	Recall	F1
Ours	0.948	0.94	0.95	0.94
Majority	0.944	-	-	-

K=7 was used for final submission, and results can be seen in **??**

The task C sub-task was to compare two minutes and classify whether they belonged to the same meeting.
The methodology for this task was also kept similar to the previous one.
We created a dataset using Jaccard score [20] in Eq 1 and cosine similarity score [21] in Eq 2 for each minute pair and also stored the corresponding ground truth labels. And then, this data was fed to a KNN model, which was then predicted on the dev set using this dataset. And could classify whether two minutes belonged to the same meeting or not.
After this, we used this model on the test set to classify the given pairs into true or false. The final values were recorded in a CSV file, then submitted. For this sub-task, we tried different values of K 7 and then selected the one that gave the best results.
We found that 7 gave us good results by performing these experiments, so we kept K as 7 for running on test set8.

Table 7: *Task- C (Dev-Set)*

K	Accuracy	precision	recall	F1
1	0.881	0.88	0.88	0.88
3	0.905	0.89	0.90	0.90
5	0.923	0.92	0.92	0.91
7	**0.915**	**0.91**	**0.92**	**0.90**
9	0.913	0.90	0.91	0.90

Table 8: *Task- C (Test-Set (k=7))*

System	Accuracy	Precision	Recall	F1
Ours	0.923	0.92	0.92	0.92
Majority	0.936	-	-	-

4. Conclusion And Future Work

In this paper, we have described our system submission for AutoMin @ Interspeech 2021 on automatic minuting and analyzing and comparing meeting data. The proposed system utilizes transfer learning techniques to leverage a pre-trained transformer-based model and fine-tune that model according to our needs. We have also discussed the approach for the subtasks where we are comparing different minutes and transcripts. In the future, we will try to enhance the proposed system by applying topic segmentation techniques so that the segments we make are coherent. It might improve accuracy as we get topic-wise summaries that can then combine. For the subtasks, we can try to apply various other scoring techniques and pre-processing to increase the system's accuracy. Also, we would try to combine these systems in a GAN-like architecture so that the system can generate summaries and try to analyze those and improve its performance.

5. References

[1] Y. Mehdad, G. Carenini, F. Tompa, and R. T. Ng, "Abstractive meeting summarization with entailment and fusion," pp. 136–146, Aug. 2013. [Online]. Available: https://aclanthology.org/W13-2117

[2] M. Li, L. Zhang, H. Ji, and R. J. Radke, "Keep meeting summaries on topic: Abstractive multi-modal meeting summarization," pp. 2190–2196, Jul. 2019. [Online]. Available: https://aclanthology.org/P19-1210

[3] T. Ghosal, O. Bojar, M. Singh, and A. Nedoluzhko, "Overview of the first shared task on automatic minuting (automin) at interspeech 2021," in *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, 2021, pp. 1–25. [Online]. Available: http://dx.doi.org/10.21437/AutoMin.2021-1

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

[5] E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[6] 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," *CoRR*, vol. abs/1910.10683, 2019. [Online]. Available: http://arxiv.org/abs/1910.10683

[7] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 328–11 339.

[8] 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," *arXiv preprint arXiv:1910.13461*, 2019.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[11] S. Almasian, D. Aumiller, and M. Gertz, "Bert got a date: Introducing transformers to temporal tagging," *arXiv preprint arXiv:2109.14927*, 2021.

[12] E. Hovy and D. Marcu, "Automated text summarization," *The Oxford Handbook of computational linguistics*, vol. 583598, 2005.

[13] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[14] J. Steinberger, "Text summarization within the lsa framework," Ph.D. dissertation, PhD Thesis. University of West Bohemia in Pilsen. Pilsen, 2007.

[15] C. Mallick, A. K. Das, M. Dutta, A. K. Das, and A. Sarkar, "Graph-based text summarization using modified textrank," in *Soft computing in data analytics*. Springer, 2019, pp. 137–146.

[16] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.

[17] T. Nomoto and Y. Matsumoto, "A new approach to unsupervised text summarization," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 26–34.

[18] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: pre-training with extracted gap-sentences for abstractive summarization," *CoRR*, vol. abs/1912.08777, 2019. [Online]. Available: http://arxiv.org/abs/1912.08777

[19] 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," *CoRR*, vol. abs/1910.13461, 2019. [Online]. Available: http://arxiv.org/abs/1910.13461

[20] G. Ivchenko and S. Honov, "On the jaccard similarity test," *Journal of Mathematical Sciences*, vol. 88, no. 6, pp. 789–794, 1998.

[21] F. Rahutomo, T. Kitasuka, and M. Aritsugi, "Semantic cosine similarity," in *The 7th International Student Conference on Advanced Science and Technology ICAST*, vol. 4, no. 1, 2012, p. 1.

[22] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2003, pp. 986–996.

6. Generated Samples

Following is an example of minutes generated by our model sampled from the data for Task A:

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

SUMMARY-

- i've emailed [person8], if [person8] would be joining as well. don't think there is any way to switch off your webcams in the alfa view, unless you simply put a ducktape over ovet them.
  see the presentation platform slowly taking its form.. in some cases and also, there is a chance of having the real video mixers.
  a text client would be useful also for debugging. if anything comes out that is also as a side effect that is also text client, it would be very useful..
  czech machine translation still has some bugs. kt translation platform does not handle batch eh the batch example very well.

- ehm some checks unintelligible¿ that relate to partial sentences.
  i
- if one sentence in your chat testing data does not end in the period or a full stop, it will pend the next sentence because it thinks that them those two together make a full sentence..

- don't see any big benefit from fixing the batch processing mode.
- instead, send unintelligible¿ audio files to [person1] and [person1] did the asr offline.

- if [person12] sends unintelligible¿ audio files to [person1] and [person1] did the asr offline, that would be like the way to get the transcriptions..

- cnn's john sutter talks to sutter about the offline asr.

- sutter wants to test the mission translation.

- sutter wants to use a text only client..

- i believe what happens is for the dutch people unintelligible¿ so you already have segment unintelligible¿.

- the offline asr does that in a different process.

- there is even maybe no segmentation happening in word..

- the chat is filling up with these messages but no asr worker being available.

- i think our main concern is not the quality of the translation, but the mismatch of the segmentations.

- i would like to ask [person12] to send the ted files, which do have their translations..

- i think that we are just seeing again, the same bug.
  but we do not know what exactly the bug is that [person1] will be fixing sooner or later.

- at some point, they will both be redone, but i can't promise that will be next week..mt clients in- and the workers, including the [project2] one, expect already the text.
  but but the asr workers are not reliable to to to debug it or unintelligible¿ to debug.

- there is no immediate plan to to have such client, that would be able to digest ctm..

- cnn's john sutter looks at the asr output to see if there is a bug.

- the asr workers are registered as non-ideal even when they have finished their jobs.

- sutter says the bug is still triggered..

- slt [organization5] is the sort of everything except the actual neural networks part that is doing the work.

- the segmentation worker is a sequence labeling-other_noise¿labels each word with either opf insert coma or insert the full stop.

- slt [organization5] is connected to the mediator, doing some preprocessing for mt, for example, like ppe, and that sort of thing..

- if the mediator dies, their worker should survive and reconnect again.

- if there are other things, the output of the asr which will get from you offline, will be different from the output that will get from this improved pipeline that you maybe using.

- if there are other things, than the output of the asr which will get from you offline, will be different from the output that will get from this improved pipeline that you maybe using..

- the presentation platform will receive all those streams. we can choose the main stream for each language.

- the actual solution is to have a kind of combinatory explosion of all the possible match matching path..

- ehm is proposing that all audio inputs are translated into all target languages at all times.

- if the output from one of the re-speaking cabins or the floor is bad, this operator should kill the client that is unavailable.
- killing the client is useful only to ehm in computational power on servers which was the workers..

- for june it will not be available the preview of the subtitle.
- but the selection of the stream should be available.

- i can make a blind choice, within the presentation platform control, i can make a blind choice..

- the final user will be the client of the subtitle solution.
- the client will connect to a particular stream of publishing subtitles.
  's possible that this will happen more often than normal.. cnn's jarrett bellini talks with cnn's mike downey. he says he's unsure if 400 can receive the signal from his wifi. downey: "we have a year to to find better one.

- we could also add wires" downey: "i think that's everything for today.i think that's everything for today".

No need for this sample And, on the following page there is a true positive instance predicted by our model, for TASK-C :

Minute:A)

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

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

Milestones

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

Minute:B)

Organizational stuff

- Monthly call will be on Thursday, 5 PM LOCATION1 time
  - At least PERSON14 and PERSON10 should take part
  - PERSON14 will care about including PERSON6 into the mailing list
- PERSON6's coming to LOCATION1
  - It is very desirable that PERSON6 comes to LOCATION1 in person
  - Visa issues due to Covid situations

PROJECT2

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

Progress on PROJECT6

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

PROJECT4

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