




Four Approaches to Developing Autonomous Facilitator Agent for Online and Face-to-Face Public Debate

Shun Shiramatsu^(✉) , Ko Kitagawa, Shota Naito, Hiroaki Koura, and Chao Cai

Graduate School of Engineering, Nagoya Institute of Technology,
Nagoya, Japan
siramatu@nitech.ac.jp

Abstract. Recently, the importance of public debate is increasing both globally and locally for addressing sustainability problems such as pandemics, climate change, and economic crisis. To support such public debate, software agents need to be developed to facilitate discussions, for example, to recommend relevant information by detecting stagnation and flaming in online public debate, to invite debate participants from SNS, or to record face-to-face public debates. In this study, we prototyped four software agents for facilitation: (1) an agent for detecting stagnation and flaming while quantifying the degree of discussion progress in a Web-based debate, (2) an agent for providing relevant information in accordance with the preceding context of a Web-based debate, (3) an agent for finding people who are interested in the content of the discussion and inviting them to a public debate from Twitter, and (4) an agent for recording a face-to-face public debate and supporting users' reviewing of the debate. In this paper, we overview these four agents and evaluation experiments and present the feedback from the participants in an event organized by Facilitation Association of Japan.

Keywords: Discussion facilitation · BERT · Public debate · Civic tech

1 Introduction

Societies worldwide are currently facing various threats to their sustainability, e.g., rapid climate change, the COVID-19 pandemic, natural disasters. Local societies in Japan are also facing sustainability problems such as low birth rate and aging population. In tackling these problems, people need to actively participate in public debate and collaboration.

However, it is not so easy for people to participate in such collaborations because they do not always have enough background knowledge. For example, since hackathons for civic tech activities require diverse participants who has various skills [11], there should be participants who have less background knowledge about the focused on social issues such as IT engineers. Discussion facilitation is

thus important for enabling people to constructively participate in public debates on sustainability issues.

We aim to develop software agents for helping facilitation of online and face-to-face public debates. In this paper, we introduce the following four prototype software agents for facilitating public debate in Japanese.

1. An agent for detecting stagnation, flaming, and deviation from the topic while quantifying the degree of discussion progress in an online debate. To post the facilitator's questions at appropriate times, a facilitator agent needs to detect when the debate is not progressing.
2. An agent for providing relevant information in accordance with the preceding context of an online debate.
3. An agent for finding people who are interested in the content of an online debate and inviting them to a public debate from Twitter.
4. An agent for recording face-to-face debate and supporting users' reviewing of the debate.

The first three agents are for online debate and the last one is for face-to-face debate. This paper overviews the experimental results and presents the feedback from the participants in an event organized by Facilitation Association of Japan (FAJ).

2 Related Works

Online debate systems called COLLAGREE and D-Agree [6, 10] are the basis of this study. Ikeda et al. [5] developed a facilitator agent with a rule-based question generation for online debates on COLLAGREE. However, the timing at which their agent posts the question was not carefully considered. Since their agent just periodically posts the questions, sometimes the agent's posts were excessive. Shibata et al. [10] developed an agent for automated questioning on D-Agree. This agent was used in a social experiment of public debate on the Nagoya City Next Comprehensive Plan in 2018. However, this agent did not consider appropriate timing for automated posts because it just periodically posts the questions.

We have proposed a method to quantify the degree of discussion progress on the basis of the structure of the issue-based information system (IBIS) for online debates in Japanese [8]. However, our previous method considers not the content of a post in the debate but only the node type of IBIS structure extracted from the post. We have also proposed a method to estimate Twitter users' interests and to invite online debate participants from Twitters [1]. However, our previous experiment did not investigate the versatility of the method because the experiment was conducted for only one particular topic of debate.

3 Four Software Agents for Discussion Facilitation

This section overviews the four facilitator agents we prototyped and results of evaluation experiments.

3.1 Agent Estimating the Degree of Discussion Progress

We aim to quantify the degree of discussion progress (DDP) toward the final goal of the debate in order to estimate appropriate timing the facilitator should intervene. To detect such appropriate timing, it is not enough to observe only the number of utterances because even if there are many remarks, they may be the result of flaming or deviate.

We improve our previous IBIS-based method [8] for quantifying the DDP. To consider the content of posts in online debate, we incorporate the bidirectional encoder representations from transformers (BERT) [2]. For the training data of BERT, we used 17 discussion threads in Japanese collected by a social experiment using COLLAGREE in 2013 [6], in which 13 subjects rated the argument progress of each post on a six-point Likert scale from 0 to 5. For each post, we evaluated two types of the DDP: one for the divergence phase of a debate and the other for the convergence phase. Since three annotators evaluated one discussion thread, we averaged them together and normalized the range of the DDP to be $[0, 1]$. These average values are used as reference data for training and testing. This training dataset is used both for IBIS-based calculation [8] and our BERT-based one.

The IBIS-based DDP d_{ibis} is a summation of the weights of IBIS nodes extracted from the preceding debate content. The weight of a node, which is determined only by the IBIS node types (task, idea, merit, and demerit), is optimized by the genetic algorithm [8]. The BERT-based DDP d_{bert} is calculated by regression using BERT. This regression is anomalously implemented on the basis of a BERT model fine-tuned for classifying 6-point Likert scale. Before the fine-tuning, the BERT model is pre-trained using Japanese Wikipedia with SentencePiece [7].

To complementarily use these two calculations of DDP, we define the DDP d as the weighted summation of d_{ibis} and d_{bert} as follows:

$$d = \alpha d_{\text{ibis}} + (1 - \alpha) d_{\text{bert}},$$

where $0 \leq \alpha \leq 1$.

As the evaluation experiment, we calculate the correlation coefficient between the estimated DDP and the reference data. As the result showed in Table 1, the DDP for the divergence phase is accurately estimated by d and d_{bert} . Especially, d with $\alpha = 0.5$ indicates strong positive correlation since $r = +0.69$.

Table 1. Correlation coefficient between the estimated DDP and the reference data (the average of three experiment participants' subjectively evaluated DDP)

Method	Corr (divergence)	Corr (convergence)
IBIS note type	+0.47	+0.30
BERT	+0.62	+0.44
Weighted sum	+0.69 ($\alpha = 0.5$)	+0.42 ($\alpha = 0.1$)

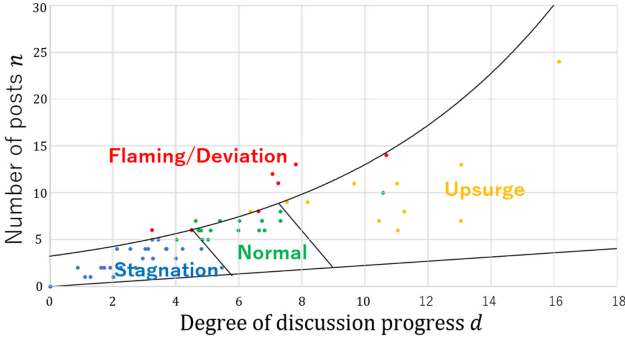


Fig. 1. Classification model for four types of discussion state

Here, we examined how reliable each experimental participant was, since the value of progression in the training data was the average of the three experimental participants. Specifically, we also calculated the correlation coefficients between the mean value of progression and the degree of progression assessed by each experimental participant. The mean correlation coefficient was $r = +0.67$ for the degree of progress in the divergence phase and $r = +0.74$ for the convergence phase. This indicates that the performance of DDP estimation for the divergence phase ($r = +0.69$) is comparable to that of the average human experiment participants. However, it was also suggested that the performance of the convergence phase was significantly inferior to that of humans. This could be attributed to the fact that most of the 17 discussion threads used in the training data did not actually converge towards consensus building.

Furthermore, using the DDP estimation of divergent phases, we prototype a classification model of discussion states shown in Fig. 1. In this figure, the horizontal axis represents the estimated DDP and the vertical one represents the number of recent posts. The plots in the figure represent 65 moments in Slack debates which the agent needs to determine whether it posts some questioning or not. On the basis of the assumption that the facilitator should intervene when the amount of change in DDP is low in relation to the number of posts, this model classifies Web-based debate into four discussion states: Stagnation, Normal, Upsurge, and Flaming/Deviation. The colors of plots in Fig. 1 represent the discussion state manually determined by a human annotator. The experimental results show a precision of 75% in an open test and 89% in a closed test. The model shown in Fig. 1 is obtained by the closed test.

3.2 Agent Providing Relevant Information

When a Web discussion is stagnant, providing information related to an online debate content may help participants to think about what they will post next. In this study, we implement a software agent recommending relevant information for this purpose. To provide relevant information, it is necessary to first determine

the search query from the preceding context and then select the paragraphs and segments to be presented from the Web content obtained by search engines such as Google.



Fig. 2. An example of relevant information provision for a debate in Japanese on Slack (Color figure online)

To determine the search query, the agent calculates the score of the term frequency- inverse document frequency (TF-IDF) of the words appearing in each post by integrating the decay ratio γ and extracts the words with the highest score as the search query. In addition, we also tried to determine the search query by predicting the words that appear in the next statement with BERT. However, the search query determination method by BERT was not adopted because the words at the top of TF-IDF with the accumulated decay ratio γ were more similar to the search queries chosen by human experiment participants than the words predicted by BERT.

Using the extracted queries for Google search, the agent extracts segments as relevant information to provide in the online debate from the top 10 pages of the search result. In this case, we adopted the approach of finding and presenting segments close to the IBIS node type in the search results, assuming possible IBIS node types as a response to the previous statement. Specifically, using the training data also used in the Subject. 3.1, we trained a classifier that predicts the relationship between the IBIS nodes included in the immediately preceding and subsequent utterances with BERT. The relationships we use are classified into five ones: advantages of the recent idea, disadvantages of recent idea, solutions to the recent issue, examples of the recent idea, and reasons for the recent idea. We use this classifier to predict the relationship between recent posts in online debates and segments consisting of four adjacent text sentences in the search

results. The agent chooses the relationship with the most classified segments from the above five relationships, extracts the top three scored segments from the segments classified into the chosen relationship, and presents them as the relevant information (the red dotted frame in Fig. 2).

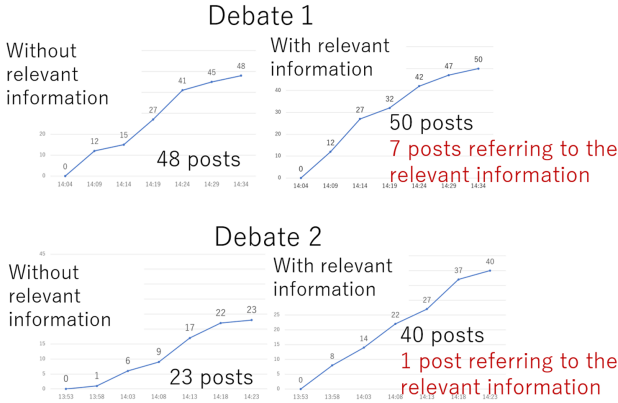


Fig. 3. Comparison of the number of posts between Slack debates with and without relevant information provision

Figure 3 compares the number of posts in online debates with and without the relevant information provision by our method. The theme of debate 1 was “Reduction of food loss,” and the theme of debate 2 was “Ms. Greta Thunberg, a teenage climate change activist.” As can be seen from the figure, there was no significant difference in the number of contributions with and without the presentation of relevant information. This suggests that the effect of providing relevant information has not been verified because the debate does not noticeably stagnate without a longer period of time. Qualitatively, the accuracy of segment selection presented as relevant information needs to be further improved because relevant information was found that was not necessarily in line with the content of the discussion.

3.3 Agent Inviting Participants from Twitter

When conducting an online debate on a social issue, the discussion sometimes stagnates if the number of interested participants is small. To attract more interested people to the Web discussion, we suppose that it is effective to invite them through social networking services (SNS). We have developed a software agent to find Twitter users who are interested in Web discussion topics and generate invitation messages for them [1]. This agent calculates the cosine similarity between tweets and a debate topic using BERT. There are two types of vectors for calculating the cosine similarity: the distributed representation output by the bert-as-service [12] and the output vector of a BERT model for predicting

hashtags from text. The similarity between a tweet and a discussion topic is calculated as a weighted summation of the two kinds of similarities. A Twitter user’s score is defined as a summation of the scores of the user’s top three tweets of the user.

Our previous experiment [1] was conducted for only one particular public debate, i.e., the Nagoya City Next Comprehensive Plan in Japanese, which was conducted in 2018 on an online debate system called D-Agree [10]. For the hashtag estimation, 91 hashtags relevant to the debate topic were prepared. Experiment participants evaluate pairs of a Twitter user and a discussion thread from two aspects: “Is the target user interested in the agenda of the target online debate?” and “Is the target user likely to participate in the target online debate?” As a result, the agent could more accurately estimate first aspect on Twitter users’ interests than the second one on users’ participation possibility.

We conduct an additional experiment on Slack. The debate topic is changed to the privacy protection to investigate the versatility. As a result, we found that enough variety of hashtags is needed for estimating Twitter users’ interests. Moreover, the tendency that the interests is more accurately estimated than the participation possibility was commonly observed. We found that this tendency was influenced by the subjective observation of “even if a Twitter user seems to be interested in the agenda, he/she seems less likely to participate in the debate when the user does behave seriously on Twitter” through interviewing the experiment participants. This finding indicates that the invitation agent should consider not only the target user’s interest in the debate agenda but also the characteristics of the user’s behavior on Twitter.

3.4 Agent Supporting Review of Debate

To promote collaboration and co-creation among a region’s residents, it is important to discuss not only through the Web but also in face-to-face workshops. We aim to develop an agent facilitating face-to-face debate by combining Hylable Discussion [9] and Google Cloud Speech-to-Text [3]. However, Hylable Discussion currently specializes in post-discussion analysis, so it is not possible to obtain the results of analysis in real time during the discussion. For this reason, we first implement a software agent recording face-to-face debates and supporting reviews of the debate for facilitators to reflect on face-to-face discussions. FAJ sometimes conducts “Fishbowl discussion”, i.e., the participants and observers of the discussion are divided and the observers take notes and reflect on the discussion. The user interface generated by this agent has a function similar to that of the observer’s notes, and we aim to make it possible to look back more exploratorily.

Hylable Discussion analyzes the transitions in the volume of each participant’s speech and the tendency of turn-taking on the basis of the results of the auditory scene analysis, i.e., sound localization and sound separation. However, speech recognition is not performed. Therefore, we adopted an approach in which the results of speech recognition by Google Cloud Speech-to-Text from the separated sounds obtained by Hylable Discussion are displayed on a graph of speech

volume transitions. The prototype user interface is shown in Fig. 4. On the left is the transition of the amount of speech for each participant, with the results of speech recognition overlaid on top of it. However, it is not possible to display the results of speech recognition of all expressions for a long discussion, so the important remarks that should be displayed need to be selected.

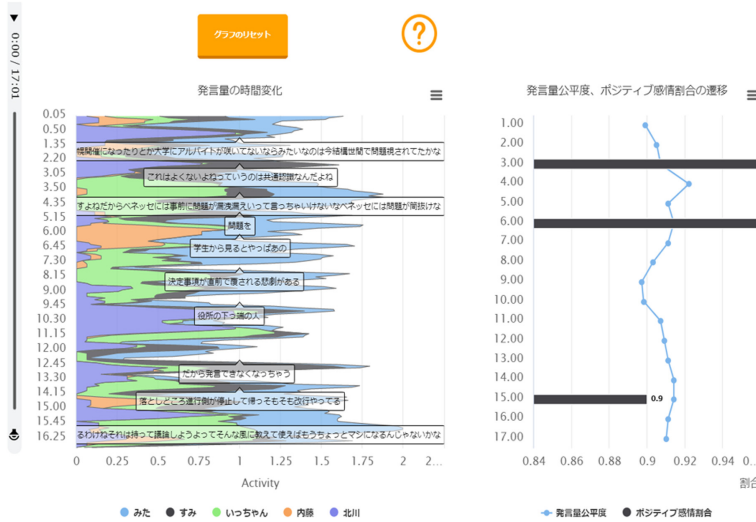


Fig. 4. A review support interface for face-to-face discussions

Through interviews with FAJ facilitators, we learned that facilitators should focus on the process rather than the content of the discussion. The agent automatically selects and displays an utterance at the turning-points where the distribution of the amount of utterances changes significantly. Furthermore, the turning-points in the discussion process are represented by overlaid icons corresponding to the emotions estimated from the phonological information. By clicking on the displayed speech recognition results or icons, discussion participants can listen to the corresponding speech at the corresponding time. Since some speech recognition errors are also included, a mechanism is needed that allows the participants to somehow correct the recognition errors.

In the right side of Fig. 4, the transitions of the fairness of the amount of speech and the transitions of the ratio of positive to negative emotions are shown. This also aims to be used as a visualization method for facilitators to understand the discussion process. Furthermore, we are planning to use this interface for not only facilitators but also people who do not participate in a face-to-face debate to understand the content of the debate.

4 Feedback from Facilitators

We conducted an online workshop on our four prototype agents for discussion facilitation on May 10th, 2020 under the collaboration with FAJ. The online workshop was titled “AI × Facilitation: How far has the research gone? Where should we go using this?” (translated from Japanese) [4]. Over 50 Japanese participants, who were mostly FAJ members, listened to the presentation about the four agents and discussed their potential needs.

The feedback from the participants was written in a Google Spreadsheet in Japanese after the presentation. On the agent estimating DDP, a participant wrote “The timing of interventions usually bothers me,” which represents a need for this agent. Another participant suggested that the DDP can be used for real-time visualization of discussion status. Furthermore, there was a remark pointing out that debates sometimes need to stagnate.

On the agent providing relevant information, a participant wrote “Textbook-wise, it’s good to be able to share the necessary information before the divergence.” From this feedback, we need to consider the appropriate timing to provide relevant information. Another participant wrote that such kinds of search tasks are more suitable for artificial intelligence (AI) than human facilitators.

On the agent inviting Twitter users, a participant wrote “It’s scary when debate trolls are invited in and we get into a bad discussion.” Another participant wrote that actual use is needed to judge whether this prototype agent is useful or not.

On the agent supporting the review of face-to-face debate, multiple participants pointed out the necessity of visual processing for recognizing non-verbal behaviors or emotions of debate participants. Another participant wrote that the user interface for reviewing the debate can be useful in the final stages of consensus building. Furthermore, there was a remark that such a quantitative analysis of the amount of utterance is a suitable task for AI.

5 Conclusion and Future Perspective

We introduced prototypes of three facilitator agents for online debate and one for face-to-face debates. The experiment results showed that the degree of discussion progress (DDP) estimation has a relatively strong correlation with human’s subjective estimation in the divergence phase of debate. The experiment results also showed that the accuracy for providing relevant information needs to be improved. Moreover, longer debate experiments are needed, e.g., several days. The experiment results on the invitation agent indicated that we need to consider not only SNS users’ interest but also their behavior before inviting them. We also prototyped an agent supporting the review of face-to-face debate while finding the turning-point by calculating the distribution of participants’ utterances.

We are planning to improve these facilitator agents in accordance with the feedback on them from the participants in an online workshop organized by FAJ. Especially, to improve the agent for relevant information provision, we are

developing a system for gathering social issues and collaborative activities among people from Web articles. As another future work, since we are currently practicing social distancing due to the COVID-19 pandemic, we are also considering how to develop functions for supporting facilitation on online meeting tools.

Acknowledgement. We appreciate Ms. Kayoko HAYASHI and Mr. Shigeru ICHIKI for managing FAJ’s online workshop and for providing advices on designing the discussion review support agent. We also thank Dr. Takeshi MIZUMOTO, the president of Hylable Inc., for supporting the use of the Hylable Discussion. Moreover, we appreciate Prof. Takayuki ITO for providing the discussion corpora with COLLAGREE and D-Agree. This work was partially supported by JSPS KAKENHI (17K00461) and JST CREST (JPMJCR15E1).

References

1. Cai, C., Shiramatsu, S.: Estimating Twitter user’s interest for inviting potential participants into Web-based debate platforms based on BERT. In: Proceedings of the 2nd International Conference on Artificial Intelligence in Information and Communication, pp. 602–607 (2020)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding, pp. 4171–4186 (2019)
3. Google: Speech-to-text conversion powered by machine learning. <https://cloud.google.com/speech-to-text>
4. Hayashi, K., Ichiki, S.: 175th FAJ Chubu Branch Extraordinary Regular Meeting: “AI × Facilitation” (2020). <https://www.faj.or.jp/base/chubu/event/2020510aka/>. (in Japanese)
5. Ikeda, Y., Shiramatsu, S.: Generating questions asked by facilitator agents using preceding context in web-based discussion. In: Proceedings of the 2nd IEEE International Conference on Agents, pp. 127–132 (2017)
6. Ito, T., Imi, Y., Ito, T., Hideshima, E.: COLLAGREE: a facilitator-mediated largescale consensus support system. In: Collective Intelligence 2014 (2014)
7. Kikuta, Y.: BERT pretrained model trained on Japanese Wikipedia articles (2019). <https://github.com/yoheikikuta/bert-japanese>
8. Kitagawa, K., Shiramatsu, S., Kamiya, A.: Developing a method for quantifying degree of discussion progress towards automatic facilitation of web-based discussion. In: Lujak, M. (ed.) AT 2018. LNCS (LNAI), vol. 11327, pp. 162–169. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-17294-7_12
9. Matsuoka, M., Mizumoto, T.: Toward a better discussion in English: quantitative perspective of feedback. In: Extended Summaries of the 26th Korea TESOL International Conference, p. 54 (2018)
10. Shibata, D., Moustafa, A., Ito, T., Suzuki, S.: On facilitating large-scale online discussions. In: Nayak, A.C., Sharma, A. (eds.) PRICAI 2019. LNCS (LNAI), vol. 11671, pp. 608–620. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29911-8_47
11. Shiramatsu, S., Tossavainen, T., Ozono, T., Shintani, T.: Towards continuous collaboration on civic tech projects: use cases of a goal sharing system based on linked open data. In: Tambouris, E., et al. (eds.) ePart 2015. LNCS, vol. 9249, pp. 81–92. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22500-5_7
12. Xiao, H.: bert-as-service documentation. <https://bert-as-service.readthedocs.io/>