Team The Turing TESTament@AutoMin 2021: A Pipeline based Approach to Generate Meeting Minutes Using TOPSIS

Umang Sharma[1], *Muskaan Singh*[2], *Harpreet Singh*[1],

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

umangsharma.cs@gmail.com, msingh@idiap.ch, akalharpreet@gmail.com

Abstract

In this paper, we present our submission for AutoMin Shared Task@INTERSPEECH 2021. The objectives in this task were divided into three tasks, with the main task to create a summary based on a transcript from a meeting. The other two tasks were to compare minutes and transcripts to find out if they were from the same meeting or not. We propose a pipeline-based system that extracts the important sentences from the transcript using features and then a topsis algorithm to summarize. It creates a flexible system that can provide a set of sentences from any given transcript that can best describe it based on selected features and heuristic evaluation metrics. The proposed system presents readable, grammatically correct, and fluent minutes for given meeting transcripts. We make our codebase accessible here `https://github.com/umangSharmacs/theTuringTestament`.

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

1. Introduction

The Covid pandemic resulted in a sudden shift from the physical to the virtual world. Every meeting is hosted online, and the emphasis on work from home increases. In an era dominated by information and all the duties that come from managing said information, it becomes paramount to store it and access it properly. Thus, meetings being held online lead to another issue: uploading all the meetings onto a database for accessibility. However, due to limited storage options, a need for summarizing meetings such that they can be deemed as important or unimportant became essential.

Currently, such pre-trained systems exist and can be used for many meetings; however, they require memory to be of any significant use due to their pre-trained nature. The Turing TESTament system at AutoMin 2021 [1] aims to provide similar and, at times, better results without being a burden on the host machine that it would be running on. It uses a lazy learner approach, similar to K-Nearest Neighbours, and hence does not require any prior training. Moreover, its heuristic and feature-focused approach means that it has the flexibility to be used on multiple types of texts, rather than just transcripts.

The System aims to provide this flexibility by utilizing multiple Features and categorizing them using TOPSIS. While used for various Multiple Criteria Decision-Making problems, this approach has not been applied for single document summarising and therefore is worth exploring.

2. Task A: Generating Automatic Minutes

This is the main task where, given a multiparty meeting transcript, one has to automatically create minutes. The generated summaries are compared with original summaries using the ROUGE [2] 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).

Every summary has corresponding annotated meeting minutes(summary) that contain the AGENDA of the meeting and summaries based on the different participants and groups of the meeting. In the transcript, the speaker is indicated using curved brackets/parentheses ”()”, located at the beginning of their dialogue. Mentions of any participant or special entities are indicated using squared brackets/parentheses ”[].”

2.1. Methodology- pipelining approach

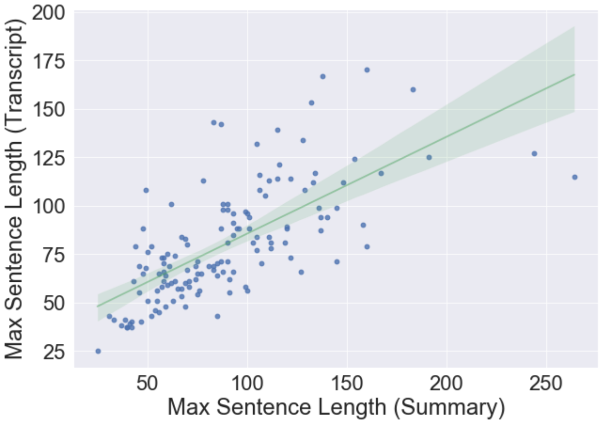
This section describes the pipelining approach we have used for our submission. We preprocessed the text using certain features and then applied topsis for extracting the important sentences from the transcript.

- Sentence Length The number of word tokens in a sentence has been termed the Sentence Length, where a word token comprises anything within two empty spaces. According to the AMI datasets, there is a high correlation between the sentence lengths of a transcript and its corresponding extractive summary (0.68 Pearson's Correlation), as can be seen in Fig.1, hence the inclusion of this feature.
- Vocabulary/ Word frequency (Unigram BoW): For each transcript, a Bag of Words is created by preprocessing the data, i.e., removing stopwords, stemming the remaining words, converting them to lower case, and finding the frequency of such words. Each sentence in the transcript is then scored based on the presence of these words.
- Numerical Data: This feature has been included because numbers such as statistics, age, years, etc. may be important for a transcript, and thus the sentences containing them should be included in the summary.
- Topics: Using Latent Dirichlet Allocation(LDA) [3], n

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*

Figure 1: *Regression Plot showing the correlation between the maximum sentence length between the transcript and its respective summary for the AMI dataset.*



topics are generated containing m words each. These are then converted to another vocabulary, out of which sentences are scored in the transcript. Here, individual topics have not been considered due to the repetition of (confusing) topics. Simply the presence of a word in the topic is noted rather than their frequency to minimize the impact of multiple similar topics that may arise if either n or m is very high.

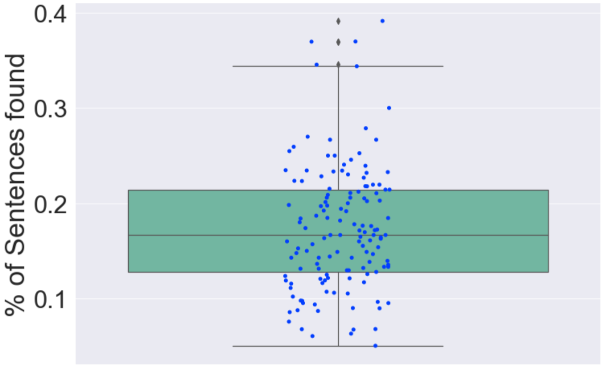
- Proper Nouns: A simple POS tagger is used to find sentences containing Proper Nouns and score them based on their frequency. Much like with *Numerical Data*, this feature has also been included on the basis that sentences containing Proper Nouns may be important and should be added to the summary.
- Affirmations: During a meeting, there are moments where several people agree with what is said by using words such as 'Okay', 'Yeah', 'Hmm' etc. By finding these affirmative phrases, they help isolate sentences to which the participants of the meeting agreed. Using this, an average of 17% of sentences were found that were included in the summary of the AMI dataset (Figure 2). Thus this feature has been included.

First, the sentences containing these affirmative words are isolated. If more than one such sentence appears consecutively, i.e., if more than one participant agreed with something said, then the three sentences before the first affirmative sentence are scored. TOPSIS is an MCDA method that uses Euclidean distance to rank alternatives. It has been used in this model to rank each sentence. After calculating the above features, a matrix A_{m*n} is obtained with n features (Six in this case) and m sentences. After which the following steps are taken,

- Vector Normalisation: The Normalised Matrix is calculated as

$$A_{normalised} = [r_{ij}]_{m*n} \tag{1}$$

Figure 2: *Boxplot showing the number of sentences found using this feature in the extractive summary of the AMI dataset*



where $r_{ij} = f_{ij}/\sqrt{\Sigma_{i=1}^{m} f_{ij}^2}$

- Calculate Weights: In order to calculate the weight of each feature n, Shannon's Entropy has been used (refer paper), by, Calculating entropy values,

$$e_j = -(\ln(m))^{-1} \Sigma_{i=1}^{m} p_{ij} \ln(p_{ij}) \tag{2}$$

where

$$p_{ij} = \frac{f_{ij}}{\Sigma_{i=1}^{m} f_{ij}}$$

- The weight is then calculated as,

$$w_j = \Sigma_{k=1}^{n} d_j / d_k \tag{3}$$

where,

$$d_j = 1 - e_j$$

Then finally, the Weighted normalized matrix is calculated where $v_{ij} = w_j * r_{ij}$

- Finding Euclidean Distances: After the Weighted normalized matrix has been calculated, then for each sentence, its Euclidean distance from the Positive Ideal solution and the Negative Ideal solution is found.

$$S_i^+ = \Sigma_{j=1}^{n} (v_{ij} - v_i^+)^2 \tag{4}$$

and

$$S_i^- = \Sigma_{j=1}^{n} (v_{ij} - v_i^-)^2 \tag{5}$$

where v_i^+ and v_i^- are PIS and NIS for the ith feature

- Get TOPSIS score and Rank : Once the Euclidean distances from the PIS and NIS for each sentence have been found, their TOPSIS score can be calculated as,

$$T_i = \frac{S_i^-}{(S_i^- + S_i^+)} \tag{6}$$

and then can be ranked based on this score.

2.1.1. Experimental Details

Applying TOPSIS ranks every sentence based on its features. Then, the System applies an iterative approach to remove the worst performing sentences, compare the new set of sentences with the original transcript, and use the Rouge and Meteor metrics to find the best possible subset of sentences from the original transcript. The removal of sentences is based on the range of scores per iteration. If the standard deviation between the previous five ranges is below 0.01, then the number of sentences to remove is increased; else, it is decreased. This helps to go through stagnant plateaus quicker. With each iteration, the System tries to maximize the Heuristic,

$$Heuristic = \frac{r_1 * r_2 * r_L * m}{Length^x} \tag{7}$$

where r_1, r_2, r_L, and m are the Rouge1, Rouge2, RougeL, and Meteor scores of the current set of sentences versus the original transcript, and $Length$ is the size of the current set of sentences. In the Heuristic, the x hyperparameter is used to punish summaries with larger lengths. For AutoMin this parameter is set to 3, i.e. the Length is being cubed. The Table 2 shows average results based on the Hyperparameter x which punishes summaries with higher lengths. If we consider any given Transcript as a set of sentences,

Table 2: *Hyperparameter Details*

Hyperparameter X	Rouge1	Rouge2	RougeL	Meteor
1	0.50	0.49	0.50	0.21
2	0.70	0.69	0.70	0.34
3	0.51	0.50	0.49	0.20

$$B = \{S_1, S_2, ..., S_m\}$$

then after applying TOPSIS, a certain number of worst performing sentences are removed, say k such that set B becomes,

$$B_1 = \{S_1, S_2, ..., S_{m-k}\}.$$

The Rouge and Meteor scores are calculated against the original set of B. The System tries to maximize the Rouge and Meteor scores while minimizing the size of the set, i.e., maximizing the Heuristic for every iteration, for a total of, say, t iterations.

$$B_{summary} = max\{H_{B_1}, H_{B_2}, ..., H_{B_t}\}$$

where $H_{B_1}, H_{B_2}, ..., H_{B_t}$ are Heuristic scores of set B_1, B_2 up to B_t respectively. This ensures that the best set of sentences from the transcript is extracted.

2.2. Results and Analysis

The proposed model significantly lower computation cost and memory usuage for generating minutes. We evaluate our results using ROUGE [2] or Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a scoring algorithm that calculates the similarity between a generated summary and a collection of reference summaries. 'N' here stands for the N-gram that would be used. For this model, ROUGE-1, ROUGE-2, and ROUGE-L have been used. METEOR [4] or Metric for Evaluation of Translation with Explicit ORdering is an evaluation metric that depends on the weighted harmonic mean of unigram precision and recalls where the recall is weighted higher than the precision. In addition to standard word matching, it also matches synonyms and thus provides a more accurate evaluation score. The model uses an iterative approach to remove the worst performing sentences by ranking them based on TOPSIS.

The weights for each feature are calculated using the approach mentioned in *An integrated Shannon entropy and TOPSIS for product design concept evaluation based on bijective soft sets*[5].

The methodology of utilizing Multiple-Criteria Decision Analysis (MCDA) techniques like TOPSIS for summarization has been used before:

1. *Generating an Overview Report over Many Documents*[6] The authors have tried to utilize TOPSIS to summarise and create an Overview Report for multiple related documents in this paper. TOPSIS has been used to determine the best overview using Saaty's pairwise comparison.
2. *Automatic Summarization of Textual Document*[7] Much like The Turing TESTament, the System introduced in this paper is also based on Feature Engineering and uses TOPSIS to Aggregate the Sentence features. The features used here were much more position-dependent, such as the Degree centrality criterion, Closeness centrality criterion, and Sentence position criterion. This is because the summaries were generated from Documents. However, we have refrained from using such features since meetings do not always follow a standard structure, and thus, introducing Positional features may result in biases.

Table 3 shows the Automatic results for the System as well as the Average results of other submitted Systems for comparison. It can be seen from these results that, on average, The Turing TEStament system did not perform well on the ROUGE-2 metric. Table 4 shows the Human Evaluation scores. The minutes were rated based on Adequacy, Grammatical Correctness, and Fluency on the Likert Scale of 1-5, where 1 signifies the worst and 5 signifies the best score.

Table 3: *Automatic Evaluation scores.*

Model	ROUGE-1	ROUGE-2	ROUGE-L
The Turing TESTament	0.157	0.046	0.092
Average	0.166	0.095	0.095

Table 4: *Human Evaluation Scores.*

Model	Adequacy	Grammatical Correctness	Fluency
The Turing TESTament	2.75	2.786	2.214
Average	2.335	3.111	2.626

As it can be seen from Table 4, minutes produced by The Turing TESTament were deemed Adequate and Grammatically Correct by the Annotators. On the other hand, they were not considered Fluent. The Turing TESTemant is an extractive summarising system. Thus, the minutes may not be as fluent as an abstractive summarising system.

Moreover, the results of our experiments have been mentioned below. The System was compared with the current state-of-the-art models: BERT, Pegasus, and T5.

3. Task B and C: Semantic Similarity

In this task, given a transcript and a summary, one has to find out whether the given summary is created from the provided transcript or not. Similarly, given two summaries, one has to find out whether they are created from the same transcript or not. This is important to uncover how minutes created by two different persons for the same meeting may differ in content and coverage. The dataset provided for Task B contains 280 dev folders, 972 test folders, and 566 training folders. Each folder contains a summary and a transcript. The dataset provided for Task C contains 378 dev folders, 1431 test folders, and 555 training folders. Each folder contains two summaries. In Task B, Cosine Similarity was used to calculate whether the given summary matches with the given transcript i.e. if the summary is from the transcript or not.

$$similarity(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \tag{8}$$

For the given summary and transcript, calculate their embeddings using sentence transformers. The *bert-base-nli-mean-tokens* model [8] was used. It maps all sentences to a 768 dimensional dense vector space. For each sentence embedding in the summary and transcript, their cosine similarities were calculated. A high similarity between the sentences indicated took the summary from the given transcript.

Table 5: *Training Dataset for Task B Results*

Threshold	0.60	0.70	0.75	0.80	0.90
Accuracy	0.33	0.41	0.48	0.56	0.80
F1-Score	0.37	0.47	0.55	0.62	0.77

On the training dataset, the threshold was consequently set as 0.90. These results are shown in Table 5. Much like Task B, in Task C too, cosine similarity along with the *bert-base-nli-mean-tokens* model was used to find out whether the two given summaries were taken from the same transcript or not.

A higher similarity would indicate that indeed took the given two summaries from the same transcript. A similar experiment was done here for Task C for calculating the threshold value, and its results can be seen in Table 6.

Table 6: *Training Dataset on Task C Results*

Threshold	0.60	0.70	0.75	0.80	0.90
Accuracy	0.26	0.40	0.50	0.60	0.79
F1-Score	0.28	0.48	0.59	0.67	0.79

3.1. Results and Analysis

As it can be seen from Table 7 and Table 8, The Turing TESTament performed very well when comparing the summaries with the original transcript because it aims to maximize the heuristic against the provided transcript. However, against the provided summaries, the current state-of-the-Art models such as BERT, Pegasus, and T5 provided better results, and The Turing TESTament system lagged. The System aims to give a summary without access to any baseline summary. This has been discussed further in the next section. Table 9 shows the Accuracy of the Submitted System. In the above table, Cosine Similarity was able to predict nearly half of the summaries and transcripts correctly.

Table 7: *Average Rouge and Meteor scores against given transcripts*

Model	Rouge1	Rouge2	RougeL	Meteor
BERT	0.20	0.20	0.21	0.07
Pegasus	0.20	0.18	0.19	0.06
T5	0.17	0.15	0.15	0.05
The Turing TESTament	0.49	0.47	0.48	0.06

Table 8: *Average Rouge and Meteor scores against given summaries*

Model	Rouge1	Rouge2	RougeL	Meteor
BERT	0.26	0.05	0.11	0.22
Pegasus	0.27	0.06	0.11	0.23
T5	0.31	0.08	0.14	0.24
The Turing TESTament	0.17	0.05	0.09	0.06

Table 9: *Task B and Task C Results*

	Task B	Task C
Correct Predictions	400	748
Total Predictions	972	1431
Accuracy	0.411	0.522

4. Conclusions

As the previous section showed, we can use feature engineering for automatic minuting depending on the selected features used. Moreover, using a heuristic with scores other than Rouge for different types of meetings and texts can also create different summaries.

Thus, the System is flexible enough to accommodate changes based on different types of texts. Moreover, as the System aims to provide the best set of sentences from any given text based on the heuristic used, it can be used as a preliminary condenser to remove unwanted sentences from a text.

4.1. Limitations and Future Work

However, it is not to say that this approach does not have its limitations.

1. The biggest limitation of this System would be that the features to use and the method to rank the individual sentences would be at the user's discretion.
2. In the present System, the sentences extracted to create the summary often lead to a high word count. This is because only one speaker sometimes speaks in a meeting or

conversation, leading to longer sentences than average. Therefore, adding a word count hyperparameter can lead to short summaries.

Further work would be based on the above-mentioned limitations and would aim to answer the following questions:

1. How can different heuristics, metrics, and features affect the generated summary?
2. How this System performs in conjunction with a pre-trained State of the Art model like T5? Would provide the generated summary as input to the T5 model yield better results?

Thus, the System can also be used as a layer for different deep learning models, improving the time and efficiency of finding a summary as the number of sentences will reduce.

5. References

[1] 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

[2] C.-Y. Lin, "Rouge: a package for automatic evaluation of summaries," in *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*, July 2004, pp. 74–81. [Online]. Available: https://www.microsoft.com/en-us/research/publication/rouge-a-package-for-automatic-evaluation-of-summaries/

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003. [Online]. Available: http://portal.acm.org/citation.cfm?id=944937

[4] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909

[5] V. Tiwari, P. K. Jain, and P. Tandon, "An integrated shannon entropy and topsis for product design concept evaluation based on bijective soft set," *Journal of Intelligent Manufacturing*, vol. 30, no. 4, pp. 1645–1658, Apr 2019. [Online]. Available: https://doi.org/10.1007/s10845-017-1346-y

[6] J. Wang, H. Zhang, C. Zhang, W. Yang, L. Shao, and J. Wang, "Generating an overview report over many documents," *CoRR*, vol. abs/1908.06216, 2019. [Online]. Available: http://arxiv.org/abs/1908.06216

[7] F. Ahmad, Y. ., and A. Ahmad, "Automatic summarization of textual document," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, pp. 2486–2491, 09 2020.

[8] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *CoRR*, vol. abs/1908.10084, 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

6. Generated Samples

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

- And whose- (PERSON2) Oh, ok.
- (PERSON1)[PERSON3] is not coming?(PERSON1) Aha, she is ill.
- (PERSON2) Yea, yeah, yeah, right.(PERSON1) Oh yey, what is, is it Covid?(PERSON2) Oh, [PERSON3]joined.(PERSON1) Yeah, [PERSON3]joined.(PERSON3) Oh, actually I 'm-
- (PERSON1) Oh, yey.(PERSON3)¡unintelligible¿ since-
- (PERSON2)¡unintelligible¿ list maybe¡unintelligible¿.(PERSON3) Yeah, anyway was just lying down I can hear- I'm getting bored actually now.
- To, [PERSON1] do you have anything specific to discuss today?(PERSON2) Ok, so [PERSON1] I was like looking through the task that we need to do.So, Ah I have connected to the person whom [PERSON4] reffered.And he reverted to me back with a link and say that you can do it there.So I'm get to see that how it looks like to do with the [ORGANIZATION1], ah, website.What I was wondering like if we can quickly do it on a Github or something.So, let me first look into that um, like, how does it works, uh, to do with¡unintelligible¿and something.¡unintelligible¿ do it in Github and do it very quickly.
- (PERSON2)¡unintelligible¿ like do much more timing to, into these things.So one thing is setting up the website,but the website should have, uh, the content like, ah, the call for participations.
- (PERSON1) So the website of Ah, the¡unintelligible¿, of the, of the special event.Like, let me write it first version, and then, um, like maybe you, [PERSON3] and [PERSON4] should comment on that.Yeah, so I will share with both of you a¡unintelligible¿ e-mail for that we will resending to people we know and we want to-
- (PERSON1) I don't know -I don't have any, I don't have any connections in this summarization field.(PERSON2) No, no, not require, not require [PERSON1], not require to like only for summarization, ok?(PERSON1) But how-
- (PERSON2) We have to show we have a good program committee.(PERSON2) Like no one is going to check, you know.¡laugh¿
- (PERSON2) Yeah.(PERSON1) They will know that they are good, but at the same time they will not know that they are summarization people.So even big and like agree to do uh review at least one or two papers.
- (PERSON2) You can just ask them, because eventually, like finally, it depends on how many papers we are getting.And then we can invite more people based on how many papers we are getting, okay?So that is¡unintelligible¿ requirement, but for in the website we have to show at least, I will say approximately around more than 20 people in the program committee.So like

we have many good uh contexts, like, for example, from [ORGANIZATION1] only we can have two or more than three or four people.

- (PERSON2) And so let us let me write the draft e-mail and share it with you.Just Google Docs is kind of uh it does not work today.
- (PERSON1) I don't know, but [PERSON4] told that today Google doesn't work.¡laugh¿
- (PERSON2) Yes, I can also see that.Uh so the thing is like a for the website we need this thing uh to be ready.So, he is yet to respond back that we should do the individual membership, or we should do institutional membership so that we have to know what to do.(PERSON1) I think I 'm a little bit lost and¡unintelligible¿.(PERSON2) Yeah, so we have to also like get the¡unintelligible¿membership right.So uh for that, I have mail to that-
- (PERSON1) Aha, okay, I see on my¡unintelligible¿, ok.
- (PERSON2) And I was looking at the, into some challenges of for the last year and their websites.(PERSON1) Ok.
- (PERSON2) Wait for the draft and, yeah.And uh maybe like uh, I will also copy on to it so that he can also invite some more people into the program committee.
- (PERSON1) Definitely, definitely um copy to [PERSON4].Because, yes he is the person who could have know about somebody.Uh it seems a lot of work to do.Because, you know, like um, usually in workshop, like we used to have a big organizing committee, and some people are from¡unintelligible¿.
- You're also, you're also not from the summarization background?So that is, anyway, I did not thought like, um, like making a share task would be that much challenging.Well, I¡unintelligible¿ that it ischallenging since why I was so much unsure about making workshop that.
- (PERSON2) Yeah, the actually, like we are also doing uh some workshops with¡unintelligible¿.But I, like I am mostly looking into the like publicity and the website.And also a bit on writing the paper uh for that uh, but I did not like, um,¡unintelligible¿ share task I guess it there is so many things that we need to take care of.So that is why we have to focus on-
- (PERSON1) Yes, we don't have, we don't, we don't actually have scientific input.(PERSON2) Yes and I believe, like if to make it a workshop we definitely need at least one or two people who have a background in summarization.Because originally, we planned and make a work, a shirt task.It means to give the participants a totally new task and asked them to make mixed experiment and then compare the results.I really, I can 't help beat frustrated how it is it is possible to do.
- (PERSON2) Yeah, I think uh like uh, even if we call it a workshop, but it will be mostly challenge specific.But it is also actually not a workshop it's a satellite event.At the satellite event maybe¡unintelligible¿
- (PERSON2) Right, right, that's the exact word.¡unintelligible¿ not call it a workshop.(PERSON1) Yes, let's not call it a workshop.Uh so, but uh like they did not like, they didn't like responded affordably so they were actually not available.
- (PERSON2) Yes, yes, many, me and [PERSON4] we, we addressed at least six to seven people from summarization community.(PERSON2) But they are very, like they are very famous professors, including [PERSON6],¡unintelligible¿ and-
- (PERSON1) And they did not answer?(PERSON2) No, they did answer, but they said, they were very busy.(PERSON1) Aha, ok.
- (PERSON2) But the thing is that we ¡unintelligible¿to like the like what unquote very famous people from some relation who were like, you know.
- (PERSON2) No, I 'm going to ask him to be a part of the program committee because he declined to be a part of the organizing committee.
- (PERSON1) Yeah,¡unintelligible¿ is definitely a good idea.He should be, if he agrees, he is a good part of an organizing committee.But feel uh [ORGANIZATION2] is is organizer, so it does not make sense, like someone from [ORGANIZATION1], right?(PERSON2) As a¡unintelligible¿ does it make sense like someone from [ORGANIZATION1] to deliver the keynote?So if it is somebody, why not if it could be, if it is an authoritative person who knows about the topic, but I do not know the person.
- (PERSON1) As as I think about summarization, I knew just a couple of very young students who began to do the year.Mm, so let me ask, let me ask [PERSON4]¡unintelligible¿.(PERSON1) Should I, should you, do you want to ask him now?(PERSON2) Ok.
- (PERSON1) No, he's in the next room.(PERSON2) I think just let me-
- (PERSON1) Ok, let¡unintelligible¿, write him an e-mail, yes.(PERSON2) Yes, I try to figure out what are the other things that we need to take care of.(PERSON2) Yeah,¡unintelligible¿, ok, bye.

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
