IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol. 003, LIIR070
DOI: https://doi.org/10.52731/liir.v003.070

Semi-automatic Summarization of Spoken Discourse for Recording Ideas using GPT-3

Yuki Yoshimura [*] , Shun Shiramatsu [*] , Takeshi Mizumoto [†]

Abstract

Although sticky notes are generally used to record and structure ideas stated in a face-to-face workshop, participants sometimes forget to write down their stated ideas due to the inconvenience. This study aims to apply speech recognition to record the ideas stated in a face-to-face workshop using a model created by fine-tuning the GPT-3 to choose the utterances to be recorded and then paraphrase spoken utterances. In the evaluation experiment, we compared the effects of including preceding and following utterances in addition to the summarizing and paraphrasing target utterance, and demonstrated that including only preceding utterances in addition to the summarizing and paraphrasing target utterance in the model input resulted in an F1 value over 0.8 for the selection of utterances to be recorded and ROUGE-1 up to 0.48 for paraphrasing utterance content.

Keywords: Discussion support, spoken discourse, GPT-3, summarization, paraphrase.

1 Introduction

Although sticky notes are generally used to record and structure ideas stated in a face-to-face workshop, participants sometimes forget to write down their stated ideas due to the inconvenience. In this study, we aim to address this issue by applying audio recognition to record these stated ideas. Our idea is that recording and visualizing the key ideas will facilitate the metacognition of workshop participants because they can check their ideas as they relate to the current context. Moreover, we aim to develop a human-machine collaborative method for recording spoken discourse because we assume that human participant's interpretation is indispensable to make the record easy-to-understand. Specifically, we address the following two issues.

1. **How to choose the utterances to be recorded.** A spoken discourse generally contains many redundant utterances that should be omitted. Therefore, we need to develop an automatic or semi-automatic system to choose which ideas to record.

[*] Nagoya Institute of Technology, Aichi, Japan
[†] Hylable Inc., Tokyo, Japan

2. **How to paraphrase spoken utterances in an understandable way.** A spoken utterance is sometimes difficult to understand without the context. Therefore, we need to develop a system to paraphrase the audio recognition results in a way that can be easily understood.

Section 2 of this paper overviews related research, and Section 3 presents the system we developed to record the ideas stated in a face-to-face workshop. In Section 4, we report the results of evaluation experiments and discuss their relevance. We conclude in Section 5 with a brief summary and mention of future work.

2 Related Works

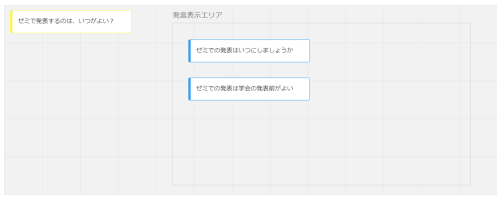
Our system employs Miro [1], an online whiteboard, and Semantic Authoring platform [2][3] for visualizing a structure of spoken discourse because we assume that semi-automatic process with human-machine collaboration is required for recording understandable meeting minutes. Miro and Semantic Authoring can be platforms for such human-machine collaboration. Although there exist several studies about summarization of spoken meeting records using automatic speech recognition, human-machine collaboration we are focusing on has not been considered. For example, Song et al. proposed a CNN-based system for automatic meeting transcription, summarization called SmartMeeting [4], Koay et al. proposed a BART-based summarization system with a sliding-window approach [5] and Alexandra et al. proposed a phrasal query-based summarization system with BART[6]. Although these studies dealt with similar target to our one, they did not dealt with human-machine collaborative summarization of spoken discourse.

3 Proposed System

3.1 User Experience and System Structure

The system proposed in this work is an improved version of the one we previously developed [7]. It is intended to be used for group discussions in workshops. Its user experience is shown in Figure 1. The workshop organizer writes the agendas on yellow cards of the Miro, before the group discussion begins (Figure 1(a)). Facilitator in each group facilitates the discussion, referring to the prepared agendas. In addition, egg-shaped recorders from Hylable Inc. are placed at each group's table or other discussion location to acquire speech signals of each group. As the discussion begins, the summarized and paraphrased text of each utterance is wrote on blue cards of Miro and displayed on a frame, named "the summarized and paraphrased text display area", of the Miro(Figure 1(b)). The operator of each group, who does not participate in the discussion, takes the summarized and paraphrased text displayed on it and connects it to the agendas or other summarized and paraphrased text (Figure 1(c)). Texts not selected by the operator disappear on Miro after a certain amount of time (Figure 1(d)). The operator repeats this process to structure the discussion and it helps discussion participants look back(Figure 1(e)).

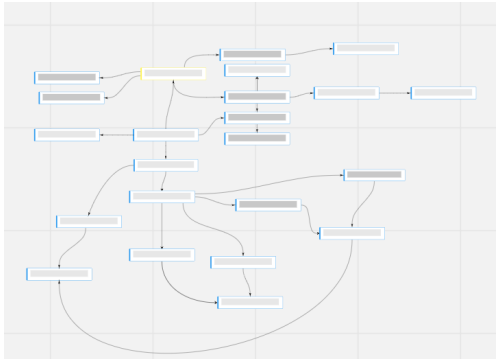
(a) Preparation of the agenda



(b) Display of summarized and paraphrased texts

(c) operator selects summarized and paraphrased text and connects it to the agenda.

(d) Summarized and paraphrased texts not selected disappear.



(e) Structured Discussion

Figure 1: The user experience of the system

System structure is shown in Figure 2. The speech signals acquired using an egg-shaped recorder is converted into text using Microsoft Azure's Speech to Text[8]. This text is speech recognition results and it is summarized and paraphrasing using fine-tuned GPT-3[9]. This text is used as Summarized and Paraphrased Text and displayed on Miro using REST API of Miro.

3.2 Summarizing and Paraphrasing

3.2.1 Summarizing and Paraphrasing Speech Recognition Results

The speech recognition results include speech recognition errors, fillers such as "uh"' "um," and "well," and remarks unrelated to the content of the discussion such as greetings and short responses like "OK." and "I see". Since displaying all such text would be too heavy a load on the operator, we reduce it by excluding in advance information that is not necessary, by summarizing and paraphrasing the statements and then display it.

Table 1 shows examples of summarized and paraphrased speech recognition results (tasks of the summarizing and paraphrasing model are presented later). Since the discussion

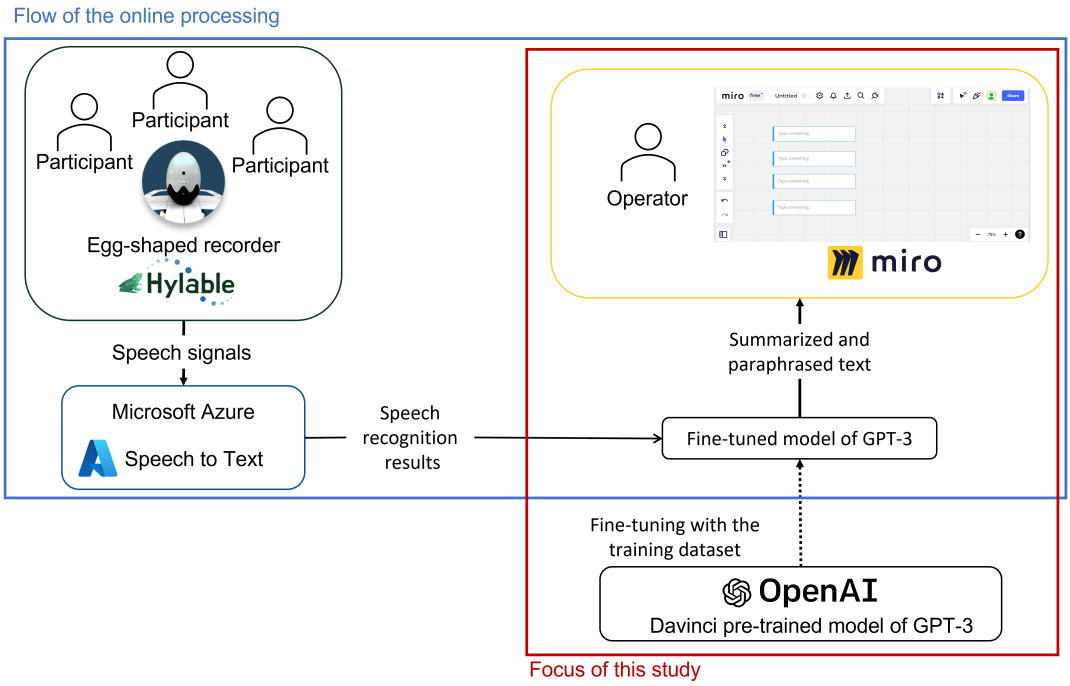


Figure 2: Structure of proposed system.

was in Japanese, we present Japanese-language text here and provide, an English translation in each row of the table.

Table 1: Examples of summarized and paraphrased speech recognition results.

Participant	Speech	Speech recognition results	Summarized and Paraphrased Text
A	ゼミでの発表は いつにしましょうか。 (When is a good time to present at a seminar?)	ゼミでの発表は いつにしましょうか。 (When is a good time *to present at a seminar*?)	ゼミでの発表は いつにするか。 (When is a good time to present at a seminar?)
B	そうですね。 (Well.)	そうですね。 (Well.)	[Null] [Null]
	うーん、学会の 発表前がいいですね。 (Uh, I think it is better to do that before the conference presentation.)	うーん、国会の 発表前がいいですね。 (Uh, I think it is better to do that before the announcement of the congress.)	ゼミでの発表は、 学会の発表前がよい。 (*Presentations at the seminar* should be made before the conference presentation.)

As shown in Table 1, discussion participants A and B had the following conversation.

A When is a good time to present at a seminar?

B Well. Uh, I think it is better to do that before the conference presentation.

Now, suppose "When is a good time to present at a seminar?" is one utterance and "Well. Uh, I think it is better to do that before the announcement of the congress." is divided into two utterances, "Well." and "Uh, I think it is better to do that before the announcement

of the congress.". This one utterance is appropriately separated by Azure's Speech to Text from the speech recognition results (note that it is not always divided into each sentence).

Let's take a close look at each speech recognition results starting with "When is a good time to present at a seminar? ". In the English translation, there is no difference between the speech recognition results and the Summarized and Paraphrased Text, but in Japanese, the polite-form is changed to the non-polite form, as this is generally more suitable for summarizing and paraphrasing. Next, regarding "Well", it is a filler and therefore does not need to be displayed on the whiteboard. In this case, the output of the summarizing and paraphrasing model will be "Null." Like this, utterances that do not include ideas to be recorded, e.g., fillers, interjections, backchannels, and greetings, are given a "Null" label and are not displayed on the Miro. The final part is "Uh, I think it is better to do that before the announcement of the congress.", where we can see that this speech recognition results obviously had a speech recognition error during the conversion from the speech signals. This error probably occurred because "conference" is pronounced as "gakkai" in Japanese and "congress" is pronounced as "kokkai", which are quite similar. Therefore, as with the second utterance, the filler is removed, and then the speech recognition error is corrected and summarized and paraphrased. The missing subject, "presentation at the seminar," can be inferred from the previous utterance, so it is also added. Conversely, with the first utterance "When is a good time to present at a seminar? ", "when" can be inferred from the latter utterance. Whether to include both or only one of the preceding and following utterances will be compared in the evaluation experiment.

The items above demonstrate the tasks performed by the summarizing and paraphrasing model.

- Conversion from polite-form to non-polite form
- Correction of speech recognition errors
- Removal of fillers, interjections, backchannels, and greetings
- Creation of a contextual summary

3.2.2 Summarizing and Paraphrasing Model using GPT-3

We created the summarizing and paraphrasing model by fine-tuning "Davinci", which is currently the best performing GPT-3 model.

An example of training data including both the preceding and following utterances is shown in Table 2 (note that only Japanese data were used in the training). This example is a continuation of the utterance in Table 1, where the summarizing and paraphrasing target utterance is set to "Uh, I think it is better to do that before the announcement of the congress.". The reason the summarizing and paraphrasing target utterance is not "Uh, I think it is better to do that before the *conference presentation*" is that the speech recognition results are used as the model input. Also, the length of preceding utterances should be one minute and the following utterances should be 30 seconds. This is because abbreviated words and words suggested by the indicative word are likely to be included in the preceding utterances.

In addition, as mentioned in 3.1, since the utterances that do not include ideas to be recorded, e.g., fillers, interjections, backchannels, and greetings, need to be automatically omitted, such utterances, we assign the "Null" label.

Table 2: Example of training data.

Preceding utterances	ゼミでの発表はいつにしましょうか (When is a good time to present at a seminar?) そうですね。 (Well.)
Summarizing and paraphrasing target utterance	うーん、国会の発表前がいいですね。 (Uh, I think it is better to do that before the announcement of the congress.)
Following utterance	次、学会の発表があるの誰でしたっけ。 (Who has the conference presentation next?)
Summarized and Paraphrased Text	ゼミでの発表は、学会の発表前がよい。 (Presentations at the seminar should be made before the conference presentation.)

4 Evaluation and Discussion

4.1 Experiment Settings

We conducted evaluation experiments using the following three patterns of whether to include both or only one of the preceding and following utterances.

1. Summarizing and paraphrasing target utterance

2. Preceding utterances, Summarizing and paraphrasing target utterance

3. Preceding utterances, Summarizing and paraphrasing target utterance, Following utterances

Data were from a meeting held on May 30th, 2022 for 39 minutes and 15 seconds. There were 179 utterances each in second pattern and third pattern and 175 in first pattern due to elimination of duplicates and the 5-split cross-validation we performed. The hyperparameters were: batch size = 1, number of epochs = 4, and learning rate = 0.1.

The evaluation was based on the following two points.

1. **Rejection** : When there is no need for summarization and paraphrasing, as in the case of fillers, interjections, backchannels, and greeting, the summarizing and paraphrasing model outputs “Null.” Therefore, we evaluated whether these were rejected, as a binary classification (Null or not) problem.

2. **Summarization and paraphrasing** : This is the evaluation of the summarized and paraphrased text when it is determined to summarize and paraphrase. ROUGE-1[10] is used for the evaluation.

We also compared the values of the “temperature” parameter when outputting the summary after the model was created. This parameter takes values between 0 and 1, with higher values being more creative.

4.2 Results and Discussion

First, we present the results of the rejection.

Table 3: The rejection scores.

(a) Accuracy

	Temperature		
	0	0.5	1
target	0.89	0.88	0.90
pre + target	**0.91**	**0.92**	**0.91**
pre + target + post	0.82	0.82	0.79

(b) Precision

	Temperature		
	0	0.5	1
target	0.71	0.70	0.75
pre + target	**0.82**	**0.87**	**0.83**
pre + target + post	0.65	0.64	0.60

(c) Recall

	Temperature		
	0	0.5	1
target	**0.84**	0.81	**0.84**
pre + target	0.81	**0.83**	0.81
pre + target + post	0.67	0.70	0.57

(d) F1

	Temperature		
	0	0.5	1
target	0.75	0.73	0.78
pre + target	**0.81**	**0.84**	**0.81**
pre + target + post	0.76	0.64	0.57

target : Summarizing and paraphrasing target utterance
pre + target : Preceding utterances, Summarizing and paraphrasing target utterance
pre + target + post : Preceding utterances, Summarizing and paraphrasing target utterance, Following utterances

Figure 3 shows the confusion matrix for each temperature, and Table 3 lists the accuracy, precision, recall, and F1 values for each pattern. The confusion matrix shows the results for all data and the scores (accuracy, precision, recall, and f1) are the averages of the five models created by the 5-split cross-validation.

As we can see in Table 3, when the temperature was 0 or 1, the recall was higher for summarizing and paraphrasing target utterance only, but all other scores were highest for the combination of preceding utterances and summarizing and paraphrasing target utterance. In particular, if Summarized and Paraphrased Text is incorrectly "Null", it is not displayed on the Miro and the operator cannot view. Therefore, the combination of the preceding utterance and the summarizing and paraphrasing target utterance is the best because of the highest precision. In this combination, the Recall and F1 values are above 0.8, so the rejection is adequate. In addition, 0.5 is the best temperature for this combination.

Next, Table 4 lists the ROUGE-1 precision, recall, and F1 values for the summary.

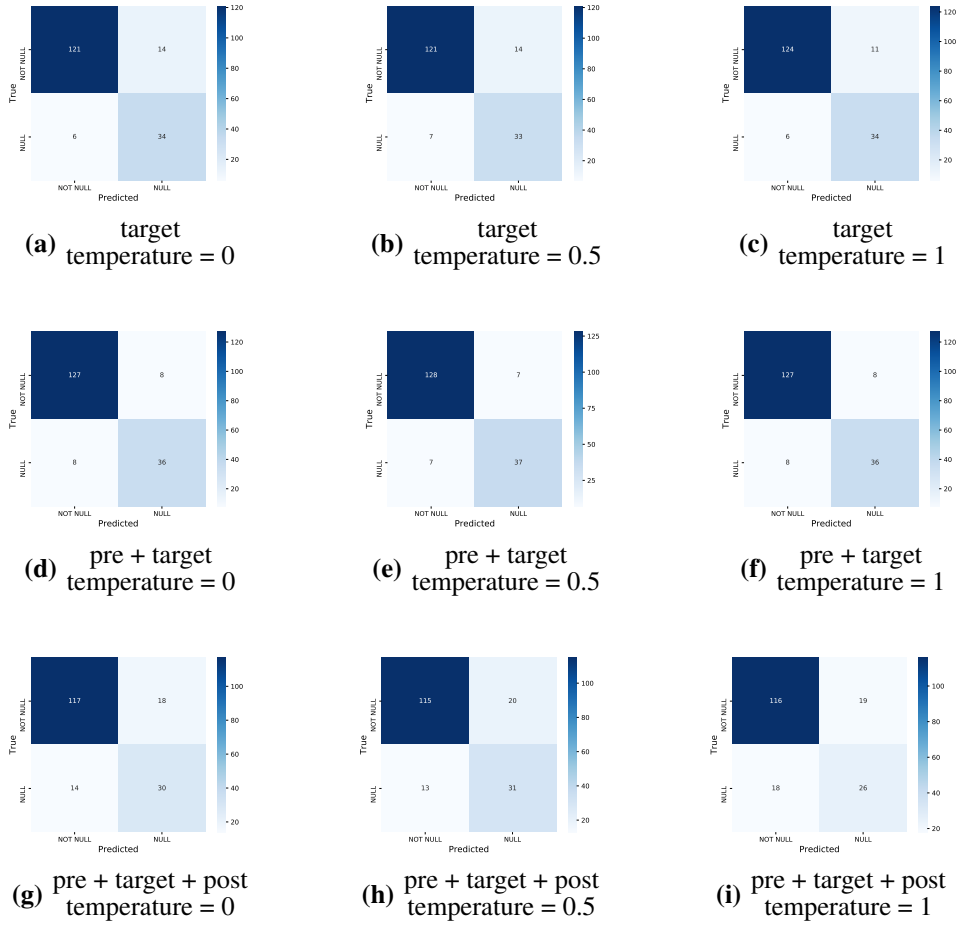


Figure 3: Confusion matrix of the binary classification (Null or not)
target : Summarizing and paraphrasing target utterance
pre + target : Preceding utterances, Summarizing and paraphrasing target utterance
pre + target + post : Preceding utterances, Summarizing and paraphrasing target utterance, Following utterances

Table 4: ROUGE-1 scores.

(a) Precision

	Temperature		
	0	0.5	1
target	0.54	0.53	0.47
pre + target	**0.55**	0.53	**0.49**
pre + target + post	0.54	**0.56**	0.45

(b) Recall

	Temperature		
	0	0.5	1
target	0.43	0.45	**0.44**
pre + target	**0.49**	**0.48**	**0.44**
pre + target + post	0.44	0.44	0.40

(c) F1

	Temperature		
	0	0.5	1
target	0.45	0.46	**0.43**
pre + target	**0.48**	**0.47**	**0.43**
pre + target + post	0.44	0.46	0.40

target : Summarizing and paraphrasing target utterance
pre + target : Preceding utterances, Summarizing and paraphrasing target utterance
pre + target + post : Preceding utterances, Summarizing and paraphrasing target utterance, Following utterances

As we can see in Table 4, summarizing and paraphrasing target utterance only or the combination of summarizing and paraphrasing target utterance plus preceding utterances and following utterances sometimes has a higher score, but overall, similarly to rejection, the combination of preceding utterances and summarizing and paraphrasing target utterance is the best. Considering the possibility that the model-generated summary is more likely to be correct than the human summary, we should focus on the F1 value. Therefore, the combination of preceding utterances and summarizing and paraphrasing target utterance is the best. In the case of this combination, temperature 0 is the best, but considering the rejection score, temperature 0.5 is considered to be the best for use in the system. In future work, it is necessary to adjust the value of temperature.

There are two possible reasons why the combination of summarizing and paraphrasing target utterance plus preceding utterances and following utterances scored lower than the combination of preceding utterances and summarizing and paraphrasing target utterance. The first reason is that the summary is more focused on the following utterances than the summarizing and paraphrasing target utterance. Table 5 shows example of that (note that since the names of people were included, they were replaced with Mr. A and Mr. B.). Int this case, The summarizing and paraphrasing target utterance is about A doing it later and who to do it next, but the Summarized and Paraphrased Text even includes B's internship that was discussed in the following utterances. Second, it is possible that the human summary was not sufficient and that the model-generated summary was correct. However, this is not main reason because it could be caused in any other combinations.

Furthermore, because of using a single meeting as training data, Summarized and Para-

phrased Text sometimes included information that was not included in the input. So we would like to improve this situation by increasing the training data.Table 6 shows example of that In this case, the model input is only the summarizing and paraphrasing target utterance, and the next meeting time is undetermined, but the Summarized and Paraphrased Text contains it.

Table 5: Example of focused on following utterances.

Preceding utterances	じゃえっと今日はえっと、まずは。A さんからえっと？まあ、今の方針というかですね。どんなことをやってるかっていうのを共有してください。はいえっと今、画面の共有がなんか google ドライブにあげてたと思うんですけど、なんか例えば入れなくなっちゃってで見れなくて。だって？ (Let's see, today, let's start with Mr. A. Let's see...? Well, I guess you could say it's our current policy. Please share with us what you are doing. I think I have been sharing my screen on google drive, but I can't see it because I can't access it anymore. Because?)
Summarizing and paraphrasing target utterance	えっとあとできるかわからないんですけどにしてもらってで、あとでやりましょう。じゃあ次はえっと？ (Well, I don't know if I can do it, but let's do it later. Then, what's next?)
Following utterances	そうですね。じゃあ B くんのインターンシップの件はちょっと、もうちょっと詰めていきましょうか？えっと結構あの。まあ、あのフルタイムというかですね。9 時 5 時でえっとお願いするという想定になっています。どうもですね。こちらの大学の一箇所に集められてえーとで、まあオンラインであの研究インターンシップす人るは、そこのある集められた教室で従事するということになるそうです。はい。 (Yes, that's right. So, let's talk a little more about Mr. B's internship, shall we? Let's see, it's pretty much..... Well, we're thinking of asking him to work 9:00 to 5:00, full time. Thank you very much. The interns will be gathered in one place at the university, and the online interns will work in one of the classrooms. Ok.)
Summarized and Paraphrased Text	画面の共有が見れなくなったため、後で行う。次は B くんのインターンシップの件を詰めていく。フルタイムで 9 時か 18 時かでお願いするという想定になっている。大学の一箇所に集められ、オンラインで研究する人はそこのある集められた教室で従事する。 (Will do so later as I can no longer see the screen sharing. Next, we will finalize B's internship. The assumption is that he will be asked to work full time, either 9:00 or 18:00. They will be gathered in one place at the university, and those who do their research online will be engaged in one of the gathered classrooms there.)

Table 6: Example of Summarized and Paraphrased Text including information not included summarizing and paraphrasing target utterance.

Summarizing and paraphrasing target utterance	じゃあ次回決めましょうか？はい。 (Shall we decide next time then? Yes.)
Summarized and Paraphrased Text	次回は 9 月 18 日木曜日の午後 6 時からとする。 (The next meeting will be Thursday, September 18 at 6:00 p.m.)

5 Conclusion

In this study, a model for summarization of spoken discourse was created using GPT-3, with the aim of applying speech recognition technology to record ideas expressed in face-to-face workshops. The model was evaluated in two aspects: selection of utterances to be recorded and summarization and paraphrasing. In the selection of utterances to be recorded, the F1 value exceeded 0.8 when the prior context of the utterance to be summarized and paraphrased was added to the input of the model. In summarizing and paraphrasing, the ROUGE-1 had a maximum F1 value of 0.48. In both respects, the goal of reducing operator burden was achieved. The model of the summary will be improved by increasing the training data and by preparing hyperparameters.

We consider that the discourse structure can increase the F1 value of summarization of spoken discourse. Therefore, we would also like to work on using semantic authoring.

Acknowledgments

This work is supported by NEDO (JPNP20006) and JST CREST (JPMJCR20D1).

References

[1] Miro, https://miro.com/online-whiteboard/, 25 August 2022.

[2] Koiti Hasida“Decentralized, Collaborative, and Diagrammatic Authoring,”, the 3rd International Workshop on Argument for Agreement and Assurance, 2017.

[3] Shun Shiramatsu, Yasunobu Igarashi, “A Preliminary Consideration toward Evidence-based Consensus Building through Human-Agent Collaboration on Semantic Authoring Platform,” , Proceedings of the 15th International Conference on Knowledge, Information and Creativity Support System, 2020, pp. 122-125.

[4] Song, Y., Jiang, D., Zhao, X., Huang, X., Xu, Q., Wong, R.C., and Yang, Q. , “ SmartMeeting: Automatic Meeting Transcription and Summarization for In-Person Conversations.”, Proc. the 29th ACM International Conference on Multimedia., 2021, pp 27772779

[5] Koay, J.J., Roustai, A., Dai, X., and Liu, F. , “ A Sliding-Window Approach to Automatic Creation of Meeting Minutes.”, arXiv, preprint, 26 April 2021, arXiv:2104.12324.

[6] Ganesh, Prakhar and Saket Dingliwal, “Abstractive Summarization of Spoken and Written Conversation.” arXiv, preprint, 5 February 2019, arXiv:1902.01615.

[7] Gen Sato, Shun Shiramatsu, Yuki Yoshimura, Tomoko Omori, Takeshi Mizumoto, “Applicability of Automatic Selection Mechanism for Speech Recognition Results using GPT-3 to Face-to-face Discussion among Elementary Students, Special Interest Group on Crowd Co-creative Intelligence”, Vol.2022, No.CCI-009, 2022, pp. 5-8,

[8] Speech to text, https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/, 25 August 2022.

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei, “Language models are few-shot learners.” , arXiv, preprint, ,28 May 2020, arXiv:2005.14165.

[10] Lin, Chin-Yew, and Eduard Hovy, “Automatic evaluation of summaries using n-gram co-occurrence statistics.”, Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics, 2003, pp.150-157.