Timing Prediction of Facilitating Utterance in Multi-party Conversation

Tomonobu Sembokuya[1] and Kazutaka Shimada[1]

Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Iizuka Fukuoka 820-8502 Japan
shimada@pluto.ai.kyutech.ac.jp

Abstract. Supporting consensus-building in multi-party conversations is a very important task in intelligent systems. To conduct smooth, active, and productive discussions, we need a facilitator who controls a discussion appropriately. However, it is impractical to assign a good facilitator to each group in the discussion environment. The goal of our study is to develop a digital facilitator system that supports high-quality discussions. One role of the digital facilitator is to generate facilitating utterances in the discussions. To realize the system, we need to predict the timing of facilitating utterances. To apply a machine learning technique to our model, we construct a data set from the AMI corpus, first. For the construction, we use some rules based on the annotation of the corpus. Then, we generate a prediction model with verbal and non-verbal features extracted from discussions. We obtained 0.75 on the F-measure. We compared our model with a baseline method. Our model outperformed the baseline (0.7 vs. 0.5 on the AUC value). The experimental results show the effectiveness of our model.

Keywords: Multi-party conversation · Timing Prediction · Facilitation.

1 Introduction

In collaborative work, people need to discuss several topics for decision-making on a meeting, namely multi-party conversation. It is a very important task in intelligent systems to support consensus-building in conversations with multiple participants. Participants in a discussion often struggle to identify the most suitable solution for a decision on a meeting agenda because there are generally many alternatives and criteria related to making a decision. As a result, they often fail to make a satisfying decision. It leads to the failure of the discussion. To conduct smooth, active and productive discussions, they need an effective facilitator who controls the discussion appropriately. However, it is impractical to assign an effective facilitator to each group in the discussion environment due to a lack of human resources. Although a project manager needs to appropriately handle a discussion in business meetings, he/she might not have remarkable facilitation skills, such as asking questions to gain additional information and asking follow-up questions to further expand participants' understanding. Ordinary people in a group discussion might subconsciously need help from others to

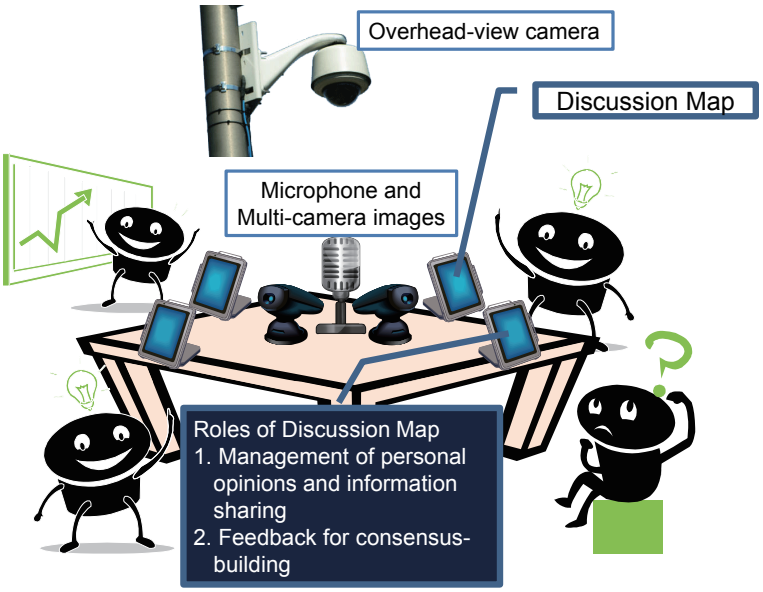

Fig. 1. Overview of our digital facilitator system with discussion maps. The goal of our study is to develop a system that behaves like a facilitator by using several modalities, such as speech and image inputs.

generate a good decision. Therefore, a system that supports consensus-building plays a very important role in discussion.

The goal of our study is to construct a system that cooperatively supports consensus-building and management of conversation for high-quality discussion. The system is referred to as a digital facilitator: a collaborative agent for participants of discussions. Figure 1 shows the overview of our system. We are developing a prototype system to support real discussions [6]. The system estimates the current state of a discussion and then generates sentences and charts that describe it. This is a part of our digital facilitator system. However, the generation timing in the current system depends on participants' clicks on the system: a passive control of the system. Therefore we need facilitator's knowledge, behavior, and patterns to realize a good digital facilitator: an active control from the system.

In this paper, we focus on timing about intervention or facilitation by the digital facilitator. We define an utterance by a participant that behaves like a chairperson on the discussion as "Facilitating Utterance." We propose a prediction model of the timing of such utterances by using a machine learning technique. The contributions of this paper are as follows:

- We design a guideline for constructing training data from the AMI corpus for a timing prediction task of facilitating utterances. It is based on dialogue act tags and social role tags in the corpus.
- We propose a timing prediction model using verbal and non-verbal features for facilitating utterances. We compare the effectiveness of the features experimentally.

2 Related Work

Shiota et al. [14] have reported an analysis of characteristics of facilitators in two multi-party conversations corpora: the AMI corpus [3] and the Kyutech corpus [19]. In the analysis, they generated decision tree models to classify each participant into a facilitator and a non-facilitator in the corpora. Omoto et al. [11] have reported the analysis of facilitating behavior of the exemplary facilitator from measured non-verbal and para-linguistic data. They defined four actions for facilitation: convergence, divergence, conflict, and concretization, and then analyzed the conversations on the basis of these factors. However, these studies analyzed conversations from a macro perspective. We need to determine the timing of facilitation as a function of our digital facilitator system, namely a micro perspective.

Lala et al. [7] have proposed an approach to attentive listening, which integrates continuous backchannels with responsive dialogue to user statements to maintain the flow of conversation in spoken dialogue tasks. They constructed a prediction model based on a logistic regression approach. The task is that a backchannel would occur in 500ms or not. In addition, they improved their system by incorporating a statement response model on the four different response types and a flexible turn-taking model. They evaluated the system with the autonomous android, Erica, as a pilot study. Skantze [16] has proposed a turn-taking model using LSTM for spoken dialogue systems. The model predicted the speech activity for an upcoming fixed time window. They also evaluated how the hidden layer in the network can be used as a feature vector for turn-taking decisions in human-robot interaction data. The target of these studies is a dialogue with two persons. On the other hand, our task is to predict facilitating utterances in multi-party conversations and discussions.

3 Data Construction

We need a data set for a prediction model based on machine learning. For the purpose, we utilize the AMI corpus and the tag sets.

3.1 AMI Meeting Corpus

The AMI corpus [3] is one of the most famous meeting corpora. It consists of scenario and non-scenario meetings. In this paper, we handle scenario meetings.

In the scenario task, participants pretended members in a virtual company, which designs remote controls. Each participant played each role: project manager (PM), industrial designer (ID), user-interface designer (UI), and marketing expert (ME).

The AMI corpus contains numerous annotations, such as topic tags and dialogue acts. In this paper, we focus on the dialogue act tags. The dialogue acts denote speakers' intentions, such as "inform" and "backchannel." The number of dialogue act tags is 15.

Some researchers annotated social role tags for 59 meetings on the scenario portion of the AMI corpus [12, 17]. Each meeting was segmented into short clips by long pauses: pauses longer than 1 second. One social role was assigned to each speaker in each segment by annotators. Each annotator for the tagging was asked to watch the entire video segment and assign a speaker to a role on the basis of a list of specified guidelines. The number of social role tags is five, and the roles are summarized as follows:

- Protagonist: a speaker that takes the floor, drives the conversation, asserts its authority, and assumes a personal perspective.
- Supporter: a speaker that shows a cooperative attitude demonstrating attention and acceptance as well as providing technical and relational support.
- Neutral: a speaker that passively accepts ideas from the others without expressing his/her ideas.
- Gatekeeper: a speaker that acts as a group moderator. He/she mediates and encourages the communication.
- Attacker: a speaker who deflates the status of others, expresses disapproval, and attacks other speakers.

In this paper, we handle the 59 meetings with the social roles.

3.2 Facilitating Utterance

Our task is to predict the timing of facilitating utterances in conversations. Therefore, we need to determine which utterances correspond to facilitating utterances in the conversations to apply a machine learning method for the prediction. For the purpose, we configure tree-based rules to determine the facilitating utterances. Figure 2 shows the flowchart of the determination process.

First, we check whether each utterance is spoken by a participant with the Gatekeeper tag. If so, move to the next step. If not so, we regard the utterance as a Non-facilitating utterance. From the definition in Section 3.1, the Gatekeeper tag is the most important factor for the judgement of facilitating utterances. Although project managers (PM) often have a similar role with Gatekeeper in discussions, we focus on the Gatekeeper tags only. The reason is that participants with other roles (ID, UI, and ME) often behave like a facilitator.

Next, we focus on specific dialogue act tags. Table 1 shows examples of the dialogue act tags in the AMI corpus. We use "Inform", "Suggest", "'Offer", and "Elicit-*" for the determination. We can correctly remove utterances with these tags as backchannel utterances of Gatekeepers, e.g., "Uh, I see."

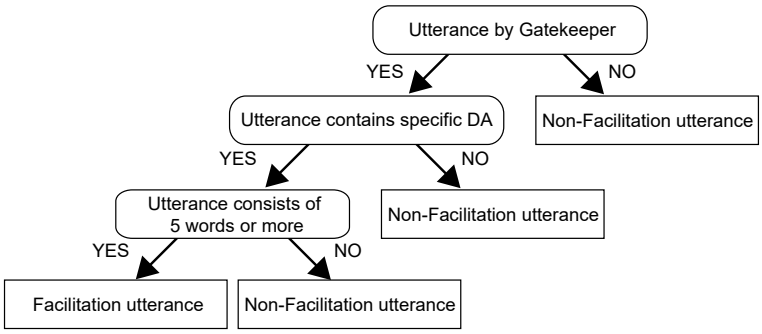

Fig. 2. Flowchart for the determination process. The data set is constructed by using social role tags and dialogue act tags in the AMI corpus.

Table 1. Dialogue acts on AMI meeting corpus.

DA tag	Meaning
Backchannel	Response such as "yeah"
Stall	Filled pauses
Fragment	Utterance that does not convey a speaker intention
Inform	Giving information
Suggest	Expressing an intention relating to the actions of another individual, the group as a whole, or a group in the wider environment
Offer	Expressing an intention relating to his or her own actions
Assess	Comment that expresses an evaluation
Elicit-*	Requests about the DA; e.g., if * is "Inform", it denotes a request that someone else give some information.

Finally, we select utterances with five words or more as facilitating utterances. By using this rule, we can remove utterances for giving information from the facilitating utterance list, e.g., the utterance "No." with the Inform tag. The threshold, five or more, was determined experimentally[1].

4 Method

We explain our method and the task in this section. Figure 3 shows the overview of our method. The 1st and 14th utterances with the orange color in the figure are instances of facilitating utterances by the rules in Section 3.2.

We regard utterances within S_p sec. from the current utterance as one cluster. Then, we assign the "+1" label to the cluster that contains a facilitating

[1] Five is the mode value of utterances with the Gatekeeper and specific DA tags.

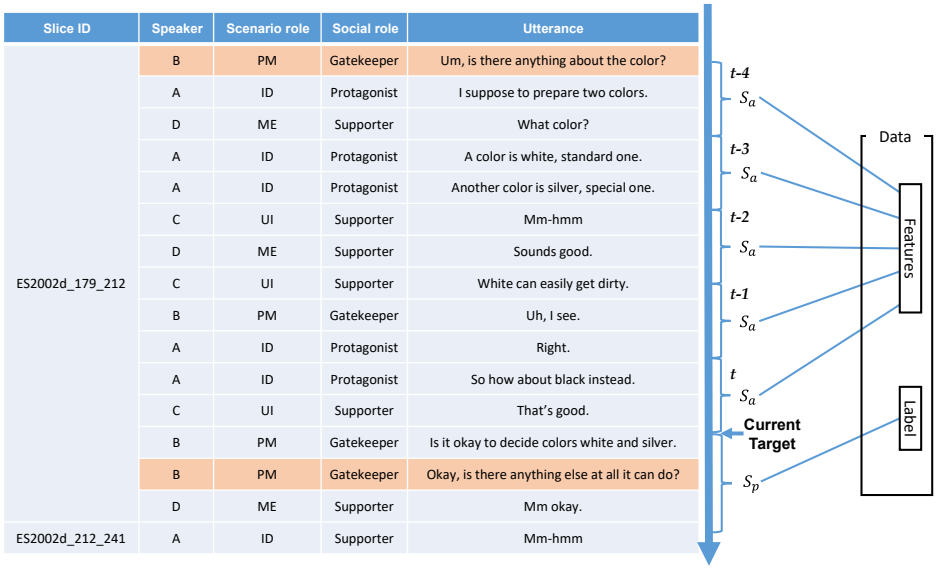

Fig. 3. Overview of our timing prediction model. Our model extracts features from the previous utterances in each S_a and concatenates the features for the prediction model. The label denotes +1 if the S_p contains a facilitating utterance and -1 if the S_p does not contain a facilitating utterance.

utterance, otherwise -1. Our model predicts the label by using features extracted from utterances in the previous S_a range. We concatenate features of five S_a.

We use Support Vector Machines (SVMs) as the classifier. We utilize LIBSVM [4] for the implementation. The parameters are default settings, and the kernel is RBF. For SVMs, we extract the following features:

f1) Average of word embedding
The word embedding is a vector representation of each individual word which is pre-trained by some of the syntactic and semantic relationships in the language [9]. We use a model trained from Wikipedia and Web news. We utilize fastText [2] for the implementation. We calculate the average value of the embedding vectors of words that appear in S_a.

$$Emb_{ave}(t, S_a) = \frac{\sum_{x_i \in Words(t,S_a)} Emb(x_i)}{S_a} \quad (1)$$

where $Emb(x_i)$ is the embedding of a word x_i based on fastText. t denotes time in the discussion and the unit of S_a is seconds.

f2) Average of words in each S_a
Participants tend to frequently utter his/her thoughts and opinions in heated

discussion. They also tend to not frequently utter his/her thoughts and opinions in non-hated discussion. It indicates that the number of utterances is one important feature for the prediction. Here, we utilize the average number of words in S_a as a feature for the prediction model.

$$NumWords_{rate}(t, S_a) = \frac{NumWords(t, S_a)}{S_a} \tag{2}$$

where $NumWords$ is the number of words in S_a in t.

f3) Ratios of overlap and silence
In a similar situation to f2, overlaps and silences occur in discussion. To capture the tendencies of each participant about the two characteristics, we introduce ratios of overlaps and silences as the features. The feature values are also the average values of overlap length and silence length in S_a.

$$SilentTime_{ratio}(t, S_a) = \frac{\sum_{sl_i \in Silents(t, S_a)} sl_i}{S_a} \tag{3}$$

$$OverlapTime_{ratio}(t, S_a) = \frac{\sum_{ol_i \in (t, S_a)} ol_i}{S_a} \tag{4}$$

where sl_i and ol_i are silence length and overlap length in S_a, respectively.

f4) Number of long silences
Long silences often indicate non-heated discussion as compared with short silences. Therefore, we detect silences that are longer than a threshold and then utilize the frequency of the long silences as the feature[2].

$$NumSilents_T(t, S_a) = \sum_{sl_i \in Silents(t, S_a)} \begin{cases} 1 & (sl_i \geq T) \\ 0 & (sl_i < T) \end{cases} \tag{5}$$

where t is a threshold for a long time silence.

f5) Number of speaker changes
In heated discussion, speaker changes occur frequently. Therefore, we utilize the number of speaker changes in S_a as the feature.

5 Experiment

We evaluated our timing prediction model with the dataset described in Section 3. We set $S_p = 30$ seconds because the average time of each segment in social role annotation in the previous studies was approximately 30 seconds. We evaluated the dataset with 10-fold cross-validation. We analyzed our method and the data in terms of types of features, length of S_a, and types of dialogue acts. In addition, we compared our model with a baseline.

[2] For overlaps, we do not handle this feature because the overlap length is usually shorter as compared with the silence length in discussion.

Table 2. Comparison on features.

Features	P	R	F
Verbal	0.74	0.76	0.75
Non-Verbal	0.73	0.55	0.63
All	0.74	0.76	0.75

5.1 Discussion about Features

To discuss the effectiveness of features, we categorize the features into two types: verbal and non-verbal features. Here, the verbal features are f1 and f2, and the non-verbal features are f3, f4 and f5 described in Section 4. We generated three prediction models: a model with verbal features, a model with non-verbal features, and a model with all features. Then, we evaluated the models with precision (P), recall (R), and F-values. We set $S_a = 30$ in this experiment.

Table 2 shows the experimental result. From the result, the non-verbal features were not essentially effective for the prediction. The values of the model with verbal features and the model with all features were the same in all criteria. However, instances that the models predicted incorrectly were not completely the same. Some instances were predicted correctly by using the ratio of overlaps in f3. Therefore, non-verbal features are not always counterproductive to the prediction. We need to discuss more effective nonverbal features through detailed error analysis.

To achieve better accuracy, we need to apply other features that are obtained from speech information for the prediction model. Prosodic features, such as pitch and volume, were often used in studies about participant's role recognition [13, 18]. As a verbal feature, we utilized the average vector based on word embedding. However, the average vector lost some information, such as fillers in utterances. We need to discuss the effectiveness of some specific words and phrases for the improvement of the model.

5.2 Discussion about S_a

In Section 5.1, we set $S_a = 30$. However, it is not clear which S_a is appropriate for the prediction model. Therefore, we compared different settings about S_a ($S_a = 5, 10, 20$, and 30.) We used all features in the comparison because the setting was the best in Section 5.1.

Table 3 shows the experimental result. We obtained better results for S_a =20 and 30, as compared with the smaller values of S_a.

We analyzed the difference in the results between $S_a = 5$ and $S_a = 30$ in detail. First, we discuss the case that the setting $S_a = 30$ was better than that of $S_a = 5$. For the case that $S_a = 30$ was better, a long silence often appeared in the S_a. It indicates the decrease of the number of utterance in the S_a. It led to the decrease of the number of words. Since the verbal features were effective in our model, long S_a was important to capture the features. On the other hand,

Table 3. Comparison on S_a.

S_a [sec]	P	R	F
5	0.71	0.71	0.71
10	0.74	0.72	0.73
20	0.75	0.75	0.75
30	0.74	0.76	0.75

Table 4. Recall about each DA.

DA tag	Num	R
Inform	1305	0.75
Suggest	311	0.76
Offer	122	0.78
Elicit-Inform	178	0.79
Elicit-Offer-or-Suggestion	31	0.77
Elicit-Assessment	104	0.83
Elicit-Comment-about-Understanding	2	0.5

the setting $S_a = 5$ was unfitted and unsuitable in the case that S_a contained a long silence because the model cannot adequately capture the verbal features due to a small number of words.

Next, we discuss another situation, that is $S_a = 5$ was better. In this situation, the problem was also opposite to the setting that $S_a = 30$ was better. In other words, $S_a = 5$ performed well in the case that S_a contained many words and no long silence in S_a. Information about important words in S_a vanished by the average vector generated from too many words. Thus, the optimal length of S_a depends on the tendency in each S_a.

The goal of this model is to detect the timing of facilitating utterances for our digital facilitator system. In other words, we need to ensure real-time prediction in discussion. Thus, the small S_a is essentially suitable although the current result is the opposite. We need to investigate the optimal S_a through the addition of effective features described in Section 5.1.

5.3 Discussion about Dialogue Act

As we used Dialogue Act (DA) tags in the data construction, DA tags are closely related to social roles because DA tags denote the intention of each utterance. Hence, we compared the recall rate of each DA tags in the model with all features[3] and $S_a = 30$.

Table 4 shows the experimental result about some specific DA tags that are closely related to facilitating utterances, such as Inform and Elicit-*. "Num" in the table denotes the number of instances with each DA tag in the experimental

[3] Note that our model did not use any DA tags as features.

Table 5. Comparison with a baseline.

Model	P	R	F	AUC
Baseline	0.65	1.00	0.79	0.50
Ours	0.74	0.76	0.75	**0.70**

data set. We almost obtained balanced results, namely 0.75 or more. Although the recall rate of "Elicit-Comment-about-Understanding" was not enough, the reason was the number of instances in the data set.

The DA tag with the best recall rate was "Elicit-Assessment." In other words, the previous utterances of "Elicit-Assessment" contained much information for the prediction. The guideline of the AMI corpus [1] said,

> In an ELICIT-ASSESSMENT, the speaker attempts to elicit an assessment (or assessments) about what has been said or done so far. Sometimes a speaker seems to be making a suggestion and eliciting an assessment about it at the same time.

This definition entails a part of the roles of facilitating utterances. However, utterances with the DA tags listed in Table 4 are not always facilitating utterances. Therefore, it is insufficient to predict timing of facilitating utterances by using only the DA tags although they are important features for prediction of facilitating utterances. Moreover, DA tags cannot be applied to a real-time prediction model easily due to need of annotation by human annotators. On the other hand, our model was able to predict facilitating utterances in a relatively high accuracy without DA tag information.

5.4 Comparison with a Baseline

We compared our model with a baseline. The baseline was based on a simple and naive assumption; each S_p always contains one or more facilitating utterances. In other words, the baseline always produced +1 for each S_p, namely the majority of the label in the data set. For the evaluation, in addition to P, R, and F-values, we introduce the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC). The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC is the area under the ROC curve. Therefore the AUC of the baseline becomes 0.5 from the definition.

Table 5 shows the experimental result. The F-value of the baseline was more than that of our model due to the perfect Recall rate. However, our method outperformed the baseline in terms of the AUC value. It shows the effectiveness and appropriateness of our method as compared with the baseline.

6 Conclusion

In this paper, we proposed a model for predicting the timing of facilitating utterances for the digital facilitator. We defined an utterance by a participant that behaves like a chairperson on the discussion as facilitating utterances. We designed a guideline for constructing training data from the AMI corpus for a timing prediction task of facilitating utterances. It was based on dialogue act tags and social role tags in the corpus.

We applied verbal and non-verbal features for the timing prediction model of facilitating utterances. We evaluated our model in terms of types of features, length of S_a, types of dialogue acts, and comparison with a baseline. The verbal features were effective for the prediction. The non-verbal features also performed a certain function for the prediction. Our model also outperformed a simple baseline in terms of the AUC value. As a whole, the experimental results show the effectiveness of our prediction model.

To achieve better accuracy, we need to apply other features that are obtained from speech information, such as pitch and volume. We also need to discuss the effectiveness of some specific words and phrases for the improvement of the model. In the experiment, the S_a of the best F-value was 30. However, the optimal length of S_a depends on the tendency in each S_a. Dynamic setting of the S_a is interesting future work.

In this paper, we handled the AMI corpus as the data set. We have also developed a multi-party conversation corpus, the Kyutech corpus [19]. Shiota et al. [14] reported the difference between the AMI corpus and the Kyutech corpus, for the facilitators' behavior. Therefore, analysis and evaluation on our corpus for this timing prediction are important future work.

One big issue of our work is the validity of the rules for determination of facilitating utterances in Section 3.2. We automatically constructed the training data by using the rules. The rules were based on annotated tags in the AMI corpus and our heuristics, and were intuitively plausible. However, we need to discuss the validity more deeply through manual data analysis.

There are many approaches and studies to build mutually agreeable solutions and a consensus, such as multi-agent systems for negotiation [15] and a large-scale online discussion [5]. Studies about facilitation robots [8] and human communication skills [10] are also related to our work. The knowledge from these studies would lead to the improvement of our work.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 17H01840.

References

1. Guidelines for Dialogue Act and Addressee Annotation Version 1.0, 2005.

2. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.
3. Jean Carletta. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007.
4. Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.
5. Takayuki Ito, Yuma Imi, Takanori Ito, and Eizo Hideshima. COLLAGREE: A facilitator-mediated large-scale consensus support system. In *Proceedings of the 2nd Collective Intelligence Conference*, 2014.
6. Ryunosuke Kirikihira and Kazutaka Shimada. Discussion map with an assistant function for decision-making: A tool for supporting consensus-building. In *International Conference on Collaboration Technologies*, pages 3–18. Springer, 2018.
7. Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 127–136, 2017.
8. Yoichi Matsuyama, Iwao Akiba, Shinya Fujie, and Tetsunori Kobayashi. Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. *Computer Speech & Language*, 33(1):1–24, 2015.
9. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
10. Shogo Okada, Yoshihiko Ohtake, Yukiko I. Nakano, Yuki Hayashi, Hung-Hsuan Huang, Yutaka Takase, and Katsumi Nitta. Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 169–176, 2016.
11. Yoshimasa Omoto, Yasushi Toda, Kazuhiro Ueda, and Toyoaki Nishida. Analyses of the Facilitating Behavior by Using Participant's Agreement and Nonverbal Behavior. *Journal of Information Processing Society of Japan*, 52(12):3659–3670, 2011. (in Japanese).
12. Ashtosh Sapru and Hervé Bourlard. Automatic social role recognition in professional meetings using conditional random fields. In *Proceedings of Interspeech*, 2013.
13. Ashtosh Sapru and Hervé Bourlard. Automatic recognition of emergent social roles in small group interactions. *IEEE Transactions on Multimedia*, 17(5):746–760, 2015.
14. Tsukasa Shiota, Takashi Yamamura, and Kazutaka Shimada. Analysis of facilitators' behaviors in multi-party conversations for constructing a digital facilitator system. In *International Conference on Collaboration Technologies*, pages 145–158, 2018.
15. Carles Sierra, Nicholas R. Jennings, Pablo Noriega, and Simon Parsons. A framework for argumentation-based negotiation. In *International Workshop on Agent Theories, Architectures, and Languages: ATAL 1997, LNCS 1365*, pages 177–192, 1997.
16. Gabriel Skantze. Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In *SIGdial Conference*, 2017.

17. Alessandro Vinciarelli, Fabio Valente, Sree Harsha Yella, and Ashtosh Sapru. Understanding social signals in multi-party conversations: Automatic recognition of socio-emotional roles in the ami meeting corpus. In *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pages 374–379, 2011.
18. Felix Weninger, Jarek Krajewski, Anton Batliner, and Björn Schuller. The voice of leadership: Models and performances of automatic analysis in online speeches. *IEEE Transactions on Affective Computing*, 3(4):496–508, Fourth 2012.
19. Takashi Yamamura, Kazutaka Shimada, and Shintaro Kawahara. The Kyutech corpus and topic segmentation using a combined method. In *Proceedings of the 12th Workshop on Asian Language Resources*, pages 95–104, 2016.